Maximum entropy discrimination

Tommi Jaakkola[†]
tommi@ai.mit.edu

Marina Meila[†]
mmp@ai.mit.edu

Tony Jebara[‡]
jebara@media.mit.edu

[†] MIT AI Lab, 545 Technology Square, Cambridge, MA 02139
[‡] MIT Media Lab, 20 Ames Street, Cambridge, MA 02139

August 18, 1999

Abstract

We present a general framework for discriminative estimation based on the maximum entropy principle and its extensions. All calculations involve distributions over structures and/or parameters rather than specific settings and reduce to relative entropy projections. This holds even when the data is not separable within the chosen parametric class, in the context of anomaly detection rather than classification, or when the labels in the training set are uncertain or incomplete. Support vector machines are naturally subsumed under this class and we provide several extensions. We are also able to estimate exactly and efficiently discriminative distributions over tree structures of class-conditional models within this framework. Preliminary experimental results are indicative of the potential in these techniques.

1 Introduction

Effective discrimination is essential in many application areas including speech recognition, image classification or identification of molecular binding sites in genomic DNA. Statistical approaches used in these contexts for classification generally fall into two major categories – generative or discriminative – depending on the estimation criterion used for adjusting the model parameters and/or structure. Generative approaches rely on a full joint probability distribution over examples and classification labels whereas for discriminative methods only the conditional relation of a label given the example is relevant. While the full joint distribution in the generative approach carries a number of advantages e.g. in handling incomplete examples, the typical estimation criterion (maximum likelihood or its variatiants) is nevertheless suboptimal from the point of view of classification objective. Discriminative methods such as support vector machines [21] or boosting algorithms [8] that focus directly on the parametric decision boundary typically yield more robust classification methods, whenever they are applicable.

Full joint distributions and the benefits they convey can be, of course, exploited in discriminative approaches as well. We may, for example, interprete the posterior probability of a label given the example as a parametric decision boundary (see e.g. [10, 13]). Alternatively, we can induce suitable vector space representations for examples from generative models and feed such representations into standard discriminative techniques [11].

In this paper we provide a more general notion of discrimination, one that applies also in the contex of anomaly detection or when the classification labels themselves are uncertain or missing. Note that the utility of e.g. unlabeled examples is not obvious [22, 2, 4, 18]. Our approach towards general discriminative training relies on the well known maximum entropy principle which embodies the Bayesian integration of prior information with observed constraints (see e.g. [15]). The formalism that we apply and extend in this paper allows, for example, a feasible discriminative training of both the parameters and the structure of a class of joint probability models. The approach is not limited to probability models, however, and we extend e.g. support vector machines.

2 Maximum entropy classification

Consider first a two-class classification problem where labels $y \in \{-1, 1\}$ are assigned to examples $X \in \mathcal{X}$. Assume we have two class-conditional probability distributions over the examples, i.e., $P(X|\theta_y)$ with parameters θ_y, one for each class. The decision rule corresponding to any particular parameter setting $\{\theta_{\pm 1}\}$ follows the sign of the *discriminant function*:

$$\mathcal{L}(X|\Theta) = \log \frac{P(X|\theta_1)}{P(X|\theta_{-1})} + b \tag{1}$$

where $\Theta = \{\theta_1, \theta_{-1}, b\}$ and b is a bias term, usually expressed as a log-ratio of prior class probabilities $b = \log p/(1-p)$. The class-conditional distributions here may come from different families of distributions or we might specify the parametric discriminant function directly without any reference to probability models. The parameters θ_y may also include the model structure as seen later in the paper.

The parameters $\Theta = \{\theta_1, \theta_{-1}, b\}$ in the discriminant function should be chosen to maximize classification accuracy. Instead of finding a single parameter setting, we consider here a more general problem of finding a distribution $P(\Theta)$ over the parameters and using a convex combination of discriminant functions, i.e.,

$$\int P(\Theta)\mathcal{L}(X|\Theta)d\Theta \tag{2}$$

in place of the original discriminant function in the decision rule. The problem is now to find an appropriate distribution $P(\Theta)$. Given a set of training examples $\{X_1, \ldots, X_T\}$ and corresponding labels $\{y_1, \ldots, y_T\}$ we seek for a distribution $P(\Theta)$ that makes the least assumptions about the choice of the parameter values Θ while giving rise to a discriminant function that correctly separates the training examples. We can formalize this as a *maximum entropy* (ME) estimation problem. In other words, we maximize the entropy $H(P)$ of P subject to the classification constraints

$$\int P(\Theta)\ [\, y_t\, \mathcal{L}(X_t|\Theta)\,]\ d\Theta \geq \gamma \tag{3}$$

for all $t = 1, \ldots, T$. Here γ specifies a desired classification margin. We note that the solution is unique (provided that it exists) since $H(P)$ is concave and the linear constraints specify a convex region. Note that the preference towards high entropy distributions (fewer assumptions) applies only within the admissible set of distributions $\mathcal{P}_\gamma$ consistent with the classification constraints.

We can readily extend this formulation to a multi-class setting by introducing additional classification constraints. To see this, suppose we have instead m class-conditional probability models

$P(X|\theta_y)$, $y = 1, \ldots, m$, prior class frequencies $\{p_y\}$, and the associated pairwise discriminant functions

$$\mathcal{L}_{y,y'}(X_t|\Theta) = \log \frac{P(X|\theta_y)}{P(X|\theta_{y'})} + \log \frac{p_y}{p_{y'}} \tag{4}$$

where $\Theta = \{\theta_1, \ldots, \theta_m, p_1, \ldots, p_m\}$. We may now replace the single constraint per training example in eq. (3) with the following $m-1$ pairwise constraints

$$\int P(\Theta)\ [\,\mathcal{L}_{y_t,y}(X_t|\Theta)\,]\ d\Theta \geq \gamma,\ \ y \neq y_t, \tag{5}$$

to ensure that the training label y_t always "wins" the competition against the alternative labels $y \neq y_t$. For notational simplicity we will consider primarily only binary classification problems in the remainder of the paper but emphasize that the analogous extension to a multi-class setting can be made.

The overall ME formulation presented so far has several problems. We have, for example, made a tacit assumption that the training examples can be separated with the specified margin. This assumption may very well be violated in practice. Moreover, we may have a prior reason to prefer some parameter values over others (as well as margin constraints) which requires us to incorporate a prior distribution $P_0(\Theta, \gamma)$ into the definition. Other extensions and generalizations will be discussed later in the paper.

A more general formulation that addresses these concerns is given by the following *minimum relative entropy* principle:

Definition 1 *Let $\{X_t, y_t\}$ be the training examples and labels, $\mathcal{L}(X|\Theta)$ a parametric discriminant function, and $\gamma = [\gamma_1, \ldots, \gamma_t]$ a set of margin variables. Assuming a prior distribution $P_0(\Theta, \gamma)$, we find the discriminative minimum relative entropy (MRE) distribution $P(\Theta, \gamma)$ by minimizing*

$$D(P \| P_0) = \int P(\Theta) \log \frac{P(\Theta)}{P_0(\Theta)}\, d\Theta \tag{6}$$

subject to the (soft) classification constraints

$$\int P(\Theta, \gamma)\ [\,y_t\, \mathcal{L}(X_t|\Theta) - \gamma_t\,]\ d\Theta d\gamma \geq 0 \tag{7}$$

for all t. The decision rule for any new example X is given by

$$\hat{y} = \text{sign}\left(\int P(\Theta)\, \mathcal{L}(X|\Theta)\, d\Theta \right) \tag{8}$$

Let us make a few remarks about the definition. First, we can recover the previous ME formulation by appropriately adjusting the prior distribution $P_0(\Theta, \gamma)$ (e.g., if $P_0(\gamma)$ peaks around a specific setting of the margins). It is clear that the margin constraints are hidden in the prior distribution $P_0(\gamma)$. Second, if we assume that there is a non-zero prior probability for all γ_t taking some negative values, we guarantee that the admissible set $\mathcal{P}$ composed of all distributions $P(\Theta, \gamma)$ consistent with the classification constraints, is never empty. Thus even when the examples cannot be separated by any discriminant function in the chosen parametric class (e.g. linear), we get a valid and unique solution. Third, the penalty for violating any of the margin constraints also depends on the prior distribution P_0; whenever the mean of γ_t deviates from its prior mean under P_0, we incur a penalty

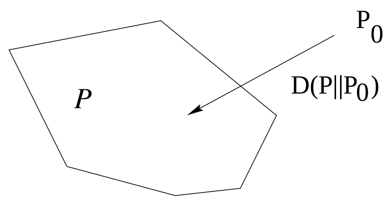


Figure 1: Minimum relative entropy (MRE) projection from the prior distribution to the admissible set.

in the form of relative entropy distance between the corresponding distributions. It is worth noting that the penalties are defined in terms of joint specifications of margins but, in certain cases, they reduce to the more typical additive penalties of violating the constraints.

The prior $P_0(\Theta, \gamma)$ playes an important role in our definition and we must choose it appropriately. Let us consider here only the prior over the margin constraints γ. Supposing again that $P_0(\Theta, \gamma) = P_0(\Theta) P_0(\gamma)$, we can, for example, set

$$P_0(\gamma) = \prod_t P_0(\gamma_t) \tag{9}$$

where $P_0(\gamma_t) = c\, e^{-c(1-\gamma_t)}$, for $\gamma_t \leq 1$. A penalty is incurred for margins smaller than $1 - 1/c$ (the prior mean of γ_t) while margins larger than this are not penalized. In the latter case, the associated constraint becomes merely irrelevant. We will see in later sections that this choice of the margin prior corresponds closely to the use of slack variables and additive penalties used in support vector machines. A number of other choices for $P_0(\gamma)$ are possible and we discuss some of them later in the paper.

An important property of the MRE solution is that it can be viewed as a relative entropy projection, the e-projection in the terminology of [1], from the prior distribution $P_0(\Theta, \gamma)$ to the admissible set $\mathcal{P}$. Figure 1 illustrates this view. Even in the non-separable case, we can view the MRE solution as a projection. This formalism readily extends to the case of uncertain or partially labeled examples as we will see later in the paper.

To solve the MRE problem, we rely on the following theorem.

