6.867 Machine learning

Final exam

December 3, 2004

Your name and MIT ID:

J. D. 00000000

(Optional) **The grade you would give to yourself + a brief justification.**

A... why not?

Problem 1

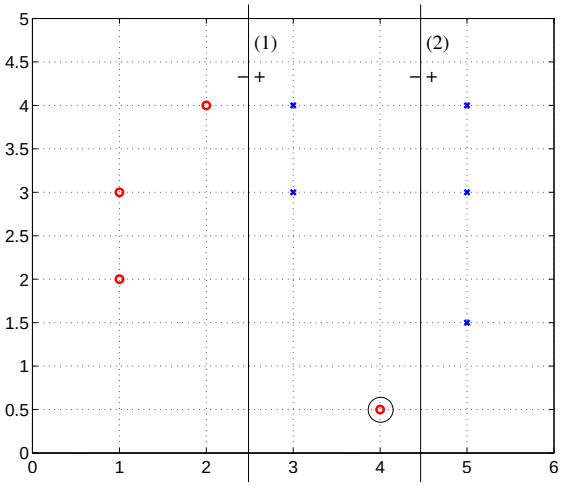


Figure 1: Labeled training points for problem 1.

Consider the labeled training points in Figure 1, where 'x' and 'o' denote positive and negative labels, respectively. We wish to apply AdaBoost with decision stumps to solve the classification problem. In each boosting iteration, we select the stump that minimizes the weighted training error, breaking ties arbitrarily.

1. **(3 points)** In figure 1, draw the decision boundary corresponding to the first decision stump that the boosting algorithm would choose. Label this boundary (1), and also indicate +/- side of the decision boundary.

2. **(2 points)** In the same figure 1 also circle the point(s) that have the highest weight after the first boosting iteration.

3. **(2 points)** What is the weighted error of the first decision stump after the first boosting iteration, i.e., after the points have been reweighted?

0.5

4. **(3 points)** Draw the decision boundary corresponding to the second decision stump, again in Figure 1, and label it with (2), also indicating the +/- side of the boundary.

5. **(3 points)** Would some of the points be misclassified by the combined classifier after the two boosting iterations? Provide a brief justification. (the points will be awarded for the justification, not whether your y/n answer is correct)

   Yes. For example, the circled point in the figure is misclassified by the first decision stump and could be classified correctly in the combination only if the weight/votes of the second stump is higher than the first. If it were higher, however, then the points misclassified by the second stump would be misclassified in the combination.

Problem 2

1. **(2 points)** Consider a linear SVM trained with n labeled points in $\mathcal{R}^2$ without slack penalties and resulting in $k = 2$ support vectors ($k < n$). By adding one additional labeled training point and retraining the SVM classifier, what is the maximum number of support vectors in the resulting solution?

   ( ) k

   ( ) $k + 1$

   ( ) $k + 2$

   (X) $n + 1$

2. We train two SVM classifiers to separate points in $\mathcal{R}^2$. The classifiers differ only in terms of the kernel function. Classifier 1 uses the linear kernel $K_1(\mathbf{x}, \mathbf{x}') = \mathbf{x}^T\mathbf{x}'$, and classifier 2 uses $K_2(\mathbf{x}, \mathbf{x}') = p(\mathbf{x})p(\mathbf{x}')$, where $p(\mathbf{x})$ is a 3-component Gaussian mixture density, estimated on the basis of related other problems.

   (a) **(3 points)** What is the VC-dimension of the second SVM classifier that uses kernel $K_2(\mathbf{x}, \mathbf{x}')$?

2

   The feature space is 1-dimensional; each point $\mathbf{x} \in \mathcal{R}^2$ is mapped to a non-negative number $p(\mathbf{x})$.

