

# 6.867 Machine learning

## FINAL exam

December 11, 2006

(2 points) Your name and MIT ID:

### Problem 1

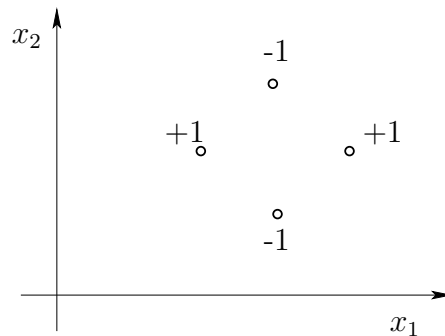


Figure 1: Labeled training points

We will consider here classifiers based on two different kernel functions

$$K_1(\mathbf{x}, \mathbf{x}') = \mathbf{x}^T \mathbf{x}'$$
$$K_2(\mathbf{x}, \mathbf{x}') = \exp(-\beta \|\mathbf{x} - \mathbf{x}'\|^2), \quad \beta > 0$$

1.1 Suppose we use a hard margin SVM (no slack) to train a classifier with the radial basis kernel  $K_2$  and the labeled points in Figure 1.

a) (2 points) Would we be able to find a separating solution? Please answer Y or N.

Y

- b) **(3 points)** If we set  $\beta$  to a very large value, how many support vectors would we get?

4

1.2 Let  $F_1$  be the set of linear classifiers of the form  $f(\mathbf{x}; \theta) = \theta^T \phi^{(1)}(\mathbf{x}) + \theta_0$  where  $\phi^{(1)}(\mathbf{x})$  is the feature mapping corresponding to the kernel  $K_1$  (linear).  $F_2$  is defined similarly based on  $K_2$ .

- a) **(2 points)** Are the two sets of classifiers nested in the sense that  $F_1 \subseteq F_2$ ? Please answer Y or N.

Y

- b) **(4 points)** Consider the labeled training points in Figure 1. In structural risk minimization, we use the 0-1 training error and the VC-dimension to guide which model (kernel) to choose. Without explicating exactly how we will train the classifiers, can we say, based on the figure and the two kernels, which one we would choose according to the structural risk minimization criterion? Briefly explain your answer.

*The VC-dimension of the SVM classifier with the radial basis kernel is infinite. The resulting complexity penalty for this classifier in the basic structural risk minimization criterion is also infinite. We would therefore always choose the linear classifier.*

## Problem 2

Consider building an ensemble of decision stumps with the AdaBoost algorithm. Figure 2 displays the labeled points in two dimensions as well as the first stump we have chosen. Stumps predict binary  $\pm 1$  values and, as linear classifiers, depend only on one of the coordinate values. The little arrow in the figure is the normal to the stump decision boundary indicating the positive side where the stump predicts  $+1$ . All the points start with uniform weights.

- 2.1 **(3 points)** Circle all the point(s) in Figure 2 whose weight will increase as a result of incorporating the first stump (the weight update due to the first stump).
- 2.2 **(3 points)** Draw in the same figure a possible stump that we could select at the next boosting iteration. You need to draw both the decision boundary and its positive orientation (as in the figure above).

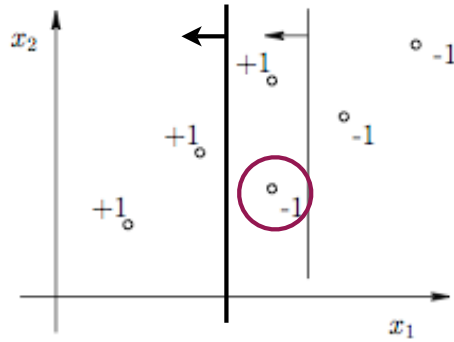


Figure 2: Labeled points and the first decision stump. The arrow points in the positive direction from the stump decision boundary.

2.2 (4 points) Will the second stump receive higher coefficient in the ensemble than the first? In other words, will  $\alpha_2 > \alpha_1$ ? Briefly explain your answer. (no calculation should be necessary).

