# 6.034 QUIZ 2

Fall 2000

| | |
|---|---|
| Name | |
| E-mail | |
| TA | |
| Recitation Instructor | |

| Problem Number | Score |
|---|---|
| Problem 1 | |
| Problem 2 | |
| Problem 3 | |
| Total | |

## Problem 1: Miscellaneous (30 points)

This problem is first because other problems were judged to take three to four times as long. Circle the **single** phrase that **best** completes the following fragments. All multiple votes will be rejected.

Progressive deepening, also known as iterative deepening, works well for games because:

- Alpha-beta allows you to go twice as deep in a given game tree.
- The branching factor varies from layer to layer.
- Almost none of the nodes in a game tree of a given depth are in the final layer.
- Almost all the nodes in a game tree of a given depth are in the final layer.
- All of the above.
- None of the above.

Alpha-beta:

- Doubles the speed of minimax.
- Is slower than minimax.
- Is incompatible with minimax.
- Is incompatible with progressive deepening
- All of the above.
- None of the above.

The topological sorting algorithm was developed to:

- Improve run time speed.
- Ensure precedence is determined by the up-to-join principle.
- Deal with loops in the inheritance tree.
- Honor ordering principles.
- All of the above.
- None of the above.

Frames have been used to:

- Enable sentence understanding.
- Enable story understanding.
- Enable metaphor understanding.
- Enable default reasoning via inheritance.
- All of the above.
- None of the above.

A key virtue of the transition-space representation is that it:

- Subsumes thematic role frames and primitive-act frames.
- Can be translated to relational-database records.
- Enables the description of states.
- Facilitates the understanding of metaphors.
- All of the above.
- None of the above.

A key virtue of the thematic-role-frame representation is that it:

- Expresses all actions in terms of a few primitive acts.
- Enables description at the story level..
- Focuses on variable-value changes.
- Captures causal relations.
- All of the above.
- None of the above.

A key virtue of semantic-transition-tree grammars is that they:

- Reduce the number of words that need to be understood.
- Simplify grammar construction by substituting recursion for explicit loops.
- Exploit the rete algorithm.
- Exploit transition-space representation.
- All of the above.
- None of the above.

Natural language database interfaces work because processed noun phrases most often become:

- Relational join operations.
- Relational selection operations.
- Relational projection operations.
- Relational sorting operations.
- All of the above.
- None of the above.

The purpose of crossover in genetic algorithms is to:

- Increase diversity.
- Model natural mutation.
- Change selection probabilities.
- Reduce the number of genotypes.
- All of the above.
- None of the above.

In neural nets:

- Biological neural nets are modeled accurately.
- Sigmoid thresholds were introduced to avoid overfitting.
- Overfitting occurs if there are too-few training cycles.
- The computation required by backpropagation  per training cycle is proportional to $n^2$, where n is the number of nodes.
- All of the above.
- None of the above.