Theorem 1 *The solution to the MRE problem has the following general form (cf. [7]):*

$$P(\Theta, \gamma) = \frac{1}{Z(\lambda)} P_0(\Theta, \gamma)\, e^{\sum_t \lambda_t [\, y_t \mathcal{L}(X_t|\Theta) - \gamma_t]} \tag{10}$$

where $Z(\lambda)$ is the normalization constant (partition function) and $\lambda = \{\lambda_1, \ldots, \lambda_T\}$ defines a set of non-negative Lagrange multipliers, one for each classification constraint. λ are set by finding the unique maximum of the following jointly concave objective function:

$$J(\lambda) = -\log Z(\lambda) \tag{11}$$

Whether the MRE solution can be found in a feasible way depends entirely on whether we can evaluate the partition function $Z(\lambda)$,

$$Z(\lambda) = \int P_0(\Theta, \gamma)\, e^{\sum_t \lambda_t [\, y_t \mathcal{L}(X_t|\Theta) - \gamma_t]}\, d\Theta d\gamma \tag{12}$$

in closed form. Given a closed form expression for $Z(\lambda)$, the maximum of the jointly concave objective function $J(\lambda)$ can be subsequently found through any standard convex optimization method such as Newton-Raphson. The resulting set of Lagrange multipliers $\{\lambda_t\}$ then define the MRE solution as indicated in the theorem. Finally, predicting a label for any new example X involves averaging the discriminant function $\mathcal{L}(\Theta)$ with respect to the marginal $P(\Theta)$ of the MRE distribution (see Definition 1). Finding this marginal as well as performing the required averaging are no more costly than computing $Z(\lambda)$. We will elaborate these calculations further in the context of specific realizations.

The MRE solution is *sparse* in the sense that only a few Lagrange multipliers will be non-zero. This arises because many of the classification constraints become irrelevant once the constraints are enforced for a small subset of examples. For support vector machines that are subsumed under the above general definition, this notion translates into a sparse representation of the separating hyperplane. Sparsity leads to immediate generalization guarantees (independent of the dimensionality of the parameter or example space):

Lemma 1 *The generalization error ϵ_g of the MRE classifier satisfies*

$$e_g \leq E\{\,\text{fraction of non-zero Lagrange multipliers}\,\} \tag{13}$$

where the expectation is over the choice of the training set.

Practical leave-one-out cross-validation estimates of the generalization error can be derived on the basis of this result (cf. [21, 12]). We may also make use of generalization error results derived for convex combination of classifiers [20] to obtain more informative generalization bounds for MRE classifiers. The details are left for another paper.

3 Practical realization of the MRE solution

We now turn to the question of actually finding the MRE solution. Consider first the following elementary but helpful lemma

Lemma 2 *Any factorization of the prior $P_0(\Theta,\gamma)$ across any disjoint sets of variables $\{\Theta,\gamma\}$ leads to a disjoint factorization of the MRE solution $P(\Theta,\gamma)$ across the same sets of variables provided that these variables appear in distinct additive components in $y_t\mathcal{L}(X_t,\Theta)-\gamma_t$.*

If we assume that the labels $\{y_t\}$ are fixed and that the prior distribution $P_0(\Theta,\gamma)$ factorizes across the components $\{\Theta\setminus b, b, \gamma\}$, then according to the lemma, the MRE solution factorizes in the same way. This factorization property allows us to eliminate e.g. the bias term from the remaining solution by means of imposing additional constraints on the Lagrange multipliers. This is analogous to the handling of the bias term in support vector machines [21]:

Lemma 3 *Assuming $P_0(\Theta,\gamma) = P_0(\Theta\setminus b,\gamma)P_0(b)$ and $P_0(b)$ approaches a non-informative prior, then $P(\Theta,\gamma) = P(\Theta\setminus b,\gamma)P(b)$ and $P(\Theta\setminus b,\gamma)$ can be found independently from $P(b)$ provided that we require $\sum_t \lambda_t y_t = 0$.*

With the help of these results, we will consider now a few specific realizations such as support vector machines and a class of graphical models.

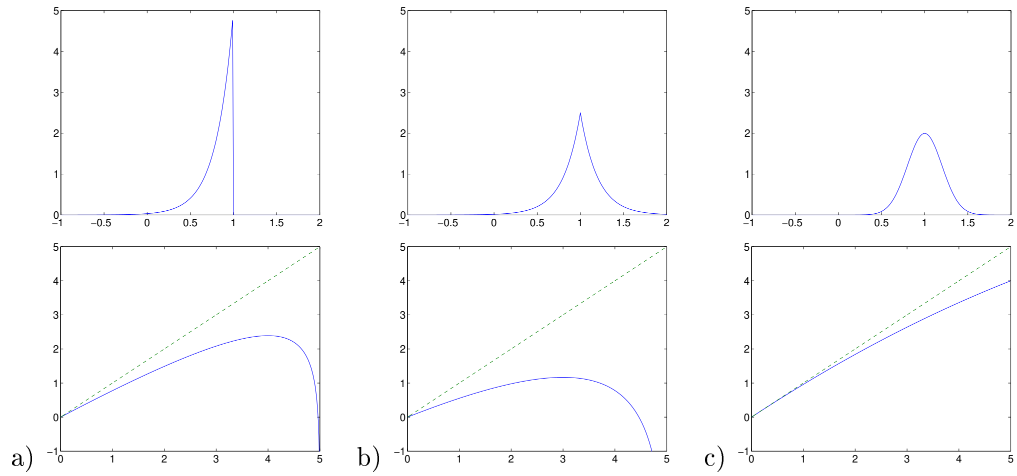


Figure 2: Three margin prior distributions (top row) and the corresponding potential terms (bottom row) from Eq. (15).

3.1 Support vector machines

It is well known that the log-likelihood ratio of two Gaussian distributions with equal covariance matrices yields a linear decision rule. With a few additional assumptions, the MRE formulation gives support vector machines:

Theorem 2 *Assuming* $\mathcal{L}(X,\Theta) = \theta^T X - b$ *and* $P_0(\Theta,\gamma) = P_0(\theta)P_0(b)P_0(\gamma)$ *where* $P_0(\theta)$ *is* $N(0,I)$*,* $P_0(b)$ *approaches a non-informative prior, and* $P_0(\gamma)$ *is given by eq. (9) then the Lagrange multipliers* λ *are obtained by maximizing* $J(\lambda)$ *subject to* $0 \leq \lambda_t \leq c$ *and* $\sum_t \lambda_t y_t = 0$*, where*

$$J(\lambda) = \sum_t [\, \lambda_t + \log(1 - \lambda_t/c)\,] - \frac{1}{2}\sum_{t,t'} \lambda_t \lambda_{t'} y_t y_{t'} (X_t^T X_{t'}) \tag{14}$$

The only difference between our $J(\lambda)$ and the (dual) optimization problem for SVMs is the additional potential term $\log(1-\lambda_t/c)$. This highlights the effect of the different miss-classification penalties, which in our case come from the MRE projection. Figures 2a) and c) show, however, that the additional potential term does not always carry a huge effect (for $c = 5$). Moreover, in the separable case, letting $c \to \infty$, the two methods coincide. The decision rules are formally identical.

The choice of the prior distribution $P_0(\gamma)$ leads to different potential terms. Figure 2 gives the following priors and their corresponding potential terms

$$\begin{array}{lll} & \text{Margin prior} & \text{Dual potential term} \\ a) & P_0(\gamma) \propto e^{-c\,(1-\gamma)}, \gamma \leq 1, & \lambda_t + \log(1-\lambda_t/c) \\ b) & P_0(\gamma) \propto e^{-c\,|1-\gamma|}, & \lambda_t + 2\log(1-\lambda_t/c) \\ c) & P_0(\gamma) \propto e^{-c^2\,(1-\gamma)^2/2}, & \lambda_t - (\lambda_t/c)^2 \end{array} \tag{15}$$

where a) is the case discussed in the theorem. Note that the resulting potential terms may or may not set an upper bound on the value of λ_t. In a) and b) λ_t is bounded by the constant c whereas in c) no such bound exists.

3.1.1 Extension

We now consider the case where the discriminant function $\mathcal{L}(X, \Theta)$ corresponds to the log-likelihood ratio of two Gaussians with different (and adjustable) covariance matrices. The parameters Θ in this case are both the means and the covariances. The prior $P_0(\Theta)$ must be the conjugate Normal-Wishart to obtain closed form integrals[1] for the partition function, Z. Here, $P(\Theta_1, \Theta_{-1})$ is $P(m_1, V_1)P(m_{-1}, V_{-1})$, a density over means and covariances (and the factorization follows from our assumptions below).

The prior distribution has the form $P_0(\Theta_1) = \mathcal{N}(m_1; m_0, V_1/k)\ \mathcal{IW}(V_1; kV_0, k)$ with parameters (k, m_0, V_0) that can be specified manually or one may let $k \to 0$ to get a non-informative prior. We used the MAP values for k, m_0 and V_0 from the class-specific data[2]. Integrating over the parameters and the margin, we get a partition function which factorizes $Z = Z_\gamma \times Z_1 \times Z_{-1}$. For Z_1 we obtain the following:

$$Z_1 \propto \ N_1^{-d/2}\ |\pi S_1|^{-N_1/2}\ \Pi_{j=1}^d \Gamma\left(\frac{N_1+1-j}{2}\right) \tag{16}$$

$$N_1 \triangleq \textstyle\sum_t w_t \quad \bar{X}_1 \triangleq \sum_t \frac{w_t}{N_1} X_t \quad S_1 \triangleq \sum_t w_t X_t X_t^T - N_1 \bar{X}_1 \bar{X}_1^T \tag{17}$$

Here, w_t is a scalar weight given by $w_t = u(y_t) + y_t \lambda_t$ for Z_1. To solve for Z_{-1} we proceed in a similar manner with the exception that the weights are set to $w_t = u(-y_t) - y_t\lambda_t$. $u(\cdot)$ here is the step function. Given Z, updating λ is done by maximizing the corresponding negative log-partition function $J(\lambda)$ subject to $0 \le \lambda_t \le c$ and $\sum_t \lambda_t y_t = 0$ where:

$$J(\lambda) = \sum_t [l_\alpha \lambda_t + \log(1 - \lambda_t/c)] - \log Z_1(\lambda_t) - \log Z_{-1}(\lambda_t) \tag{18}$$

The potential term above corresponds to integrating over the margin with a margin prior $P_0(\gamma) \propto e^{-c(l_\alpha - \gamma)}$ with $\gamma \le l_\alpha$. We pick l_α to be some α-percentile of the margins obtained under the standard MAP solution. Optimal lambda values are found via constrained gradient descent. The resulting marginal MRE distribution over the parameters (normalized by the partition function $Z_1 \times Z_{-1}$) is a Normal-Wishart distribution itself, $P(\Theta_1) = \mathcal{N}(m_1; \bar{X}_1, V_1/N_1)\ \mathcal{IW}(V_1; S_1, N_1)$ with the final λ values. Predicting the labels for a data point X under the final $P(\Theta)$ involves taking expectations of the discriminant function under a Normal-Wishart. This is simply:

$$E_{P(\Theta_1)}[\log P(X|\Theta_1)] = \text{constant} - \frac{N_1}{2}(X - \bar{X}_1)^T S_1^{-1}(X - \bar{X}_1) \tag{19}$$

We thus obtain discriminative *quadratic* decision boundaries. These extend the linear boundaries without (explicitly) resorting to *kernels*. Of course, kernels may still be used in this formalism, effectively mapping the feature space into a higher dimensional representation. However, unlike linear discrimination, the covariance estimation in this framework allows the model to adaptively modify the kernel.