*$\alpha_2 > \alpha_1$  because the point that the second stump misclassifies will have a smaller relative weight since it is classified correctly by the first stump.*

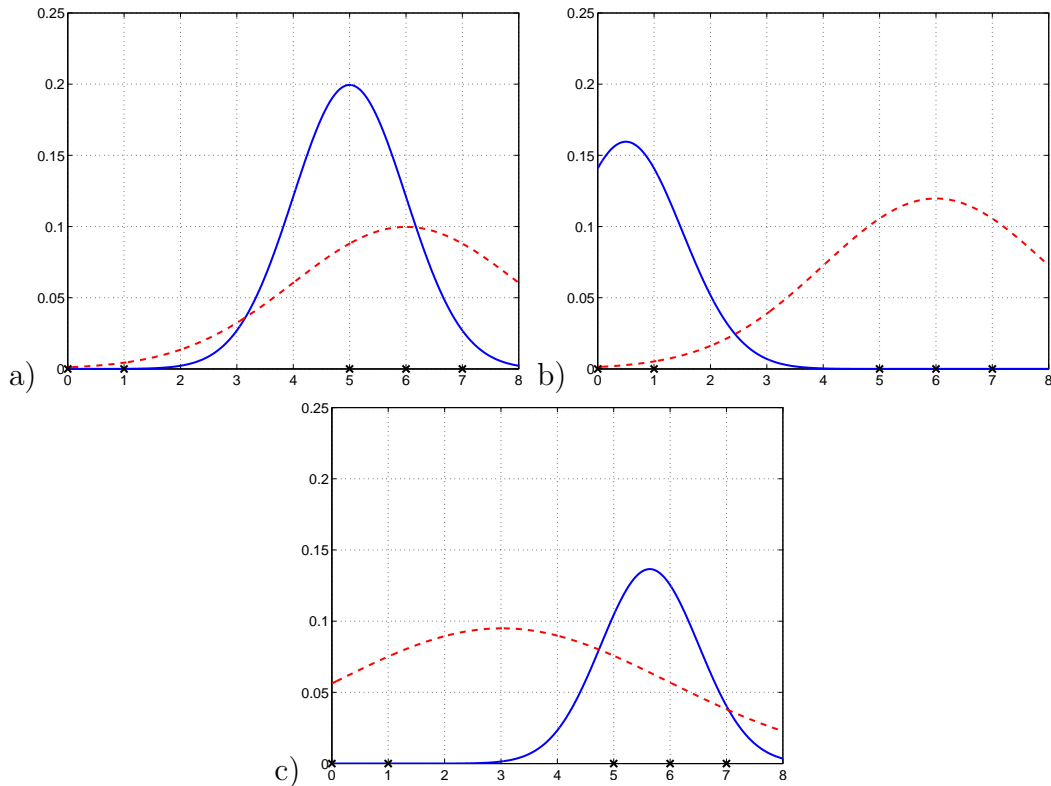
### Problem 3

Here we are estimating a mixture of two Gaussians via the EM algorithm. The mixture distribution over  $x$  is given by

$$P(x; \theta) = P(1)N(x; \mu_1, \sigma_1^2) + P(2)N(x; \mu_2, \sigma_2^2)$$

Any student in this class could solve this estimation problem easily. Well, one student, devious as they were, scrambled the order of figures illustrating EM updates. They may have also slipped in a figure that does not belong. Your task is to extract the figures of successive updates and explain why your ordering makes sense from the point of view of how the EM algorithm works.

All the figures plot  $P(1)N(x; \mu_1, \sigma_1^2)$  as a function of  $x$  with a solid line and  $P(2)N(x; \mu_2, \sigma_2^2)$  with a dashed line.



3.1 (T/F – 2 points) In the mixture model, we can identify the most likely posterior assignment, i.e.,  $j$  that maximizes  $P(j|x)$ , by comparing the values of  $P(1)N(x; \mu_1, \sigma_1^2)$  and  $P(2)N(x; \mu_2, \sigma_2^2)$ .