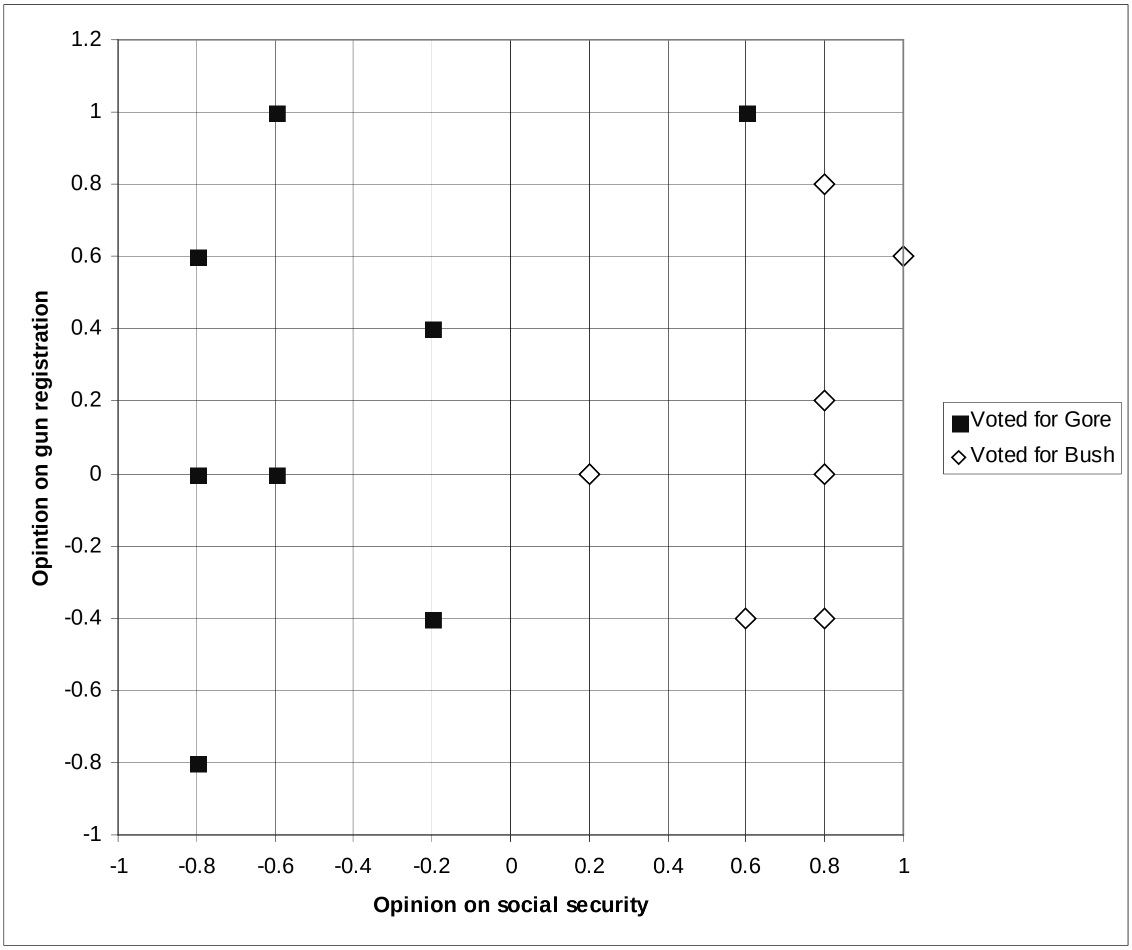
## Problem 2: Nearest Neighbor, Decision Trees and Support Vectors (30 points)

Congress has decided to ask each voter a few key questions so as to predict how each will vote. This will, of course, save everyone the troublesome and time-consuming practice of actually having to examine the ballots to figure out the election result.

They decide to start with just two questions:

a) On a scale of –1 (strongly disagree) to +1 (strongly agree) how do you feel about privatizing social security?
b) On a scale of –1 to +1, how do you feel about registering handguns?

The training sample is shown below, with 15 individuals plotted according to how they feel about these two issues, with a dark square (for Gore) and a light diamond (for Bush) indicating how they voted for president. There is also a question mark "?" on the plot, indicating one of the infamous undecided voters about which so much has been said during this election. We'll call him Mr. Undecided.



**Part A: Nearest Neighbor (10 points)**

1) What would nearest neighbor predict about the vote of Mr. Undecided, assuming the use of the standard Euclidean distance as the metric? (Your answer should be either Gore or Bush.)

2) On the plot above, **carefully** draw the **precise** boundary lines that nearest neighbor would indicate as separating the Gore part of the sample space from the Bush part. Do not include the **?** in your analysis.

3) What would 3-nearest neighbor predict about the vote of Mr. Undecided (using the same Euclidean metric)?

4) It turns out there was one other question that voters had been asked: “How do you feel about lowering the pay of Congressmen/women?” The question was not included in the publicly released data because, (according to the politicians who controlled the release), the data will not be useful in making a decision. When digging in, you find out that the actual problem was that all the answers were strongly clustered near the +1 end of the scale. You are brought in as a consultant and suggest that:

a) The politicians are correct; the data will not be useful.

b) The data can still be useful, you just need to **divide** all the values by the mean of the value.