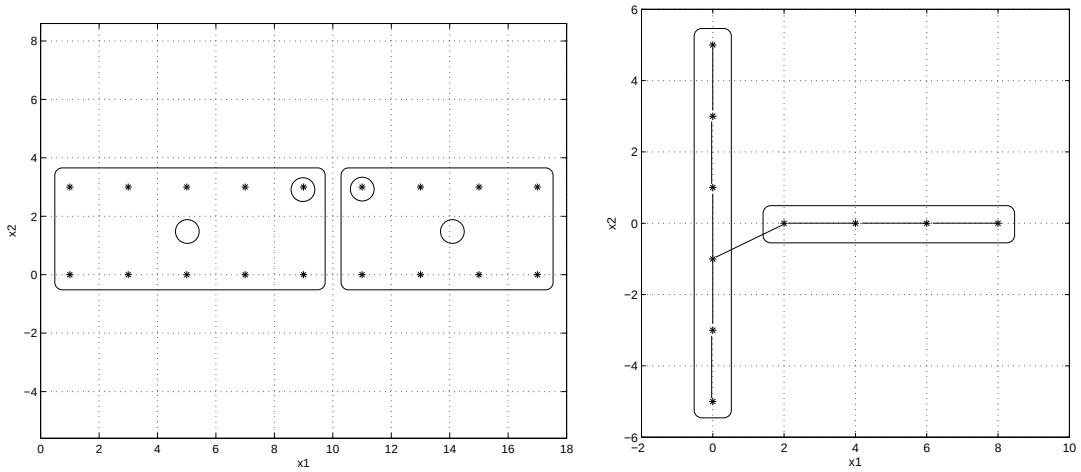
   (b) **(T/F − 2 points)** The second SVM classifier can only separate points that are likely according to $p(\mathbf{x})$ from those that have low probability under $p(\mathbf{x})$.

T

   (c) **(4 points)** If both SVM classifiers achieve zero training error on n labeled points, which classifier would have a better generalization guarantee? Provide a brief justification.

   The first classifier has VC-dimension 3 while the second one has VC-dimension 2. The complexity penalty for the first one is therefore higher. When the number of training errors is the same for the two classifiers, the bound on the expected error is smaller for the second classifier.

Problem 3



(a) (b)

Figure 2: Data sets for clustering. Points are located at integer coordinates.

1. **(4 points)** First consider the data plotted in Figure 2a, which consist of two rows of equally spaced points. If k-means clustering ($k = 2$) is initialised with the two points whose coordinates are $(9, 3)$ and $(11, 3)$, indicate the final clusters obtained (after the algorithm converges) on Figure 2a.

2. **(4 points)** Now consider the data in Figure 2b. We will use spectral clustering to divide these points into two clusters. Our version of spectral clustering uses a neighbourhood graph obtained by connecting each point to its two nearest neighbors (breaking ties randomly), and by weighting the resulting edges between points $\mathbf{x}_i$ and $\mathbf{x}_j$ by $W_{ij} = \exp(-||\mathbf{x}_i - \mathbf{x}_j||)$. Indicate on Figure 2b the clusters that we will obtain from spectral clustering. Provide a brief justification.

   The random walk induced by the weights can switch between the clusters in the figure in only two places, (0,-1) and (2,0). Since the weights decay with distance, the weights corresponding to transitions within clusters are higher than those going across in both places. The random would would therefore tend to remain within the clusters indicated in the figure.

3. **(4 points)** Can the solution obtained in the previous part for the data in Figure 2b also be obtained by k-means clustering ($k = 2$)? Justify your answer.

   No. In the k-means algorithm points are assigned to the closest mean (cluster centroid). The centroids of the left and right clusters in the figure are (0,0) and (5,0), respectively. Point (2,0), for example, is closer to the left cluster centroid (0,0) and wouldn't be assigned to the right cluster. The two clusters in the figure therefore cannot be fixed points of the k-means algorithm.

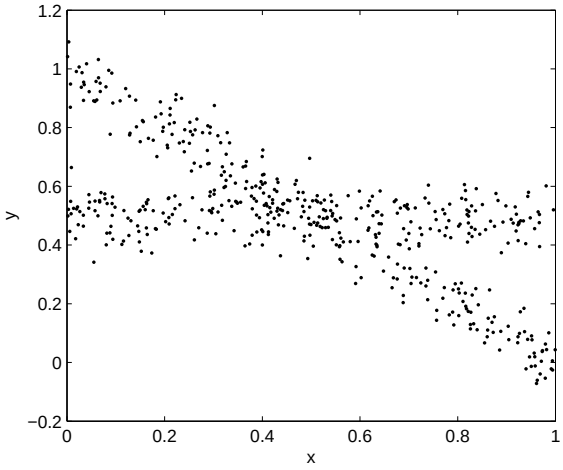


Figure 3: Training sample from a mixture of two linear models

Problem 4

The data in Figure 3 comes from a mixture of two linear regression models with Gaussian noise:

$$P(y|x;\theta) = p_1\mathcal{N}(y\,;\, w_{10} + w_{11}x,\ \sigma_1^2) + p_2\mathcal{N}(y\,;\, w_{20} + w_{21}x,\ \sigma_2^2)$$

where $p_1 + p_2 = 1$ and $\theta = (p_1, p_2, w_{10}, w_{11}, w_{20}, w_{21}, \sigma_1, \sigma_2)$. We hope to estimate θ from such data via the EM algorithm.

To this end, let $z \in \{1, 2\}$ be the mixture index, variable indicating which of the regression models is used to generate y given x.

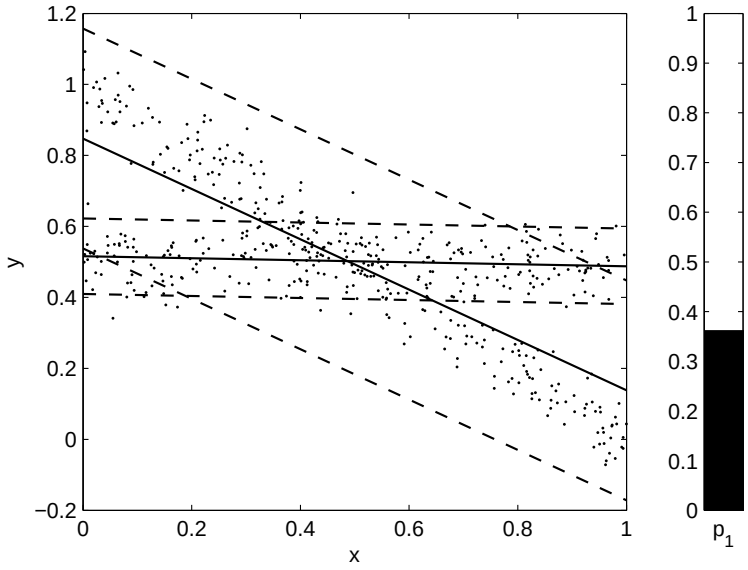
1. **(6 points)** Connect the random variables X, Y, and Z with directed edges so that the graphical model on the left represents the mixture of linear regression models described above, and the one on the right represents a mixture-of-experts model. For both models, Y denotes the output variable, X the input, and Z is the choice of the linear regression model or expert.

We use a single plot to represent the model parameters (see the figure below). Each linear regression model appears as a solid line ($y = w_{i0} + w_{i1}x$) in between two parallel dotted lines at vertical distance $2\sigma_i$ to the solid line. Thus each regression model "covers" the data that falls between the dotted lines. When $w_{10} = w_{20}$ and $w_{11} = w_{21}$ you would only see a single solid line in the figure; you may still see two different sets of dotted lines corresponding to different values of σ_1 and σ_2. The solid bar to the right represents p_1 (and $p_2 = 1 - p_1$).

For example, if

$$\begin{aligned}\theta &= (\ p_1,\ p_2,\ w_{10},\ w_{11},\ w_{20},\ w_{21},\ \sigma_1,\ \sigma_2) \\ &= (\ 0.35,\ 0.65,\ 0.5,\ 0,\ 0.85,\ -0.7,\ 0.05,\ 0.15)\end{aligned}$$