T

3.2 a) (4 points) Assign two figures to the correct steps in the EM algorithm.

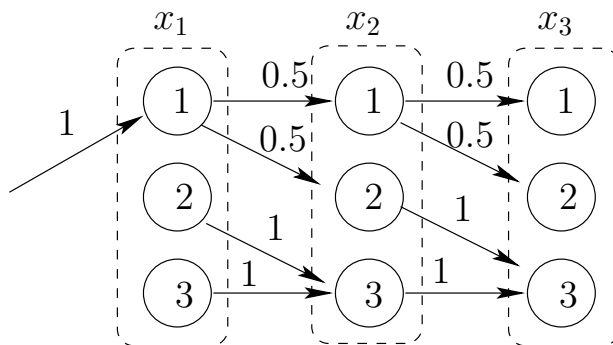
- Step 0: ( a ) – initial mixture distribution
- Step 1: ( c ) – after one EM-iteration

b) (4 points) Briefly explain how the mixture you chose for “step 1” follows from the mixture you have in “step 0”

*The two points on the left will be assigned more to the second (red) Gaussian since  $P(1)N(x; \mu_1, \sigma_1^2) < P(2)N(x; \mu_2, \sigma_2^2)$  for those points. The points on the right, except for the very last one, will be assigned mostly to the first (blue) Gaussian. As a result, the first Gaussian will become more concentrated around the two points on the right, while the second (red) Gaussian will move to the left and will have a higher variance as, in the M-step, it is estimated essentially on the basis of the spread out points  $x = 0$ ,  $x = 1$ , and  $x = 7$ .*

## Problem 4

Suppose we have a Hidden Markov model with three possible states and transitions described in the figure below. The states at times  $t = 1, 2, 3$  are represented by variables  $x_1$ ,  $x_2$ , and  $x_3$ , respectively. The distributions governing binary (0/1) outputs  $y_1$ ,  $y_2$ , and  $y_3$  are state dependent but do not depend on time. In other words, we only need to define  $P(y|x)$  common to all time points. Specifically,

$$P(y|x) : \begin{array}{c|ccc} & x = 1 & x = 2 & x = 3 \\ \hline y = 0 & 0.5 & 0.5 & 0.1 \\ y = 1 & 0.5 & 0.5 & 0.9 \end{array}$$


- 4.1 **(3 points)** There are only three possible hidden state sequences over the three time points  $t = 1, 2, 3$ . What are they?

*111, 112, and 123*

- 4.2 **(3 points)** What is the most likely hidden state sequence given the observed sequence  $(y_1 = 0, y_2 = 0, y_3 = 0)$ ? Is the answer unique?

Note first that the observation sequence does not help distinguish states 1 and 2 but would discount state 3 since  $P(y = 0|x = 3) = 0.1$ . The answer is not unique. Both 111 and 112 are equally likely according to the Markov chain and the observations do not help distinguish them. More precisely:

$$\begin{aligned} P(x_1 = 1, x_2 = 1, x_3 = 1, y_1 = 0, y_2 = 0, y_3 = 0) &= (1 \cdot 0.5) \cdot (0.5 \cdot 0.5) \cdot (0.5 \cdot 0.5) \\ P(x_1 = 1, x_2 = 1, x_3 = 2, y_1 = 0, y_2 = 0, y_3 = 0) &= (1 \cdot 0.5) \cdot (0.5 \cdot 0.5) \cdot (0.5 \cdot 0.5) \\ P(x_1 = 1, x_2 = 2, x_3 = 3, y_1 = 0, y_2 = 0, y_3 = 0) &= (1 \cdot 0.5) \cdot (0.5 \cdot 0.5) \cdot (1 \cdot 0.1) \end{aligned}$$

where the probabilities within parenthesis represent transitioning to the state at the corresponding time point and generating  $y = 0$ .