3.1.2 Experiments

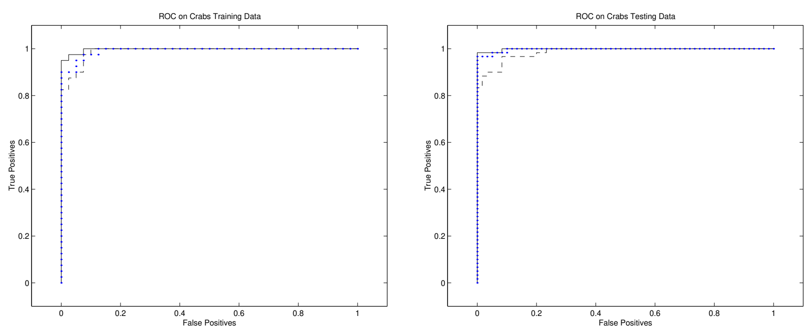
In the following, we show results using the minimum relative entropy approach where the discriminant function $(\mathcal{L}(X, \Theta))$ is the log-ratio of Gaussians with variable covariance matrices on standard 2-class classification problems (Leptograpsus Crabs and Breast Cancer Wisconsin). In

[1] This can be done more generally for conjugate priors in the exponential family.

[2] The prior here is the posterior distribution over the parameters given the data, i.e. an empirical Bayes procedure.

Method	Training Errors	Testing Errors
Neural Network (1)		3
Neural Network (2)		3
Linear Discriminant		8
Logistic Regression		4
MARS (degree $= 1$)		4
PP (4 ridge functions)		6
Gaussian Process (HMC)		3
Gaussian Process (MAP)		3
SVM - Linear	**5**	**3**
SVM - RBF $\sigma = 0.3$	**1**	**18**
SVM - 3rd Order Polynomial	**3**	**6**
Maximum Likelihood Gaussians	**4**	**7**
MaxEnt Discrimination Gaussians	**2**	**3**

Table 1: Leptograpsus Crabs



(a) Training ROC (b) Testing ROC

Figure 3: ROC curves on Leptograpsus Crabs for discriminative (solid line), Bayes / ML models (dashed line) and SVM linear models (dotted line).

addition we display a two-dimensional visualization example of the classification. Performance is compared to regular support vector machines, maximum likelihood estimation and other methods.

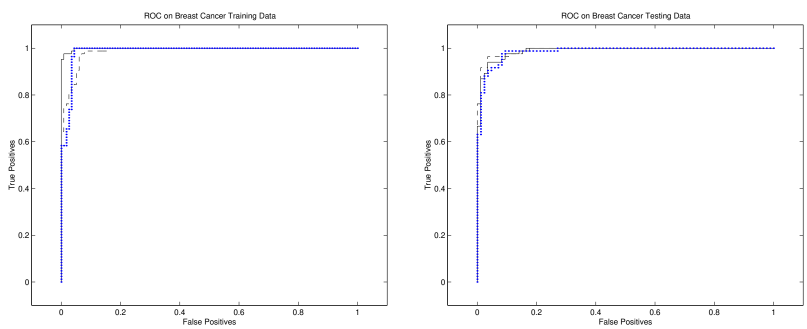
The Leptograpsus crabs data set was originally provided by Ripley [19] and further tested by Barber and Williams [3]. The objective is to classify the sex of the crabs from 5 scalar anatomical observations. The training set contains 80 examples (40 of each sex) and the test set includes 120 examples.

The Gaussian based decision boundaries are compared in Table 1 against other models from[3]. The table shows that the maximum entropy (or minimum relative entropy) criterion improves the Gaussian discrimination performance to levels similar to the best alternative models. The bias was estimated separately from training data for both the maximum likelihood Gaussian models and the maximum entropy discrimination case. In addition, we show the performance of a support vector machine (SVM) with linear, radial basis and polynomial decision boundaries (using the Matlab SVM Toolbox provided by Steve Gunn). In this case, the linear SVM is limited in flexibility while kernels exhibit some over-fitting.

In Figure 3 we plot the ROC curves on training and testing data. The ROC curve shows improved classification for maximum entropy (minimum relative entropy) case.

Method	Training Errors	Testing Errors
Nearest Neighbour		11
SVM - Linear	**8**	**10**
SVM - RBF $\sigma = 0.3$	**0**	**11**
SVM - 3rd Order Polynomial	**1**	**13**
Maximum Likelihood Gaussians	**10**	**16**
MaxEnt Discrimination Gaussians	**3**	**8**

Table 2: Breast Cancer Classification



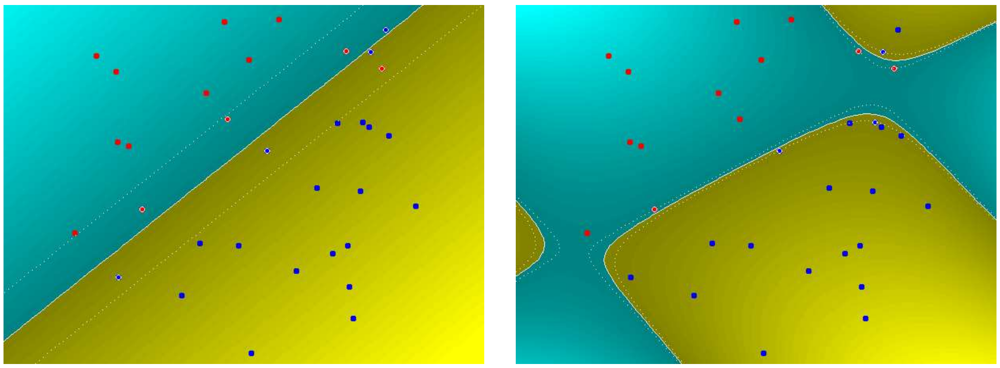
(a) Training ROC (b) Testing ROC

Figure 4: ROC curves on Breast Cancer for discriminative (solid line), Bayes / ML models (dashed line) and SVM linear models (dotted line).

Another data set which was tested was the Breast Cancer Wisconsin data where the two classes (malignant or benign) have to be computed from 9 numerical attributes from the patients (200 training cases and 169 test cases). The data was first presented by Wolberg [24]. We compare our results to those produced by Zhang [25] who used a nearest neighbour algorithm to achieve 93.7% accuracy. As can be seen from Table 2, over-fitting seems to prevent good performance for kernel based SVMs. The maximum entropy discriminator achieves 95.3% accuracy.

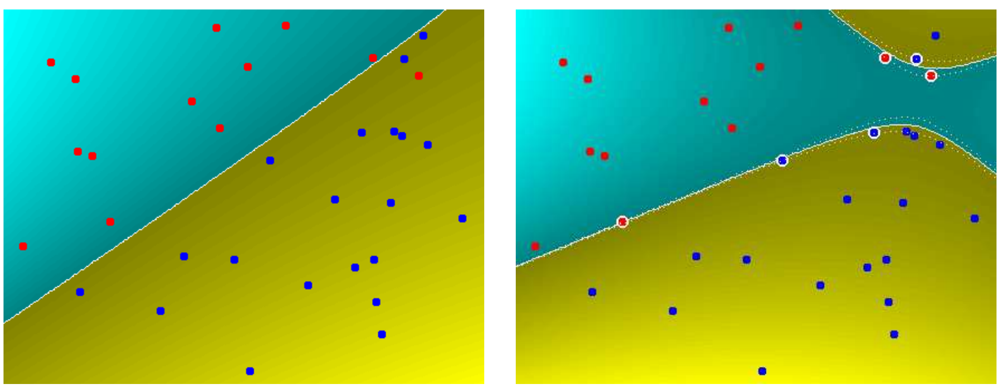
In Figure 4 we plot the ROC curves on training and testing data. The training ROC curves show improved discrimination for the maximum entropy method. ROC curves for all three methods are equivalent on testing however since we typically assume that bias is estimated exclusively from training data, the results in Table 2 are more significant.

Finally, for visualization, we present the technique on a 2D set of training data in Figure 5 and Figure 6. The SVM in Figure 5(a) attempts to achieve maximum descrimination but is limited to a linear decision boundary. It only succeeds after the application of a kernel as in Figure 5(b), where a 3rd order polynomial kernel is used. In Figure 6(a), the maximum likelihood technique is used to estimate a 2 Gaussian discrimination boundary (bias is estimated separately) which has more flexibility than the linear SVM yet fails to achieve the desired optimal classification. Meanwhile, the maximum entropy discrimination technique places the Gaussians in the most discriminative configuration as shown in Figure 6(b).



(a) Linear SVM (b) Polynomial Kernel SVM

Figure 5: Classification visualization SVMs.



(a) Max Likelihood (b) Max Ent Discrimination

Figure 6: Classification visualization for Gaussian discrimination.

3.2 The Fisher kernel classifier

Here we demonstrate that the MRE formulation proposed in this paper contains the Fisher kernel method of [11]. The Fisher kernel method provides a combination of a generative model $P(X|\theta)$ with a discriminative method such as support vector machines through defining an appropriate kernel function. The kernel function, called the Fisher kernel, can be computed from any generative model in the neighborhood of some desired e.g. maximum likelihood parameter setting θ^*. The Fisher kernel function is given by

$$K_{fk}(X, X') = U_X(\theta^*)^T F(\theta^*)^{-1} U_{X'}(\theta^*) \tag{20}$$

where $U_X(\theta)$ is the Fisher score

$$U_X(\theta) = \nabla_\theta \log P(X|\theta)_{|\theta=\theta^*}, \tag{21}$$

$F(\theta) = E\{\, U_X(\theta) U_X^T(\theta)\,\}$ is the Fisher information matrix[3] and the expectation is with respect to $P(X|\theta)$. Replacing the inner product $X_t^T X_{t'}$ between the examples in Theorem 2 with the kernel function in Eq. (20) amounts to the "simple" Fisher kernel method as explained in [11].