the plot is

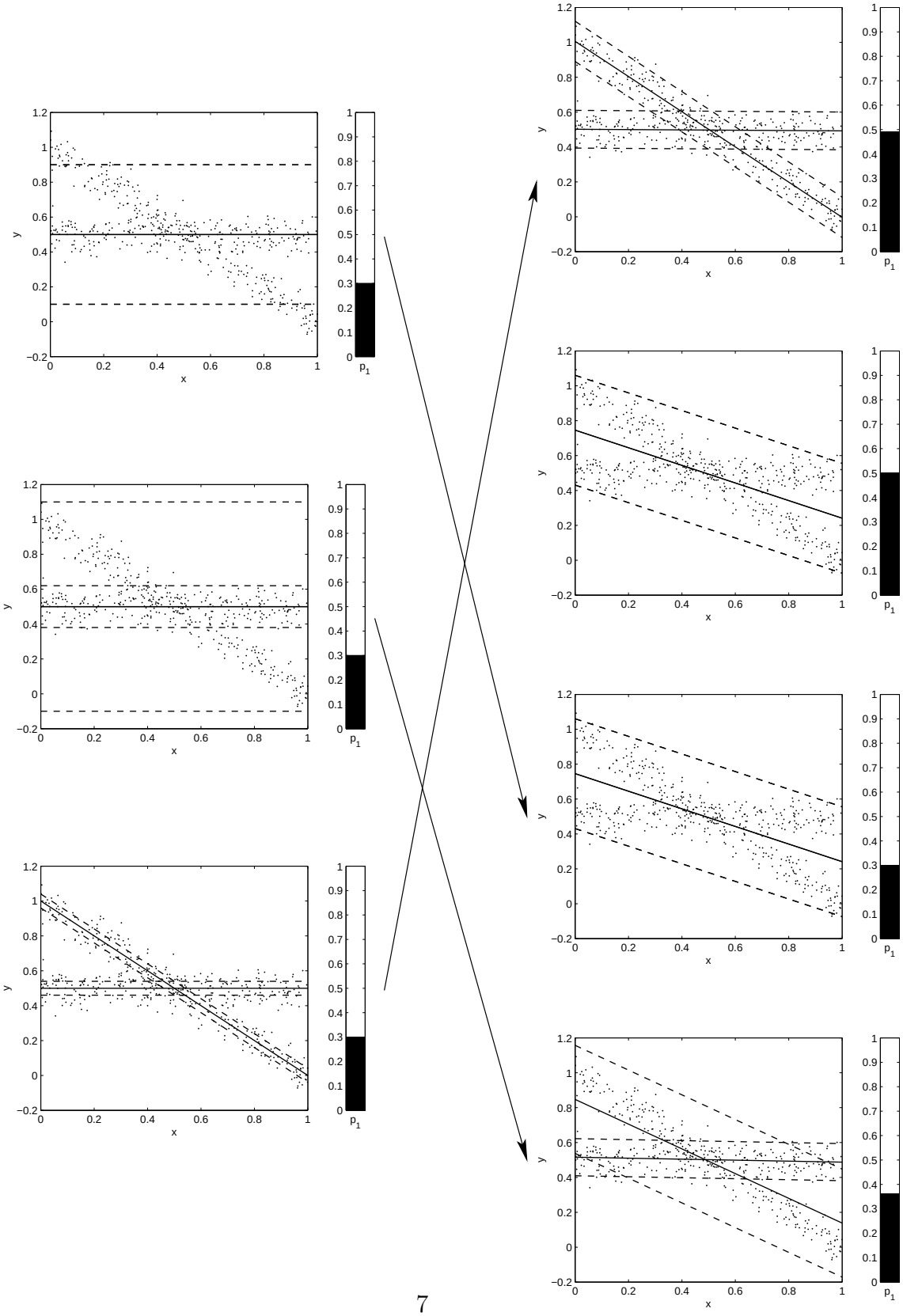


2. **(6 points)** We are now ready to estimate the parameters θ via EM. There are, however, many ways to initialize the parameters for the algorithm.

   On the next page you are asked to connect 3 different initializations (left column) with the parameters that would result after one EM iteration (right column). Different initializations may lead to the same set of parameters. Your answer should consist of 3 arrows, one from each initialization.

Initialization

Next iteration



Problem 5

Assume that the following sequences are very long and the pattern highlighted with spaces is repeated:

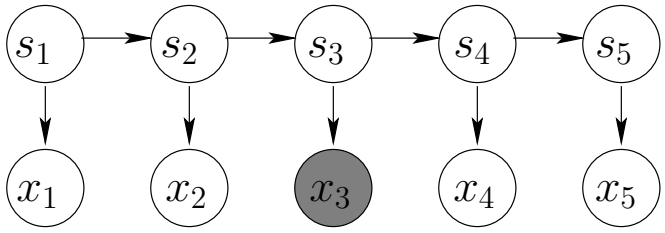
Sequence 1: `1 0 0  1 0 0  1 0 0  1 0 0 ... 1 0 0`

Sequence 2: `1  1 0 0  1 0 0  1 0 0 ... 1 0 0`

1. **(4 points)** If we model each sequence with a different first-order HMM, what is the number of hidden states that a reasonable model selection method would report?