4.3 **(4 points)** Suppose we only receive observations for the first two time points, i.e.,  $(y_1, y_2)$ , and train the HMM with the EM algorithm. The HMM is initialized with the parameters illustrated above. Select which of the following parameters could change as a result of running the EM algorithm:

- transitions from state  $x = 1$ , i.e.,  $P_{1j} = P(x_{t+1} = j|x_t = 1)$ ,  $j = 1, 2, 3$ .
- transitions from state  $x = 3$ , i.e.,  $P_{3j} = P(x_{t+1} = j|x_t = 3)$ ,  $j = 1, 2, 3$ .
- output distribution from state  $x = 1$ , i.e.,  $P(y|x = 1)$ ,  $y = 0, 1$ .
- output distribution from state  $x = 3$ , i.e.,  $P(y|x = 3)$ ,  $y = 0, 1$ .

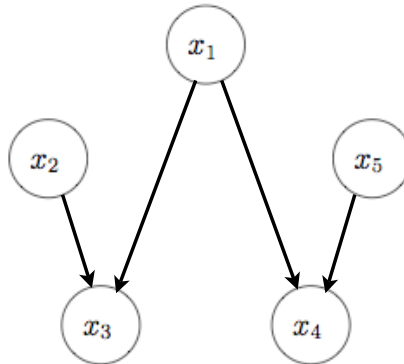
## Problem 5

We administered a short exam for  $n = 60$  students. There were only five exam questions and all of them were simple true/false questions. We were interested in finding out how the answers might depend on each other. To this end, we collected all the answers into a dataset  $D = \{(x_1^t, \dots, x_5^t), t = 1, \dots, n\}$ , where  $x_i^t$  is student  $t$ 's answer (T/F) to question  $i$ . From this data we were able to estimate a Bayesian network, graph  $G$  and the associated distribution, over the five variables  $x_1, \dots, x_5$  (answers to questions).

So, of course we misplaced the graph. But we do remember a few relevant properties that may help reconstruct the graph. In particular,

- a)  $x_1, x_2$ , and  $x_5$  were all marginally independent of each other,
- b) knowing the answer to  $x_1$  made  $x_2$  independent of  $x_4$  and  $x_3$  independent of  $x_5$ .

5.1 **(6 points)** Draw a Bayesian network that you can infer from the above constraints. Draw the edges in the figure below.



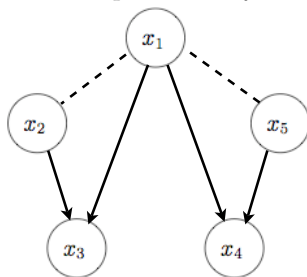
5.2 (3 points) What can you say about the form of the distribution over the five variables?

*Based on the graph we now the distribution has to factor according to*

$$P(x_1)P(x_2)P(x_5)P(x_3|x_1, x_2)P(x_4|x_1, x_5)$$

5.3 (4 points) Consider *only* students who answered  $x_3 = T$  and  $x_4 = T$ . If we looked at their answers to  $x_2$  and  $x_5$ , would we expect these answers to be independent of each other? Briefly justify your answer.

*The answers would be dependent. We could derive this formally by asking whether  $x_2$  is independent of  $x_5$  given  $x_3$  and  $x_4$ . The moralized ancestral graph is:*



*Alternatively, we can simply note that if we know  $x_3$ ,  $x_1$  and  $x_2$  are dependent (induced dependence). Similarly, if we know  $x_4$ ,  $x_1$  and  $x_5$  become dependent.*