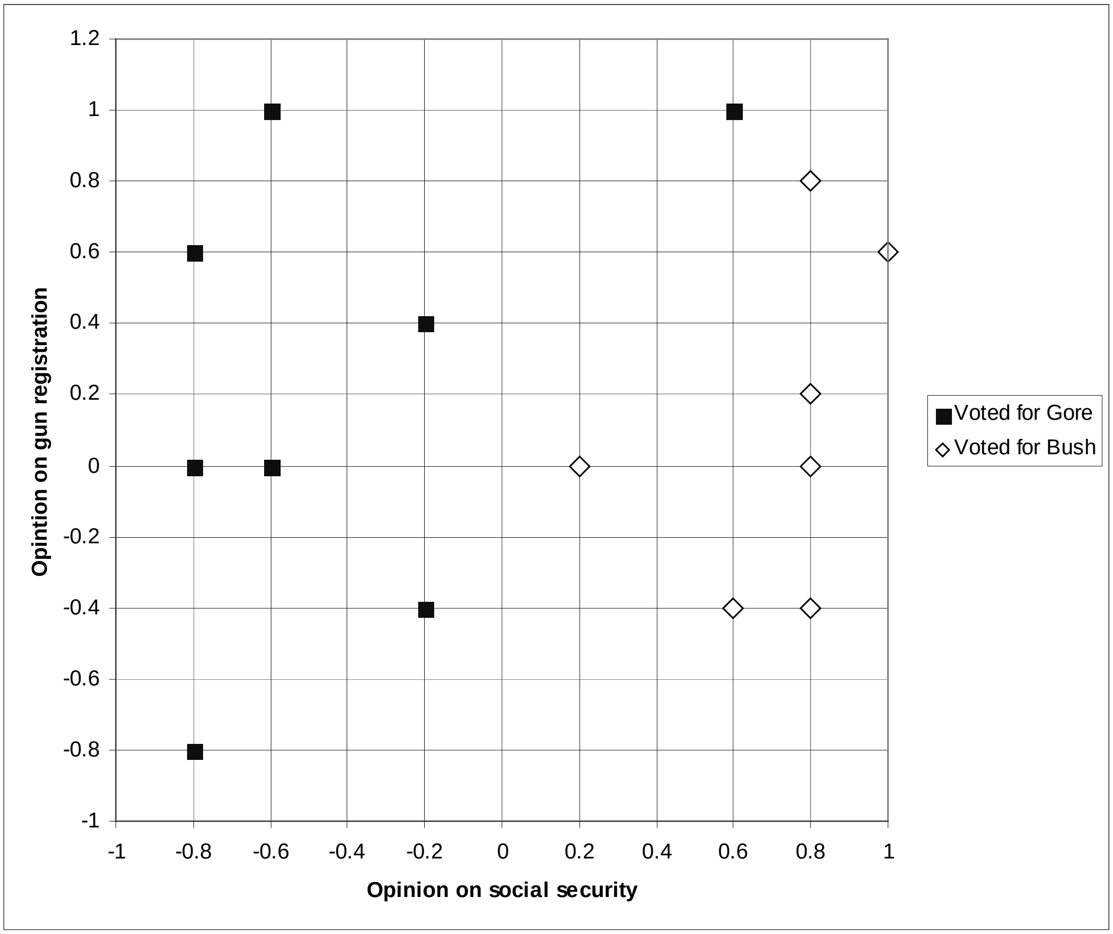
c) The data can still be useful, you just need to **subtract** from each value the mean of all the values, then **multiply** all the values by the standard deviation of the values

d) The data can still be useful, you just need to **divide** by the standard deviation.

e) The data can still be useful, but none of the choices offered above are correct.

## B: Identification Trees (8 points)

(We repeat the same data here for your convenience.)



Things seem to be going along well, when suddenly, Ralph Nader appears on the scene and suggests that nearest neighbor is wrecking the environment by wasting precious time and space. He suggests using ID trees instead.

1) You decide to try as your first test `opinion on social security < 0.` But as you know, you need to determine the average disorder of the sets produced by this test to see whether it's any good. What is the average disorder? (Your answer can include the $\log_2$ operator, you need not simplify your expression.)

2) You decide that it looks good, so you decide to complete the decision tree. Draw the ID tree and specify all tests. Do not include the **?** in your analysis.