Our goal in this section is to show that we can recover the Fisher kernel method in the MRE framework so long as the prior distribution $P_0(\theta, \gamma)$ is chosen in an appropriate way. We start with a few necessary regularity assumptions about the family of distributions $P(X|\theta)$ in some small (open) neighborhood $O(\theta^*)$ of θ^*:

1. for any $X \in \mathcal{X}$, $U_X(\theta) = \nabla_\theta \log P(X|\theta)$ is a continuously differentiable vector valued function of θ
2. $F(\theta) = E\{\, U_X(\theta) U_X^T(\theta)\,\}$ exists and is positive definite

Let us define, in addition, the differential (symmetric) relative entropy distance between the distributions $P(X|\theta)$ and $P(X|\theta^*)$

$$d(\theta, \theta^*)^2 = \frac{1}{2}(\theta - \theta^*)^T\, F(\theta^*)^{-1}\,(\theta - \theta^*) \tag{22}$$

valid whenever $\theta \approx \theta^*$. We assign a prior distribution $P_0(\theta)$ in terms of this distance[4]

$$P_0(\theta) = \frac{1}{Z(\theta^*, \beta)}\, e^{-\beta\, d(\theta,\theta^*)^2} \tag{23}$$

where β serves as a scaling parameter. This prior assigns a low probability to all θ for which the corresponding probability distribution $P(X|\theta)$ deviates significantly from $P(X|\theta^*)$. Another way to view this prior is as a local isotropic Gaussian prior distribution in the *probability manifold* induced by the family of distributions $P(X|\theta)$, $\theta \in O(\theta^*)$.

In the MRE formalism the objective is to minimize the relative entropy distance between the MRE distribution P and the prior P_0 subject to the classification constraints

$$\int P(\Theta, \gamma)\, [\, y_t\, \mathcal{L}(X_t|\Theta, \beta) - \gamma_t\,]\, d\Theta d\gamma \geq 0 \tag{24}$$

where the discriminant function $\mathcal{L}(X_t|\Theta, \beta)$ is the scaled log-likelihood ratio:

$$\mathcal{L}(X_t|\Theta, \beta) = [\, \beta^{1/2}\, \log \frac{P(X_t|\theta)}{P(X_t|\theta^*)} - b\,] \tag{25}$$

and $\Theta = \{\theta, b\}$. This discriminant function encourages parameter values θ that are indicative of the $+1$ class relative to the "null model" $P(X_t|\theta^*)$.

The following Theorem now establishes the desired connection to the Fisher kernel method.

Theorem 3 *If we replace $P_0(\theta)$ with Eq. (23) in Theorem 2 and the discriminant function with $\mathcal{L}(X_t|\Theta, \beta)$ defined above as well as let $\beta \to \infty$, then the objective function $J(\lambda)$ reduces to*

$$J(\lambda) = \sum_t [\, \lambda_t + \log(1 - \lambda_t/c)\,] - \frac{1}{2}\sum_{t,t'} \lambda_t \lambda_{t'} y_t y_{t'} K_{fk}(X_t, X_{t'}) \tag{26}$$

where $K_{fk}(X_t, X_{t'})$ is the Fisher kernel of Eq. (20).

We note that this result is merely a formal relation between the MRE principle and the Fisher kernel and does not necessarily provide any additional motivation.

[3] For many probability distributions the Fisher information matrix may not be possible to compute in closed form. However, it is the covariance matrix of the Fisher scores and thus can be easily approximated by sampling.

[4] A more precise definition of this prior would involve setting it to zero outside the open neighborhood where the regularity conditions may no longer hold. For large β, the effect of this condition vanishes and we omit it here for simplicity.

3.3 Graphical models

The MRE formulation can accomodate discriminant functions resulting from log-ratios of general graphical models. The MRE distribution, i.e. $P(\Theta)$, in this setting is over both the parameters and the structure of the model. Since the estimation is carried out in the space of distributions the distinction between discrete or continuous variables is immaterial. The framework does not, however, admit efficient solutions without restrictions on the class of graphical models. For example, assuming the structure remains fixed and that the class-conditional models have no latent variables, then the MRE distribution $P(\Theta)$ over the parameters can be obtained efficiently. This requires additional technical assumptions such as the use of conjugate priors, the parameter independence assumption of [6] and the fact that the probability model must be tractable for any fixed setting of the parameters. Although restricted, this class does include e.g. naive Bayes models, mixture of tree models and so on.

For a special class of graphical models whose structure is a tree, both the parameters and the structure can be estimated efficiently within our discriminative framework. In the remainder, we will consider such tree structured models.

First, we define a tree distribution. Let V denote the set of variables of interest, $|V| = n$, $x_v \in \mathcal{X}_v$ a particular value of $v \in V$ and $X \in \mathcal{X}$ an assignment to all the variables in V. Like any graphical model, a tree distribution is defined in two stages. First, one defines a graph (V, E), called *structure*, whose vertices are the variables in V and whose edges encode dependencies between these variables. A tree is an undirected graph over V that is connected and has no cycles. For any tree over n vertices $|E| = n - 1$. Because such a tree *spans* all the nodes in V, it is often called a *spanning tree*. Then, the *tree distribution* is defined as a product of factors corresponding to the edges and vertices.

$$T(x) \;=\; \frac{\prod_{(u,v)\in E} T_{uv}(x_u, x_v)}{\prod_{v \in V} T_v(x_v)^{\deg v - 1}} \tag{27}$$

where $\deg v$ is the *degree* of vertex v, i.e. the number of edges incident to $v \in V$ and T_{uv} and T_v denote the marginals of T:

$$\begin{aligned} T_{uv}(x_u, x_v) &= \sum_{v=x_v, u=x_u} T(X) \\ T_v(x_v) &= \sum_{v=x_v} T(X). \end{aligned}$$

When the variable x is discrete, the marginals T_{uv} and T_v can be represented as probability tables denoted respectively $\theta_{uv}(x_u, x_v)$ and $\theta_v(x_v)$. The values θ are the *parameters* of the distribution. When it will be necessary to emphasize the dependence of the tree distribution on its structure and parameters we will use the notation $T(x|E, \theta)$.

By taking the logarithm of $T(X)$ and conveniently grouping the factors one obtains

$$\log T(X) \;=\; \underbrace{\sum_{v\in V} \log T_v(x_v)}_{w_0(X)} + \sum_{uv \in E} \underbrace{\log \frac{T_{uv}(x_u, x_v)}{T_v(x_v) T_u(x_u)}}_{w_{uv}(X)} \;=\; w_0(X) + \sum_{uv\in E} w_{uv}(X). \tag{28}$$

In words, the log-likelihood is a sum of terms $w_{uv}(X)$ each corresponding to an edge (and depending only on the values of the variables u, v associated with that edge) plus a structure independent term $w_0(X)$ that depends on all the variables. All the terms are functions of the tree parameters θ.

3.3.1 Discriminative learning of tree structures

A tree model is defined by a set of discrete variables encoding its structure and a set of continuous variables representing its parameters. To use the MRE framework we must define a prior joint distribution over the structures and their associated parameters. We will assume that the structure and the parameters are independent *a priori*; moreover, we shall assume that except for the functional dependencies among the parameters that are imposed by the fact that they have to represent a valid joint distribution over $\mathcal{X}$ there are no other statistical or functional dependencies. These assumptions correspond to the *parameter independence* and *parameter modularity* assumptions of [9] (see also [6]). In our case, this means that there is a set of parameters $\theta = \{\theta_{uv}(i,j), u, v \in V, i \in \mathcal{X}_u, j \in \mathcal{X}_v\}$ associated with the edges such that in any tree model containing an edge $uv \in E$, the pairwise marginals $T_{uv}(x_u, x_v)$ are given by $\theta_{uv}(x_u, x_v)$ regardless of the presence of other edges in E and their parameter values. This simplification, in turn, allows the MRE formulation for only structures (with a fixed set of parameters or a fixed distribution over their values), for parameters only, or for both.

We start with a MRE estimation of structures only when the pairwise marginals $\theta_{uv}(x_u, x_v)$ are assumed fixed. Note that each tree nevertheless makes use of a different set of $n-1$ edges and thereby a different set of parameters. For each class or label $s \in \{1, -1\}$, we have a separate set of fixed parameters θ^s. In the experiments below, the values of these parameters were obtained from empirical (class-conditional) marginals. We assume a uniform prior over the class-conditional tree structures E_s.

Definition 2 *Given a set $(X^t, y^t), t = 1, \ldots T$ of labeled examples, a set of margin variables $\gamma = [\gamma_1, \ldots, \gamma_T]$ and a prior distribution $P_0(E_1, E_{-1}, \gamma)$ the MRE distribution $P(E_1, E_{-1}, \gamma)$ is the one minimizing $D(P\|P_0)$ subject to*

$$\sum_{E_1, E_{-1}} \int P(E_1, E_{-1}, \gamma) \left[y_t \log \frac{T(X_t|E_1, \theta^1)}{T(X_t|E_{-1}, \theta^{-1})} - \gamma_t \right] d\gamma \ \geq 0 \quad \text{for } t = 1, \ldots T \tag{29}$$

Assuming $P_0(E_1, E_{-1}, \gamma) = P_0(E_1)P_0(E_{-1})P_0(\gamma)$, Lemma 2 implies that the solution is factored as $P(E_1)P(E_{-1})P(\gamma)$ with

$$P(E_s) \ = \ \frac{1}{Z_s} e^{\sum_{t=1}^{T} s\lambda_t y_t [w_0^s(X_t) + \sum_{uv \in E_s} w_{uv}^s(X_t)]} = \frac{W_0^s}{Z_s} \prod_{uv \in E_s} W_{uv}^s \tag{30}$$

for $s = 1, -1$ and

$$W_0^s \ = \ e^{\sum_t s\lambda_t y_t w_0^s(X_t)}, \quad W_{uv}^s \ = \ \prod_{t=1}^{T} (w_{uv}^s(X_t))^{s\lambda_t y_t}, \quad s = 1, -1. \tag{31}$$

In the above the normalization constants Z_s and the factors W^s are functions of the Lagrange multipliers λ which need to be set. Provided that we can obtain the normalization constants (functions) Z_s in closed form, λ are set to maximize the dual objective

$$J(\lambda) \ = \ \gamma \cdot \lambda - \log Z_1 - \log Z_{-1}. \tag{32}$$

where, for simplicity, we have assumed a fixed setting of the margin variables $\{\gamma_t\}$.