# Additional set of figures

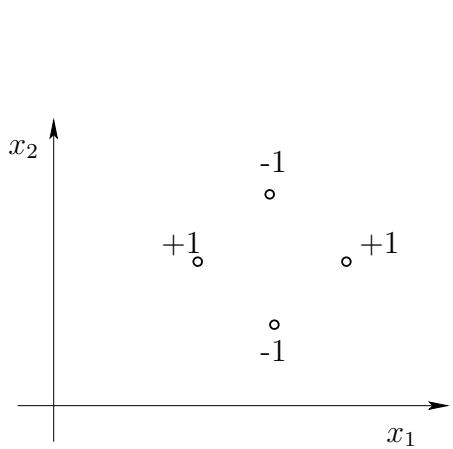


Figure 1

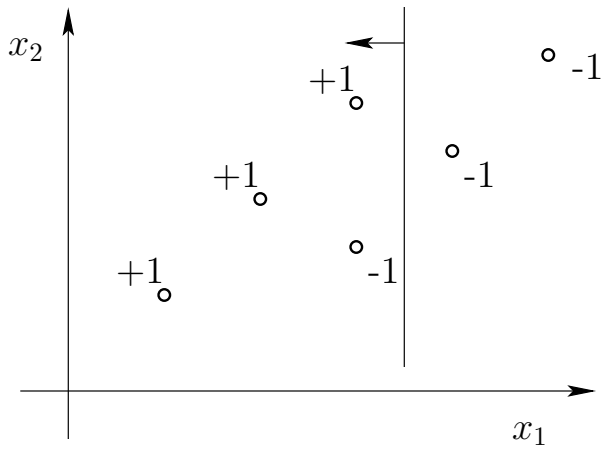
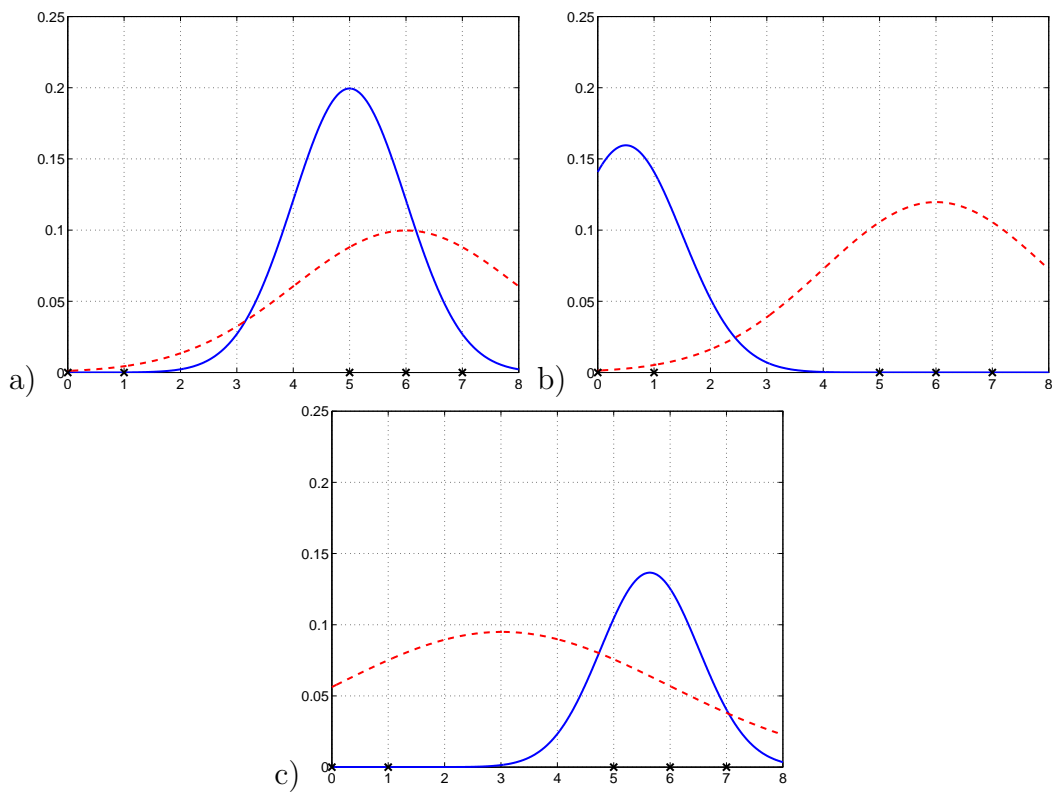


Figure 2



$$P(y|x) : \begin{array}{c|ccc} & x=1 & x=2 & x=3 \\ \hline y=0 & 0.5 & 0.5 & 0.1 \\ y=1 & 0.5 & 0.5 & 0.9 \end{array}$$