3) What does your tree predict about how Mr. Undecided will vote for president? Circle the corresponding node on your ID tree.

**Part C: More Voter Questions (4 points)**

Mr. Gore, having invented the Internet, claims to know a thing or two about technology. He says that we're asking way too few questions of the voters, and indicates that to get a decent predictive ability we should ask them at least 100 questions. You come up with 100 questions, and while you worked hard at it, you don't think all 100 questions are going to give you predictive information about the voter. Nevertheless, you forge ahead and try using both nearest neighbors and identification trees. Your initial experiments indicate (circle the most likely result):

a) Both techniques work well and work about equally well.

b) Nearest neighbors works much better than identification trees.

c) Identification trees work much better than nearest neighbors.

d) Neither works well.

**Part D: Support Vector Machines (8 points)**

All of this is about to end when you notice a small but vocal demonstration going on out in the street, with people carrying signs that say

Support
Vector Machines

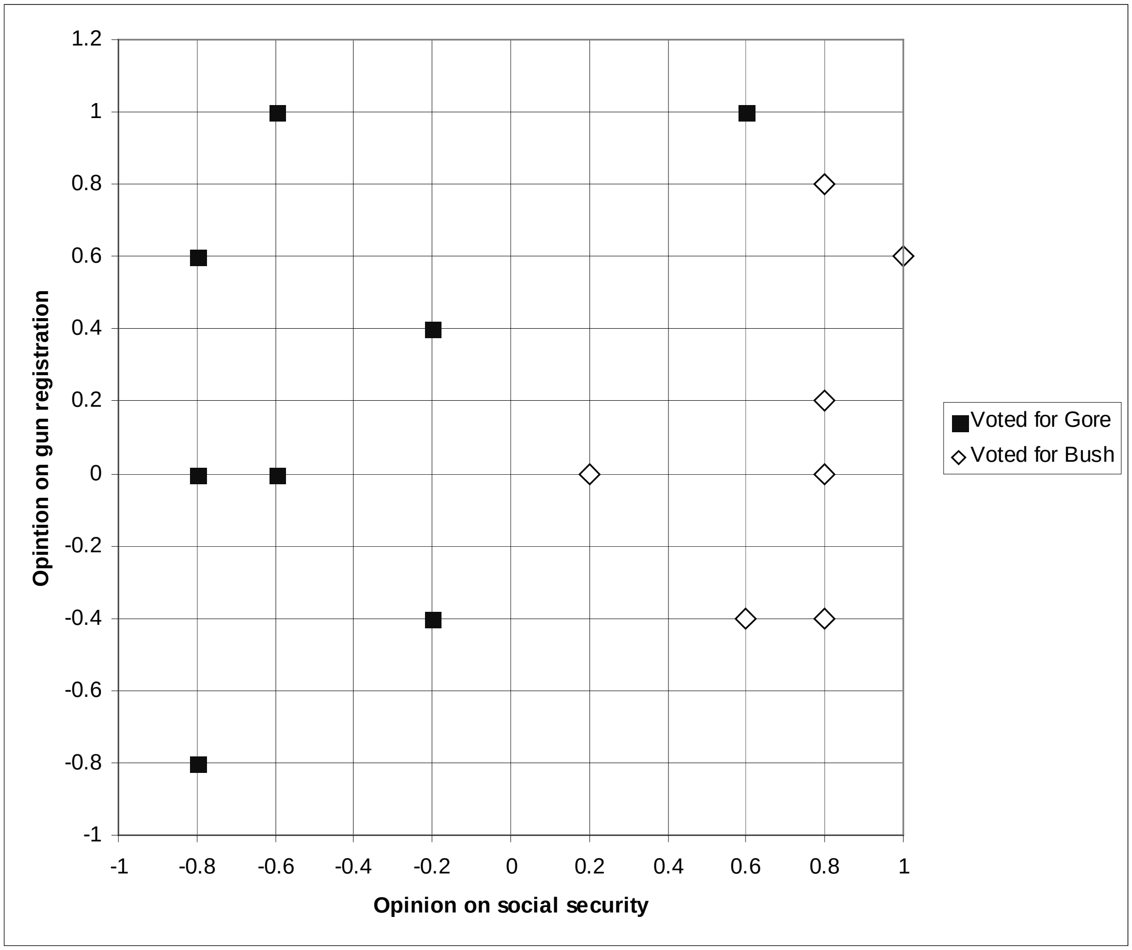
You’re a bit confused and ask a colleague, “What’s a vector machine?” He explains that you’ve parsed this a bit incorrectly, it’s called a support vector machine. This brings it all back to you, and you indicate your understanding by answering the following question. (A third copy of the same data is shown below for your convenience.)