	HMM for Sequence 1	HMM for Sequence 2
No. of hidden states	3	4

2. **(2 points)** The following Bayesian network depicts a sequence of 5 observations from an HMM, where s_1, s_2, s_3, s_4, s_5 is the hidden state sequence.



Are x_1 and x_5 independent given x_3? Briefly justify your answer.

They are not independent. The moralized ancestral graph corresponding to x_1, x_3, and x_5 is the same graph with arrows replaced with undirected edges. x_1 and x_5 are not separated given x_3, and thus not independent.

3. **(3 points)** Does the order of Markov dependencies in the observed sequence always determine the number of hidden states of the HMM that generated the sequence? Provide a brief justification.

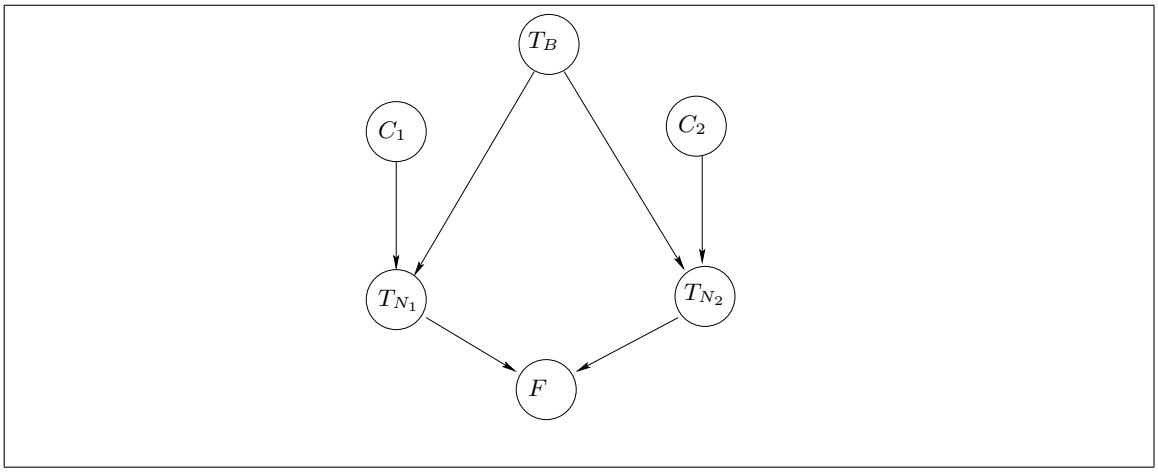
No. The answer to the previous question implies that observations corresponding to (typical) HMMs have no Markov properties (of any order). This holds, for example, when there are only two possible hidden states. Thus Markov properties of the observation sequence cannot in general determine the number of hidden states.

Problem 6

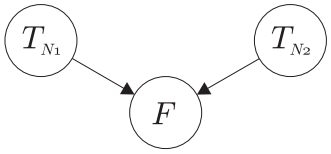
We wish to develop a graphical model for the following transportation problem. A transport company is trying to choose between two alternative routes for commuting between Boston and New York. In an experiment, two identical busses leave Boston at the same but otherwise random time, T_B. The busses take different routes, arriving at their (common) destination at times T_{N1} and T_{N2}.

Transit time for each route depends on the congestion along the route, and the two congestions are unrelated. Let us represent the random delays introduced along the routes by variables $C1$ and $C2$. Finally, let F represent the identity of the bus which reaches New York first. We view F as a random variable that takes values 1 or 2.

1. **(6 points)** Complete the following directed graph (Bayesian network) with edges so that it captures the relationships between the variables in this transportation problem.



2. **(3 points)** Consider the following directed graph as a possible representation of the independences between the variables T_{N1}, T_{N2}, and F only:



Which of the following factorizations of the joint are consistent with the graph?

$P(T_{N1})P(T_{N2})P(F\|T_{N1}, T_{N2})$	X
$P(T_{N1})P(T_{N2})P(F\|T_{N1})$	X
$P(T_{N1})P(T_{N2})P(F)$	X