3.4 Computing the normalization constant and its derivatives

The number of all possible tree structures over n vertices is n^{n-2} [23] and thus computing the normalization constants by enumerating all the tree structures is clearly not possible for reasonable

n. However, a remarkable graph theory result enables us to perform all the necessary summations in closed form in polynomial time. This is the *Matrix Tree Theorem* quoted below.

Theorem 4 (Matrix Tree Theorem)[23] *Let $G = (V, E)$ be a multigraph and denote by $a_{uv} = a_{vu} \geq 0$ the number of undirected edges between vertices u and v. Then the number of all spanning trees of G is given by $|A|_{uv}(-1)^{(u+v)}$ the value of the determinant obtained from the following matrix by removing row u and column v[5].*

$$A \;=\; \begin{bmatrix} \deg(v1) & -a_{12} & -a_{13} & \ldots & -a_{1,n} \\ -a_{21} & \deg(v_2) & -a_{23} & \ldots & -a_{2,n} \\ & & & & \\ \ldots & & & & \\ & & & & \\ -a_{n,1} & -a_{n,2} & \ldots & \ldots & \deg(v_n) \end{bmatrix} \tag{33}$$

By extending the Matrix Tree theorem to continuous-valued A and letting the weights W_{uv} play the role of a_{uv}, one can prove

Theorem 5 *Let $P(E)$ be a distribution over tree structures defined by*

$$P(E) \;\propto\; W_0 \prod_{uv \in E} W_{uv} \tag{34}$$

Then its normalization constant Z is equal to

$$Z \;=\; W_0 \sum_E \prod_{uv \in E} W_{uv} \;=\; W_0 |Q(W)| \tag{35}$$

with $Q(W)$ being the $(n-1) \times (n-1)$ matrix

$$Q_{uv}(W) \;=\; Q_{vu}(W) \;=\; \begin{cases} -W_{uv} & 1 \leq u < v \leq n-1 \\ \sum_{v'=1}^{n} W_{v'v} & 1 \leq u = v \leq n-1 \end{cases} \tag{36}$$

This shows that summing over the distribution of all trees, when this distribution factors according to the trees' edges, can be done in closed form by computing the value of a order $n-1$ determinant, operation that involves $\mathcal{O}(n^3)$ operations.

To optimize the Lagrange multipliers, we must compute derivatives of $J(\lambda)$ or, equivalently, derivates of the log-partition functions with respect to λ. It is well known that such derivatives lead to averages with respect to the distribution in question (for details see Appendix A). In our case, for example,

$$\frac{\partial \log Z_s}{\partial \lambda_t} \;=\; s y_t \; < \log T(X_t | E_s, \theta^s) >_{P(E_s)} \;=\; s y^t \left[w_0^s(X^t) + \sum_{u \neq v} w_{uv}^s(X^t) W_{uv}^s M_{uv}^s \right] \tag{37}$$

where M^s is a linear function of $Q^{-1}(W^s)$ given in Appendix A. Inverting the matrix $Q(W)$ is $\mathcal{O}(n^3)$ and this operation can be done once before the summations in equations (37). Thus, computing the derivatives of the normalization constant w.r.t all λ_t takes $\mathcal{O}(n^3 + n^2 T)$ operations and $\mathcal{O}(n^2)$ extra space.

[5] Note that A as a whole is a singular matrix.

Finally, to obtain the decision rule for any new example X we must compute averages of the log-likelihood ratio with respect to the (marginal) MRE distribution $P(E_1)P(E_{-1})$:

$$\hat{y} = \operatorname{sgn}\left\{ \textstyle\sum_{E_1,E_{-1}} P(E_1)P(E_{-1}) \log \frac{T(X|E_1,\theta^1)}{T(X|E_{-1},\theta^{-1})} \right\} \tag{38}$$

$$= \operatorname{sgn}\left\{ w_0^1(X) - w_0^{-1}(X) \;\; + <\sum_{uv\in E_1} w_{uv}^1(X)>_{P(E_1)} - <\sum_{uv\in E_{-1}} w_{uv}^{-1}(X)>_{P(E_{-1})} \right\} \tag{39}$$

where we have omitted a possible bias term b. The required averages can be computed analogously to Eq. (37) yielding e.g.

$$<\sum_{uv\in E_1} w_{uv}^1(X)>_{P(E_1)} = \sum_{u\neq v} w_{uv}^1(X) W_{uv} M_{uv}^1 \tag{40}$$

where M_{uv}^1 is the same matrix as in Eq. (37) and has been already computed in the last step of the training algorithm. Classifying a new data point therefore requires only roughly $\mathcal{O}(n^2)$ operations.

3.5 MRE distributions over tree structures and parameters

Here we describe briefly how to find the MRE distribution over both structures and parameters, i.e., $P(E_1, \theta^1, E_{-1}, \theta^{-1})$. We assume a factored prior $P_0(\theta^1)P_0(\theta^{-1})$ over the parameters and as before a uniform prior over the structures. In addition to the parameter independence and modularity assumptions used earlier, we must assume that the priors $P_0(\theta^s)$, $s = 1, -1$ are *likelihood equivalent* (i.e. they assign the same value to models having the same likelihood for all data sets). In this case, the priors over parameters are forced to be Dirichlet [9] and defined in terms of a set of *equivalent marginal counts* $\tilde{N}_{uv}^s(x_u, x_v)$ satisfying

$$\sum_{x_u} \tilde{N}_{uv}^s(x_u, x_v) = \tilde{N}_v^s(x_v) \quad \sum_{x_v} \tilde{N}_{uv}^s(x_u, x_v) = \tilde{N}_u^s(x_u) \quad \sum_{x_u x_v} \tilde{N}_{uv}^s(x_u, x_v) = \tilde{N}^s \tag{41}$$

Because the prior over parameters is independent of the structure, the MRE distribution factorizes as

$$P(E_s, \theta^s) = \frac{1}{Z_s} P_0(\theta^s) e^{\sum_t s\lambda_t y_t \log T(X_t|E_s,\theta^s)} \tag{42}$$

To evaluate the partition function Z_s, the parameters θ^s can be analytically integrated out before the summation over structures. The resulting marginal distribution over tree structures is similar to equation (35)

$$P(E_s) = \frac{W_0^s}{Z_s} \sum_E \prod_{uv\in E} W_{uv}^s \tag{43}$$

with the factors W^s are now functions of both λ and Dirichlet distribution parameters $\tilde{N}^s$ (see appendix B for exact expression).

The classification rule is also similar in form to equation (39) with the terms w^s depending on λ, the data, and the equivalent counts as described in Appendix B.

3.6 General Bayes nets

A Bayes net with given structure can be parametrized by the set of conditional distributions $P(v|\mathrm{pa}(v) = x_{\mathrm{pa}(v)})$ of a variable given a configuration of its parents. A discriminative MRE solution can be found for the parameter distribution $P(\theta^1, \theta^{-1})$ assuming complete observations. Finding the MRE distribution over structures is, however, unlikely to be feasible for other than trees (c.f. [5]).

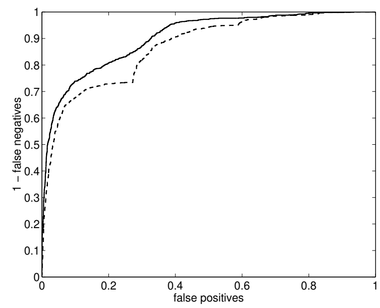


Figure 7: ROC curves for the ME discriminative classifier (full line) and the ML classifier (dashed line) for the splice junction classification problem. The minimum test errors are 12.4% and 14% respectively.

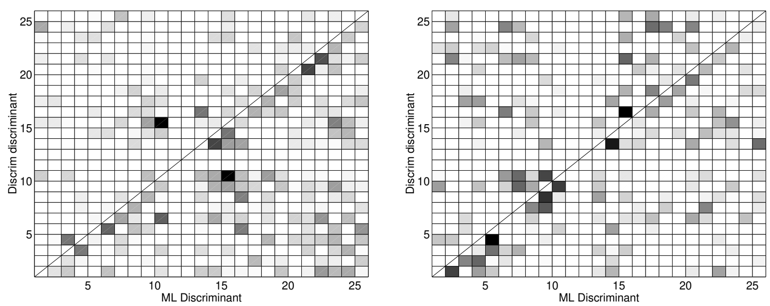


Figure 8: Logarithmic weights w_{uv} versus mutual informations I_{uv} for class 1 (a) respective -1 (b). The square in position uv, $u < v$ represents w_{uv} while its symmetric, vu represents I_{uv}. Larger values appear more back in the figures.

3.7 Experiments

We tested our model in the fixed parameter version on the detection of DNA splice sites and compared its performance to the performance of a classifier using a Maximum Likelihood (ML) tree for each class. In both cases, the tree parameters θ were the ML parameters for the corresponding class (empirical class-conditional marginals).

The domain consists of 25 variables representing sites around a (hypothetic) splice junction. The test set had 400 examples split equally between the two classes; the training set consisted of 4724 examples, about a fourth being positives ones. For simplicity, we used a fixed margin $\gamma = 4$, the largest value that allowed perfect class separation. The number of λ's that are nonzero in this example is 61 (out of 400) suggesting a performance level of about %15 according to Lemma 1. The ROC curves for the two classifiers are compared in figure 7. MRE distribution over tree structures is superior to a pair maximum Likelihood trees, although the parameter values are identical. The test set error is 14.0% for the ML classifier and 12.3% for the MRE method. The training error is 0.5% for the ML classifier and zero for the discriminative one indicating that the MRE method is resistant to overfitting.

Figure 8 compares the "edge weights" for the two classifiers. These edge weights reflect the preferences assigned to tree structures in the MRE distribution or in the (single) class-conditional maximum likelihood (ML) tree. Since the estimation criterion differs in the two cases, the most likely tree in the MRE solution does not in general equal the ML tree structure. Figure 8a) displays $w^1_{uv} = \log(W^1_{uv})$ factors corresponding to each edge uv in the MRE distribution for class 1 as well as the respective mutual information values I^1_{uv}. Since both matrices are symmetric, one can display both sets of values in a 25 by 25 square: the upper left half represents the ME weights whereas the lower right half of the square shows the mutual information. Figure 8,b shows the same results for class -1. Note that summing w^1_{uv} or I^1_{uv} across the edges of a particular tree pertains directly to the log-probability of the tree and thus the comparison is meaningful [6].