You decide experiment with a radial-basis-function kernel,

$$K(\mathbf{x}_1,\mathbf{x}_2)=e^{\frac{-\|\mathbf{x}_1-\mathbf{x}_2\|^2}{2\sigma^2}}$$

Your experiment succeeds, with a small number of samples emerging as support vectors, and all samples correctly providing values of less than or equal to -1 or greater than or equal to +1.

1. Sketch the decision boundary in the diagram below.



2. Check ***all*** the statements that are correct:

❑ There is a sigma value such that all the Gore/Bush samples will become support vectors.

❑ There is a sigma value such that the Gore/Bush samples will not be separable.

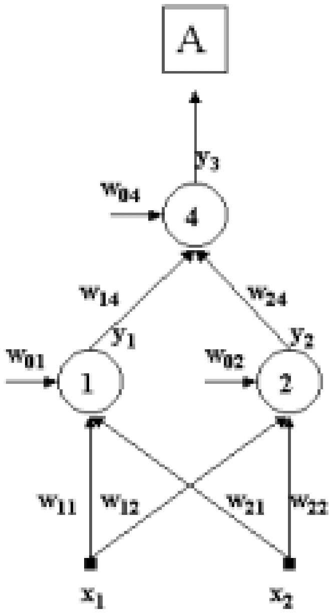
❑ There is a sigma value such that overfitting will occur.

❑ If the radial-basis-function kernel is replaced by a dot product kernel, $\mathbf{x_1} \cdot \mathbf{x_2}$, the Gore/Bush samples will be separable.

## Problem 3: Neural Nets (40 points)

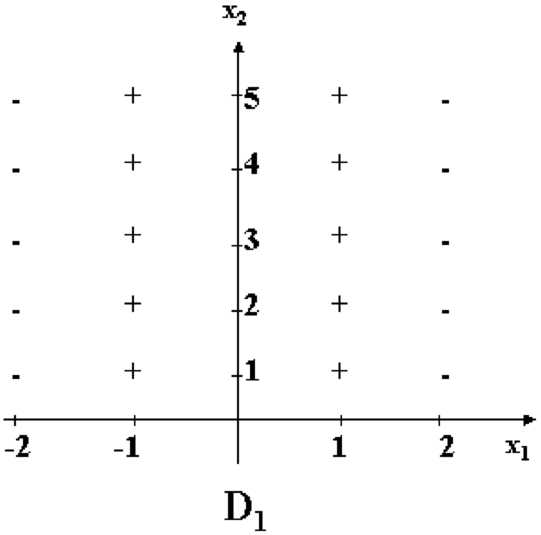
**Part A (26 points)**:

Consider the following neural network. **All three of the units are perceptron units, not sigmoid units; that is, all the outputs are 0 or 1.** Assume all **inputs** to threshold-implementing weights, such as $w_{01}$ and $w_{02}$ are -1.



**A.1 (10 points)**

Give values for the missing weights of net A so that it correctly classifies the Data in $D_1$ below. The output for each instance should be 0 for instances labeled – and 1 for instances labeled +. Note that axis ticks are **not negative samples.**



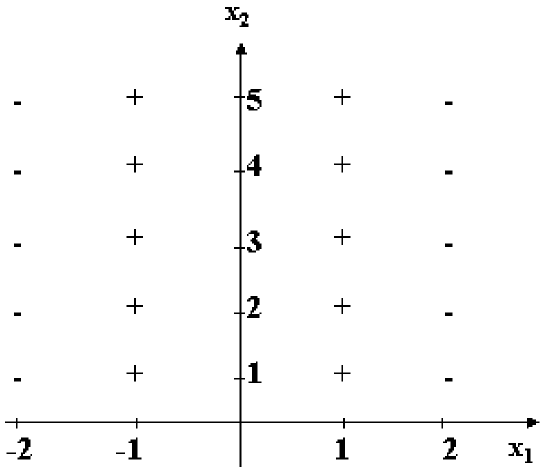
$D_1$

| | |
|---|---|
| $W_{01}$= -1.5 | |
| $W_{11}$= | |
| $W_{21}$= | |
| $W_{02}$= -1.5 | |
| $W_{12}$= | |
| $W_{22}$= | |
| $W_{04}$= | |
| $W_{14}$= | |
| $W_{24}$= | |

**A.2 (8 points)**

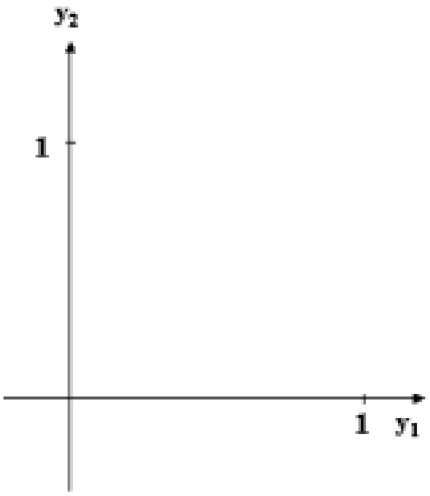
**Using the weights you computed in A.1**:

1. Carefully draw the decision boundaries for units 1 and 2 of A on the diagram below. Be sure to label each boundary with a 1 or 2.



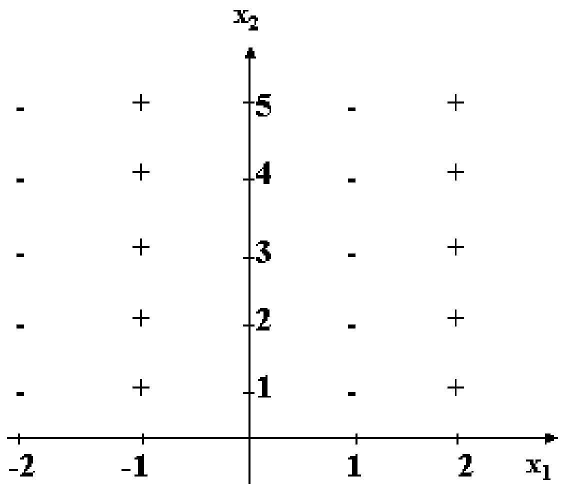
$D_1$

2. On the diagram below, draw in all combinations of outputs for units 1 and 2. Label each combination with the class(es) associated with that combination (+ or – or both), given the data set $D_{1.}$

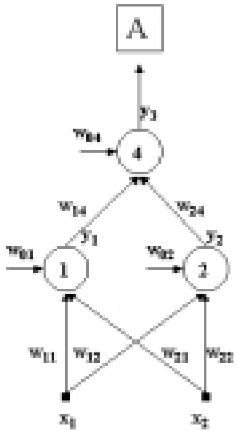


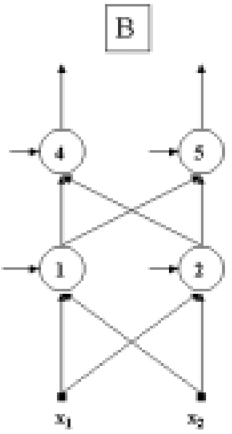
**A.3 (8 points)**

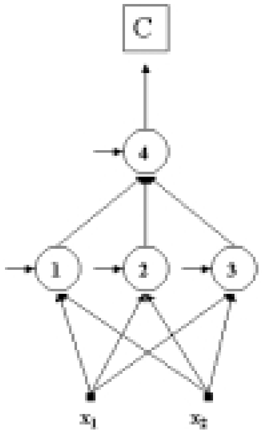
Consider the following data set, which is somewhat different from D$_1$:



D$_2$

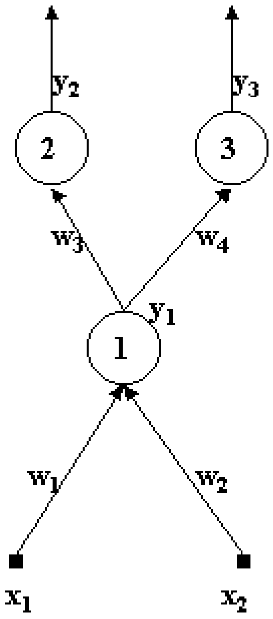


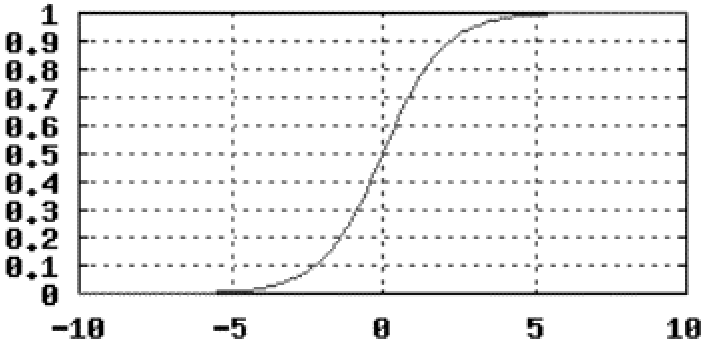




Circle ALL of the net architectures, if any, that can correctly classify this data set. **Assume that all the architectures use only perceptron units**.

**Part B (14 points)**





**Sigmoid Values**

| | | | | | |
|---|---|---|---|---|---|
| S(-5) = 0 | S(-4)=0.02 | S(-3)=0.04 | S(-2)=0.12 | S(-1)=0.27 | S(0)=0.5 |
| S(5)=1 | S(4)=0.98 | S(3)=0.96 | S(2)=0.88 | S(1)=0.73 | |

**B.**1 (6 points)

Assuming that each unit has a fixed threshold of 0.0, compute (approximately), given that the inputs are both set to 2 and the weights are $w_1$= 2, $w_2$= -2, $w_3$=4, and $w_4$=0.

- **$y_1$ =**
- **$y_2$ =**
- **$y_3$ =**

**B.2 (8 points)**

You are confronted with two situations, and you are to determine whether the weights will go up or down. In both situations, you are to use the network of part B.1, which uses sigmoid neurons.

In **situation 1**, you are to assume that the desired outputs for inputs

$x_1 = 2$,
$x_2 = 2$

are

$y_2 = 0$,
$y_3 = 1$,

that the learning rate is 1, and that the weights are as in B.1.

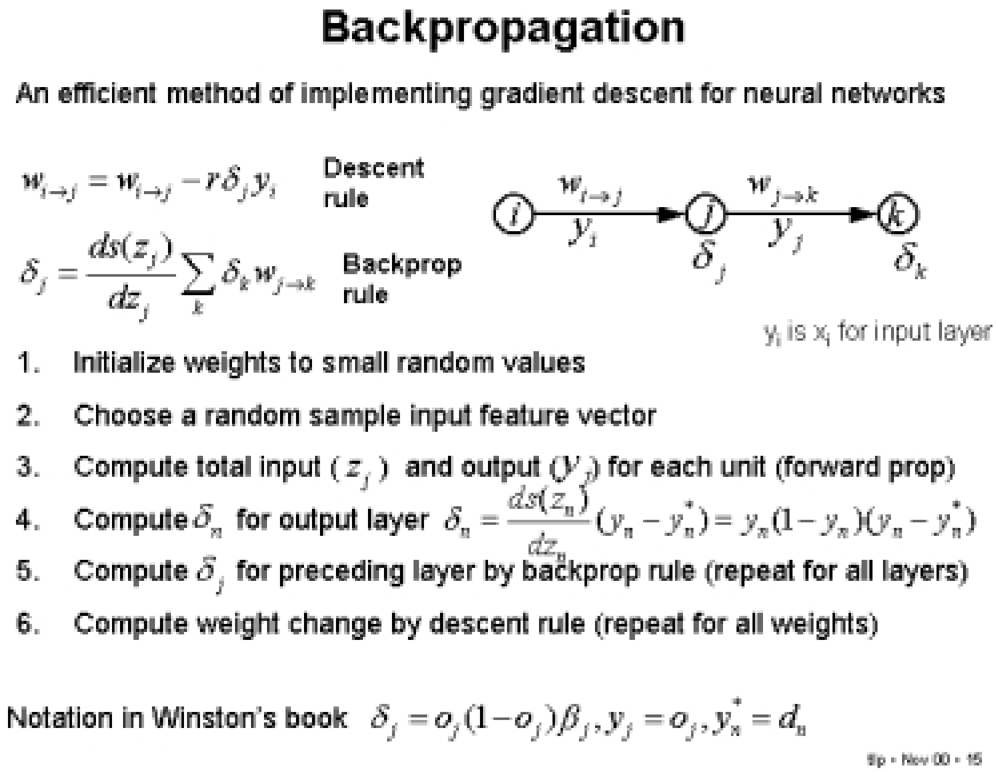
In **situation 2**, you are to assume that the desired outputs, for the same inputs, are