The figure shows that there are relatively few edges with large weights on both sides of the diagonal. This is particularly relevant for the discriminative model of the positive examples, since it shows that the MRE distribution decays rapidly around its peak. The maximum W^1_{uv} is more than 10^3 times the next largest value, clearly separating edges that are discriminative and those whose inclusion or exclusion has little effect on discrimination. This contrast is understandably less pronounced for the negative examples that represent a diverse collection of spurious splice sites.

A second important remark is that neither figure 8,a nor 8,b are symmetric w.r.t the diagonal. In other words, not all pairs of variables that exhibit high mutual information are also discriminative. Note for example that the subdiagonal band showing that adjacent variables are informative of each other is almost completely effaced under discriminative training. Our method brings out the discriminative structure of the data, which is different from its structure as a density estimator.

4 Anomaly detection

In anomaly detection we are given a set of training examples representing only one class, the "typical" examples. We attempt to capture regularities among the examples to be able to recognize unlikely members of this class. Estimating a probability distribution $P(X|\theta)$ on the basis of the training set $\{X_1, \ldots, X_T\}$ via the standard maximum likelihood (or analogous) criterion is not appropriate since there is no reason to further increase the probability of those examples that are already well captured by the model. A more relevant measure involves the level sets

$$\mathcal{X}_\gamma = \{\, X \in \mathcal{X} : \log P(X|\theta) \geq \gamma \,\} \tag{44}$$

These level sets are used in deciding the class membership, even in the context of ML parameter estimation. We therefore estimate the parameters θ to optimize an appropriate level set. As before, we cast this problem as MRE:

Definition 3 *Given a probability model $P(X|\theta)$, $\theta \in \Theta$, a set of training examples $\{X_1, \ldots, X_T\}$, a set of margin variables $\gamma = [\gamma_1, \ldots, \gamma_T]$, and a prior distribution $P_0(\theta, \gamma)$ we find the MRE distribution $P(\theta, \gamma)$ such that minimizes $D(P\|P_0)$ subject to the constraints*

$$\int P(\theta, \gamma)\, [\, \log P(X_t|\theta) - \gamma_t \,]\, d\theta d\gamma \geq 0 \tag{45}$$

for all $t = 1, \ldots, T$.

Note that this is again a MRE projection problem whose solution can be obtained as before. The choice of $P_0(\gamma)$ in $P_0(\theta,\gamma) = P_0(\theta)P_0(\gamma)$ is not as straightforward as before since each margin γ_t needs to be close to achievable log-probabilities. We can nevertheless easily find a reasonable choice e.g. by relating the prior mean of γ_t to some $\alpha-$percentile of the training set log-probabilities generated through ML or other standard parameter estimation criterion. Denote the resulting value by l_α and define the prior $P_0(\gamma_t)$ as $P_0(\gamma_t) = c\, e^{-c\,(l_\alpha - \gamma_t)}$ for $\gamma_t \leq l_\alpha$. In this case the prior mean of γ_t is $l_\alpha - 1/c$.

[6]The comparison is done upto a scaling factor and an additive constant.

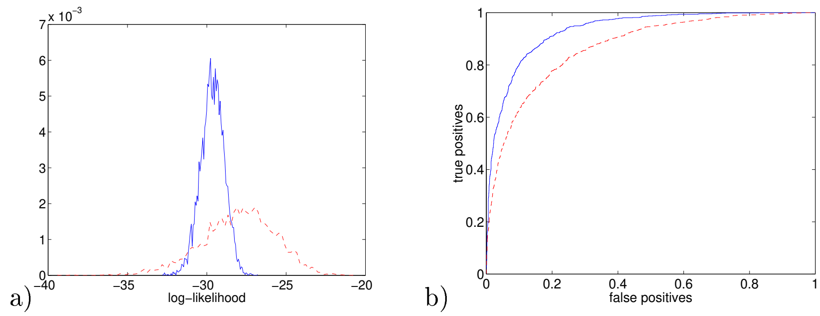


Figure 9: a) Distribution of training set log-likelihoods for the MRE model (solid line) or the Bayes model (dashed-line). b) ROC curve for the two models on an independent test set.

We have verified experimentally for a simple product distribution that this choice of prior together with the MRE framework leads to a real improvement over standard (Bayesian) approach. Figure 9 illustrates the benefit of the MRE approach for discriminating between true and spurious splice sites. The examples were fixed length DNA sequences (length 25) and we used the following product distribution of simple multinomials:

$$P(X|\theta) = \prod_{i=1}^{25} P_i(x_i|\theta_i) = \prod_{i=1}^{25} \theta_{x_i|i} \tag{46}$$

where $X = \{x_1, \ldots, x_{25}\}$, $x_i \in \{A, C, T, G\}$, and $\sum_{x_i} \theta_{x_i|i} = 1$. The model parameters $\{\theta_{x_i|i}\}$ were estimated on the basis of only true examples (7000). The estimation criterion was either Bayesian with an independent Dirichlet prior over each component distribution $\{\theta_{\cdot|i}\}$ or through the relative entropy projection method with the same prior. Figure 9a) indicates, as expected, that the training set log-likelihoods from the MRE method are more uniform and without the long tails[7]. This difference leads to improved anomaly detection as shown by the ROC curve in Figure 9b). The test set consisted of 1192 true splice sites and 3532 spurious ones.

We expect the effect to be more striking in the context of more sophisticated models such as HMMs that may otherwise easily capture spurious regularities in the data. In the next section we describe how such models can be used efficiently within the MRE framework.

4.1 Extension to latent variable models

In the presence of latent variables (missing information) we can no longer use the above formulation directly. This arises because $\log P(X_t|\theta)$ does not decompose into a sum of simple components. We can, however, achieve an efficient lower bound solution. If we let X_h be the set of latent variables, we can resort to the following variational lower bound:

$$\log P(X_t|\theta) \geq \sum_{X_h} Q_t(X_h) \log P(X_t, X_h|\theta) + H(Q_t) \tag{47}$$

where $H(Q_t)$ is the entropy of the Q_t distribution. A separate transformation has to be introduced for each training example. Note that the lower bound is reasonable in this context since the objective

[7]To compute these log-likelihoods from the MRE method, we used the MRE solution as the posterior distribution over the parameters. This is suboptimal for the MRE method given that the criterion is slightly different but suffices here for the purposes of illustration. An analogous figure with minor differences could be computed on the basis of $\int P(\theta) \log P(X|\theta) d\theta$ for the two methods. In this case, the figure would be suboptimal for the Bayesian approach.

is to guarantee that all (or most) training examples have likelihoods above some margin threshold. Whenever the lower bound exceeds the threshold, so does the original likelihood.

The MRE distribution $P(\theta, \gamma)$ is obtained under the following constraints:

$$\int P(\theta,\gamma)\left[\sum_{X_h} Q_t(X_h)\log P(X_t, X_h|\theta) \; - \gamma_t\right] d\theta + H(Q_t) \geq 0 \tag{48}$$

which are of the same form (linear) as before. Note that we have made an additional assumption that $Q_t(X_h)$ is functionally independent of the parameters θ. This assumption guarantees that the MRE distribution $P(\theta, \gamma)$ can be computed efficiently for a large class of probability models such as mixture models and HMMs. The loss in accuracy due to this simplifying assumption vanishes whenever the (marginal) MRE distribution $P(\theta)$ becomes peaked. In principle, this means that we can always find the single most discriminative setting of the parameters even with the variational bound. Roughly speaking, we incur a loss only relative to the exact MRE approach.

The overall solution to the MRE problem is no longer unique, however, but we can find a locally optimal solution iteratively as follows:

Step 1. Fix $\{Q_t(X_h)\}$ and find the MRE distribution $P(\theta, \gamma)$ as before

Step 2. Fix $P(\theta, \gamma)$ and let

$$Q_t(X_h) \propto \exp\left\{ \int P(\theta)\log P(X_t, X_h|\theta)d\theta \right\} \tag{49}$$

Both steps can be computed efficiently for a large class of models such as HMMs assuming the prior $P_0(\theta)$ is Dirichlet and factorizes across the parameters. More generally, the prior should be the conjugate prior satisfying the parameter independence assumption of [6] (see also [9]).

The iterative algorithm actually converges in the sense defined by the following theorem:

Theorem 6 *If we let $P^{(n)}(\theta, \gamma)$ be the MRE distribution after n steps of the iterative algorithm described above, then*

$$D(P^{(1)}\|P_0) \geq D(P^{(2)}\|P_0) \geq \ldots \geq D(P^{(n)}\|P_0) \tag{50}$$

The theorem is easy to understand as follows: each time we optimize any of the $Q_t(X_h)$ distributions, we maximize the associated lower bound. This maximization relaxes the corresponding constraint on the MRE distribution and allows the relative entropy to be decreased.

5 Uncertain or incompletely labeled examples

Examples with uncertain labels are hard to deal with in any standard discriminative classification method, probabilistic or not. Note the difference between labels that are inherently stochastic and those that are predictable but merely missing (the case considered here). Uncertain labels can be handled in a principled way within the maximum entropy formalism: let $y = \{y_1, \ldots, y_T\}$ be a set of binary variables corresponding to the labels for the training examples. We can define a prior uncertainty over the labels by specifying $P_0(y)$; for simplicity, we can take this to be a product distribution

$$P_0(y) = \prod_t P_{t,0}(y_t) \tag{51}$$

where a different level of uncertainty can be assigned to each example. We may, for example, set $P_{t,0}(y_t) = 1$ whenever y_t is observed and $P_{t,0}(y_t) = 0.5$ if the label is missing. The MRE solution is found by calculating the relative entropy projection from the overall prior distribution $P_0(\Theta, \gamma, y) = P_0(\Theta)P_0(\gamma)P_0(y)$ to the admissible set of distributions $\mathcal{P}$ (no longer directly function of the labels) that are consistent with the constraints:

$$\sum_y \int_{\Theta,\gamma} P(\Theta, \gamma, y) \; [\, y_t \mathcal{L}(X_t, \Theta) - \gamma_t \,] \; d\Theta \, d\gamma \geq 0 \tag{52}$$

for all $t = 1, \ldots, T$. The prior distribution $P_0(\gamma)$ in this formulation encourages decision rules that achieve large classification margins for the examples (most of the probability mass is assigned to values $\gamma_t \geq 0$). This preference towards large margins creates dependencies between the (a priori) unknown labels and the parameters Θ of the discriminant function. Consequently, even unlabeled examples will contribute to the (marginal) MRE distribution $P(\Theta)$ that specifies the decision rule. We may alternatively view the MRE formulation as a *transduction* algorithm [22] whose objective is to determine the class labels for a set of unlabeled training examples.