$y_2 = 1$,
$y_3 = 0$,

Fill in the cells of the table below with **up** and **down,** as appropriate**.**

| | **Situation 1** | **Situation 2** |
|---|---|---|
| $\mathbf{w_1}$ | | |
| $\mathbf{w_2}$ | | |
| $\mathbf{w_3}$ | | |
| $\mathbf{w_4}$ | | |

We provide a following page with helpful information that you can tear off and use for reference.



# Backpropagation

An efficient method of implementing gradient descent for neural networks

$w_{i\to j} = w_{i\to j} - r\delta_j y_i$ Descent rule

$\delta_j = \dfrac{ds(z_j)}{dz_j}\sum_k \delta_k w_{j\to k}$ Backprop rule

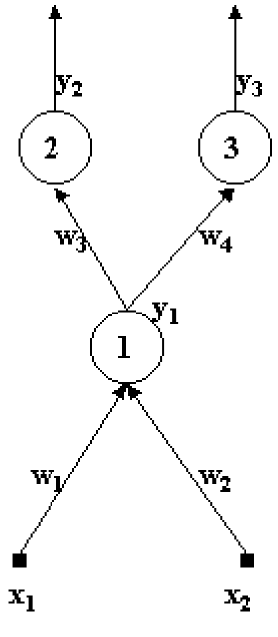
$i$
$w_{i\to j}$
$y_i$
$j$
$\delta_j$
$w_{j\to k}$
$y_j$
$k$
$\delta_k$

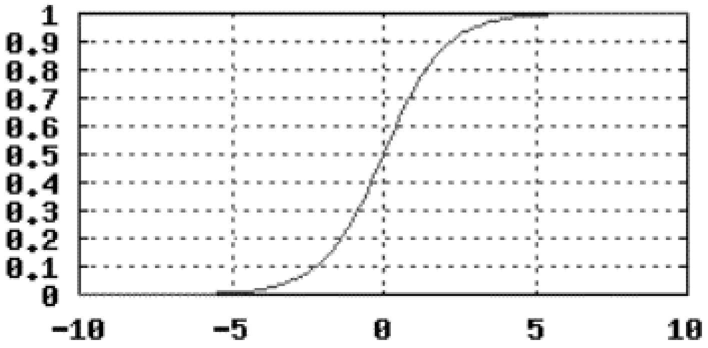
$y_i$ is $x_i$ for input layer

1. Initialize weights to small random values
2. Choose a random sample input feature vector
3. Compute total input ( $z_j$ ) and output ( $y_j$ ) for each unit (forward prop)
4. Compute $\delta_n$ for output layer $\delta_n = \dfrac{ds(z_n)}{dz_n}(y_n - y_n^*) = y_n(1-y_n)(y_n - y_n^*)$
5. Compute $\delta_j$ for preceding layer by backprop rule (repeat for all layers)
6. Compute weight change by descent rule (repeat for all weights)

Notation in Winston's book $\delta_j = o_j(1-o_j)\beta_j, y_j = o_j, y_n^* = d_n$

tlp - Nov 00 - 15





$x_1 = x_2 = 2$

$w_1$= 2, $w_2$= -2, $w_3$=4, and $w_4$=0.