While this provides a principled framework for dealing with uncertain or partially labeled examples, the MRE solution itself is not in general feasible to obtain. For example, in the context of support vector machines (for an alternative approach see [2]), the MRE distribution over the labels will be (roughly speaking) a Boltzmann machine and therefore not manageable in general via exact calculations. We can nevertheless employ efficient approximate methods to obtain an iterative algorithm for self-consistent probabilistic assignment of the uncertain labels.

5.1 Feasible approximation

To be able to deal with uncertain labels in a feasible way, we solve instead the following MRE problem with additional constraints:

Definition 4 *Given a parametric discriminant function $\mathcal{L}(X, \Theta)$, a set of margin variables $\gamma = [\gamma_1, \ldots, \gamma_T]$, a set of class variables $y = [y_1, \ldots, y_T]$, and a prior distribution*

$$P_0(\Theta, \gamma, y) = P_0(\Theta) \left[\prod_t P_0(\gamma_t) \right] \left[\prod_t P_{0,t}(y_t) \right] \tag{53}$$

we find a constrained MRE distribution $P(\theta, \gamma, y)$ of the form $P(\theta, \gamma)P(y)$ that minimizes $D(P \| P_0)$ subject to the constraints

$$\sum_y \int_{\Theta,\gamma} P(\Theta, \gamma)P(y) \; [\, y_t \mathcal{L}(X_t, \Theta) - \gamma_t \,] \; d\Theta \, d\gamma \geq 0 \tag{54}$$

for all $t = 1, \ldots, T$.

We may view this as a type of mean field approximate since the MRE distribution is forced to factorize to make the problem tractable. The solution is no longer unique but can be obtained through the following two-stage iterative algorithm:

Step 1. Fix $P(y)$ and let $p_t = \sum_y P(y) y_t$. We find $P(\Theta, \gamma)$ as the MRE solution subject to the constraints

$$\int_{\Theta,\gamma} P(\Theta, \gamma) \; [\, p_t \mathcal{L}(X_t, \Theta) - \gamma_t \,] \; d\Theta \, d\gamma \geq 0 \tag{55}$$

Note that since the prior factorizes across $\{\Theta, \gamma\}$ the MRE solution factorizes as well, i.e., $P(\Theta, \gamma) = P(\Theta)P(\gamma)$.

Step 2. Fix the marginal $P(\Theta)$ obtained in the previous step and find the MRE solution $P'(y,\gamma)$ subject to

$$\sum_y \int P'(y,\gamma) \left\{ \int_\Theta P(\Theta) \left[y_t \mathcal{L}(X_t,\Theta) - \gamma_t \right] d\Theta \right\} d\gamma \geq 0 \tag{56}$$

for all t. Update $P(y) \leftarrow (1-\epsilon)P(y) + \epsilon P'(y)$ or simply set $p_t \leftarrow (1-\epsilon)p_t + \epsilon p'_t$ where $p'_t = \sum_y P'(y) y_t$.

The fact that we include $P(\gamma)$ also in the second step is necessary since any adjustments to the labels must be compensated by an increased margin. The distribution $P(y)$ is updated via relaxation to ensure a more controlled adjustment of the labels; any large change in $P(y)$ is likely to induce a significant subsequent modification to the solution of the first step. Although the iterative algorithm remains stable even if larger changes are made, we believe the relaxation update leads to better local optima. Moreover, since the admissible set is convex and because the minimization objective (relative entropy) is also convex, the relaxation update always yields a change in the appropriate direction. The solution to either step is well defined and can be obtained in closed form assuming the problem is tractable when we have complete information about the labels. The iterative algorithm is well-behaved in the sense of the following theorem:

Theorem 7 *Let $P^{(n)}(\Theta,\gamma,y) = P^{(n)}(\Theta,\gamma)P^{(n)}(y)$ be the constrained MRE solution after n iterations. Then for all $0 \leq \epsilon \leq 1$, where ϵ is the step size used in the algorithm, we have*

$$D(P^{(1)} \| P_0) \geq D(P^{(2)} \| P_0) \geq \ldots \geq D(P^{(n)} \| P_0) \tag{57}$$

The result holds also after either step of the two-stage iterative algorithm.

5.2 Example: support vector machines

Here we provide a preliminary numeral assessment of how the above algorithm is able to make use of unlabeled examples in the context of predicting DNA splice sites with support vector machines. A detailed formulation of the algorithm for SVMs can be found in Appendix C. We generated three training sets of examples corresponding to whether 1) all the labels were known, 2) labels were provided only for about 10% randomly chosen examples and the remaining 90% were unlabeled but available, and 3) only the 10% labeled examples were used for training. The full training set in this case consisted of 500 true DNA splice sites and 500 spurious ones (false examples). The examples were fixed length (25) strings of DNA letters (A,C,T,G) which were translated into bit vectors using a four bit encoding (e.g. $A \rightarrow [1000]$). Figure 10 gives ROC curves based on an independent test set (1192 true examples and 3532 false examples) for SVMs trained with one of the three training sets. Note that when the training set is fully labeled the algorithm reduces to the standard formulation. The figures show that even the approximate formulation[8] is able to reap most of the benefit from the unlabeled examples. The finding is also robust against the choice of the kernel function as is seen by comparing Figure 10a) and 10b). The findings are preliminary.

6 Discussion

We have presented a general approach to discriminative training of model parameters, structures, or parametric discriminant functions. The formalism is based on the minimum relative entropy principle reducing all calculations to relative entropy projections. Quite remarkably, we can efficiently

[8]In our experiments, $\epsilon = 0.1$ and the iterative algorithm was run for 10 iterations. The benefit may vary as a function of ϵ and the number of iterations, particularly if ϵ is too large. The prior probability $P_0(y) = \prod_t P_{0,t}(y_t)$ over the labels were set to 0 or 1 when the label for y_t was observed and to 0.5 for the unlabeled ones.

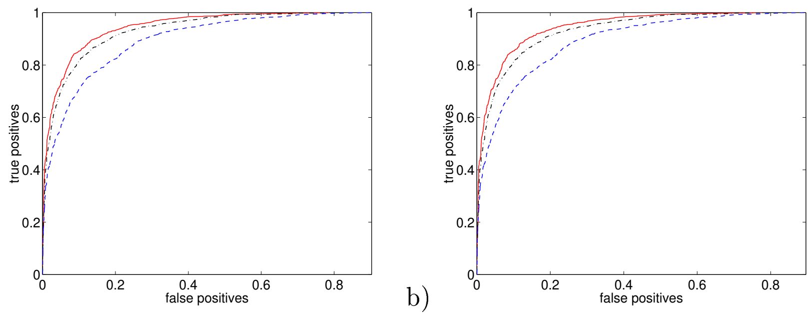


Figure 10: a) test set ROC curves based on a training set with fully labeled examples (solid line), 90% unlabeled and 10% labeled (dot-dashed), only the 10% labeled examples (dashed). In a) a linear kernel was used and in b) a Gaussian kernel.

and exactly compute the best discriminative distribution over tree structures within this framework. The MRE idea gives, in addition, a natural discriminative formulation of anomaly detection problems or classification problems involving partially labeled examples. Efficient algorithms were also given to exploit such formulations.

Acknowledgments

The authors would like to thank David Haussler for useful comments and Sayan Mukherjee and Olivier Chapelle for pointing out errors in an earlier version of this manuscript.

References

[1] Amari S-I. (1995). Information geometry of the EM and em algorithms for neural networks.

[2] Bennett K. and Demiriz A. (1998). Semi-supervised support vector machines. NIPS 11.

[3] Barber D. and Williams C. (1997). Gaussian processes for bayesian classification via hybrid monte carlo. NIPS 9.

[4] Blum A. and Mitchell T. (1998). Combining Labeled and Unlabeled Data with Co-Training . In *Proceedings of the 11th Annual Conference on Computational Learning Theory.*

[5] Chickering D., Geiger D. and Heckerman D. (1995). Learning Bayesian networks: Search methods and experimental results.

[6] Cooper G. and Herskovitz E. (1992). A Bayesian method for the induction of probabilistic networks from data. *Machine learning* 9: 309–347.

[7] Cover T. and Thomas J. (1991). *Elements of information theory.* John Wiley & Sons, Inc.

[8] Freund Y. and Schapire R. (1997). A decision theoretical generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119-139.

[9] Heckerman D., Geiger D. and Chickering D. (1995). Learning Bayesian networks: the combination of knowledge and statistical data. *Machine Learning.*

[10] Heckerman D. and Meek C. (1997). Models and Selection Criteria for Regression and Classification. Technical Report MSR-TR-97-08, Microsoft Research.

[11] Jaakkola T. and Haussler D. (1998). Exploiting generative models in discriminative classifiers. NIPS 11.

[12] Jaakkola T. and Haussler D. (1998). Probabilistic kernel regression models. In *Proceedings of The Seventh International Workshop on Artificial Intelligence and Statistics.*

[13] Jebara T. and Pentland A. (1998). Maximum conditional likelihood via bound maximization and the CEM algorithm. NIPS 11.

[14] Kapur J. (1989). *Maximum entropy models in science and engineering.* John Wiley & Sons.

[15] Levin and Tribus (eds.) (1978). *The maximum entropy formalism.* Proceedings of the Maximum entropy formalism conference, MIT.

[16] Meilă M. and Jordan M. (1998). Estimating dependency structure as a hidden variable. NIPS 11.

[17] Minka T.P. (1998). Inferring a gaussian distribution. http://www.media.mit.edu/ tpminka/papers/minka-gaussian.ps.gz

[18] Nigam K., McCallum A., Thrun S., and Mitchell T. (1999). Text classification from labeled and unlabeled examples. To appear in *Machine Learning*.

[19] Ripley B.D. (1994). Flexible non-linear approaches to classification. In V. Cherkassy, J.H. Friedman, and H. Wechsler (Eds.), ¿From Statistics to Neural Networks, pp. 105-126. Springer.

[20] Schapire R., Freund Y., Bartlett P. and Lee W. S. (1998). Boosting the margin: A new explanation for the effectiveness of voting methods *The Annals of Statistics* **26**(5):1651-1686.

[21] Vapnik V. (1995). *The nature of statistical learning theory.* Springer-Verlag.

[22] Vapnik V. (1998). *Statistical learning theory.* John Wiley & Sons.

[23] West D. (1996). *Introduction to graph theory.* Prentice Hall.

[24] Wolberg W. and Mangasarian O (1990). Multisurface method of pattern separation for medical diagnosis applied to breast cytology, *Proceedings of the National Academy of Sciences*, U.S.A., Vol. 87.

[25] Zhang J. (1992). Selecting typical instances in instance-based learning. In *Proceedings of the Ninth International Machine Learning Conference.*

A Computing averages under a factored distribution over tree structures

Lemma 4 *If $P(E)$ is given by equation (34) and f, g are functions of E additive in the edges (i.e. $f(E) = \sum_{uv \in E} f_{uv}$) then*

$$
\begin{aligned}
< f(E) >_P &= \frac{1}{Z} \left. \frac{\partial |Q(We^{\alpha f})|}{\partial \alpha} \right|_{\alpha=0} && (58)\\
< f(E)g(E) >_P &= \frac{1}{Z} \left. \frac{\partial^2 |Q(We^{\alpha f + \beta g})|}{\partial \alpha \partial \beta} \right|_{\alpha=\beta=0} && (59)
\end{aligned}
$$

This lemma can be easily proved by equating $|Q(We^{\alpha f})|$ with its definition (36) and then taking derivatives of both sides. Then, remembering that for any matrix A with elements A_{ij}

$$\frac{\partial |A|}{\partial A_{ij}} \;=\; |A|(A^{-1})_{ij} \tag{60}$$

one obtains, after conveniently grouping the terms, the result of Lemma 5:

Lemma 5 *Let $P(E)$ and Q be given by equations (34) and (36) respectively, M be a symmetric matrix with 0 diagonal defined by*

$$M_{uv} \;=\; M_{vu} \;=\; \begin{cases} \frac{1}{2}[(Q^{-1})_{uu} + (Q^{-1})_{vv} - 2(Q^{-1})_{uv})], & u, v < n \\ \frac{1}{2}(Q^{-1})_{vn} & v < u = n \end{cases} \tag{61}$$

and f a function of the structure E satisfying $f(E) = \sum_{uv \in E} f_{uv}$. Then the average of f under P is

$$< f(E) >_P = \sum_E P(E) f(E) \;=\; \sum_{u,v=1}^{n} f_{uv} W_{uv} M_{uv}. \tag{62}$$

B Integrating over the parameters $P(E_s, \theta^s)$

Let us define

$$N^s_{uv}(x_u, x_v) \;=\; \sum_{t: v = x_v, u = x_u} s\lambda_t y_t \qquad N^s_{uv}(x_v) \;=\; \sum_{t: v = x_v} s\lambda_t y_t \tag{63}$$

$$\kappa^s_{uv} \;=\; \prod_{x_u} \prod_{x_v} \frac{\Gamma(N^s_{uv}(x_u, x_v) + \tilde{N}^s_{uv}(x_u, x_v))}{\Gamma(\tilde{N}^s_{uv}(x_u x_v))} \tag{64}$$

$$\kappa^s_v \;=\; \prod_{x_v} \frac{\Gamma(N^s_v(x_v) + \tilde{N}^s_v(x_v))}{\Gamma(\tilde{N}^s_v(x_v))} \tag{65}$$

With these notations we can express W^s_{uv} and W^s_0 in equation (43) as

$$W^s_{uv} \;=\; \frac{\kappa^s_{uv}}{\kappa^s_u \kappa^s_v} \quad \text{and} \quad W^s_0 \;=\; \frac{\Gamma(\tilde{N}^s)}{\Gamma(N^s + \tilde{N}^s)} \prod_{v \in V} \kappa^s_v \tag{66}$$

In the above, $\Gamma()$ denotes Euler's Gamma function. Note that the "counts" N^s_{uv} can be either positive or negative, so that the variables κ may not be defined for arbitrary values of λ. All the above expressions exist, however, for $\lambda = 0$; in this case $W^s_{uv} = W^s_0 = 1$.

The classification rule is given by equation (39) with $w^s_{uv}(X), w^s_0(X)$ redefined as

$$w^s_{uv}(X) \;=\; \Psi[N^s_{uv}(x_u x_v) + \tilde{N}^s_{uv}(x_u x_v)] - \Psi[N^s_v(x_v) + \tilde{N}^s_v(x_v)] - \Psi[N^s_u(x_u) + \tilde{N}^s_u(x_u)] \tag{67}$$

$$w^s_0(X) \;=\; \sum_{v \in V} \Psi[N^s_v(x_v) + \tilde{N}^s_v(x_v)] - \Psi[N^s + \tilde{N}^s] \tag{68}$$

with Ψ representing the derivative of the log-Gamma function:

$$\Psi(z) \;=\; \frac{d}{dz} \log \Gamma(z) \tag{69}$$

Note the similarity with the fixed parameter case: the classification rule is still an average of a log-likelihood difference; the Ψ functions arise from averaging the log-likelihood under the MRE distribution of the θ parameters.

C Uncertain labels and support vector machines

We provide here more details about the two step feasible algorithm for dealing with partially labeled examples in the context of support vector machines. We start by defining the prior distribution over all the parameters as

$$P_0(\theta, b, \gamma, y) = P_0(\theta)P_0(b)P_0(\gamma)P_0(y) \tag{70}$$

where $P_0(\theta)$ is $\mathcal{N}(0, I)$ and $P_0(b)$ approaches a non-informative prior. By the non-informative prior we mean here a limit of $P_0(b|k) = \mathcal{N}(0, I \cdot k)$ as $k \to \infty$. The prior over the labels is assumed to factorize across the examples, i.e.,

$$P_0(y) = \prod_t P_{0,t}(y_t) \tag{71}$$

where, for example, we can set each $P_{0,t}(y_t) = 1$ whenever the corresponding label y_t is known and $P_{0,t}(y_t) = 0.5, y_t = \pm 1$ for all unlabeled examples. We use here $P_0(\gamma)$ from eq. (9); the alternatives were discussed in the text.

Let now $p_t = \sum_y P_0(y) y_t = \sum_{y_t} P_{0,t}(y_t) y_t$, where p_t is the mean value of the label. With these initializations, the two step algorithm is given as follows:

Step 1. We fix $\{p_t\}$ and find the MRE solution for $P(\theta, b, \gamma)$. Based on Lemma 3 $P(\theta, \gamma)$ and $P(b)$ can be found separately. For $P(\theta, \gamma)$ the the Lagrange multipliers are obtained by maximizing (analogously to Theorem 2):

$$J_{\theta,\gamma}(\lambda) = \sum_t [\, \lambda_t + \log(1 - \lambda_t/c)\,] - \frac{1}{2} \sum_{t,t'} \lambda_t \lambda_{t'} p_t p_{t'} (X_t^T X_{t'}) \tag{72}$$

subject to the constraint that $\sum_t \lambda_t p_t = 0$. This is no more difficult to solve than the original SVM optimization problem with hard labels.

As for the bias term b, we only need its mean relative to the MRE solution, i.e., $\bar{b} = \int P(b) b \, db$. This can be computed as the limit of the means corresponding to proper priors $P_0(b|k)$ (each MRE solution $P(b|k)$ based on $P_0(b|k)$ is a Gaussian with a well-defined mean). We omit the algebra and instead provide the answer in terms of the following averages:

$$\bar{L}_t = \int P(\theta)\,(\theta^T X_t)\, d\theta = \sum_{t'} \lambda_{t'} p_{t'} (X_t^T X_{t'}) \tag{73}$$

$$\bar{\gamma}_t = \int P(\gamma)\, \gamma_t \, d\gamma = 1 - \frac{1}{c - \lambda_t} \tag{74}$$

The desired mean $\bar{b}$ is now given by

$$\bar{b} = \arg\max_b \Big\{ \min_t (\, p_t(\bar{L}_t + b) - \bar{\gamma}_t \,) \Big\} \tag{75}$$

This setting optimizes the most critical constraints of eq. (55). In other words, $\bar{b}$ maximizes the minimum of the left hand sides of eq. (55).

Step 2. To update the MRE distribution over the labels, we fix $P(\theta, b)$ and find $P'(y, \gamma)$ subject to

$$\begin{aligned} &\sum_y \int P'(y,\gamma) \int_{\theta,b} P(\theta, b) \, [\, y_t(\theta^T X_t + b) - \gamma_t \,] \, d\theta db d\gamma \\ &\quad = \sum_y \int P'(y,\gamma) \, [\, y_t(\bar{L}_t + \bar{b}) - \gamma_t \,] \, d\gamma \geq 0 \end{aligned} \tag{76}$$

Analogously to the first step, the Lagrange multipliers are found by maximizing the corresponding $-logZ$ (algebra omitted):

$$J_{y,\gamma}(\lambda') = \sum_t \left\{ \lambda'_t + \log(1 - \lambda'_t/c) - \log \sum_{y_t=\pm 1} P_{0,t}(y_t) e^{\, y_t \lambda'_t (\bar{L}_t + \bar{b})} \right\} \tag{77}$$

Note that the Lagrange multipliers here are not tied and can be optimized independently for each t. This happens because we have assumed that the prior distribution factorizes across the examples and because the discriminant function does not tie the variables together. Each of the one dimensional convex optimization problems are readily solved by any standard methods (e.g. Newton-Raphson). The resulting MRE distribution over the labels, $P'(y)$ is given by

$$P'(y) = \prod_t P'_t(y_t) \tag{78}$$

where

$$P'_t(y_t) = \frac{1}{Z_t} P_{0,t}(y_t)\, e^{\, y_t \lambda'_t (\bar{L}_t + \bar{b})} \tag{79}$$

We can easily compute $p'_t = \sum_{y_t} P'_t(y_t) y_t$ from this result. Finally, the updates

$$p_t \leftarrow (1 - \epsilon) p_t + \epsilon p'_t \tag{80}$$

complete the second step.

The decision rule for a new example X is given by

$$\hat{y} = \text{sign} \left(\sum_t \lambda_t p_t (X_t^T X) + \bar{b} \right) \tag{81}$$

where $\{\lambda_t\}$ and $\bar{b}$ are the solutions to the first step of the iterative algorithm.