

6.867 Machine learning

FINAL exam

December 10, 2007

(2 points) Your name and MIT ID:

Problem 1

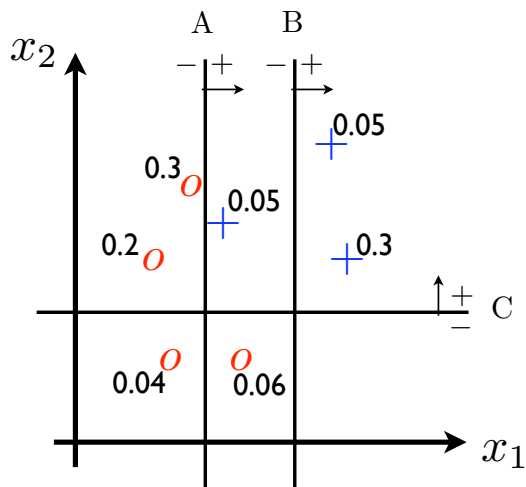


Figure 1.1: Labeled examples, weights on the examples, and three possible stumps.

In Figure 1.1 red 'o' points correspond to negative examples ($y_t = -1$) and blue '+' points are positive examples ($y_t = +1$). The figure also shows the normalized weights on the examples resulting from having run the AdaBoost algorithm for some number of iterations. There are also three decision stumps drawn in the figure, $h(\underline{x}; \theta_A)$, $h(\underline{x}; \theta_B)$, and $h(\underline{x}; \theta_C)$ or A, B and C for short.

- 1.1 (4 points) Which one of the stumps would you use at the next iteration (please answer A, B, or C)?

Briefly justify your answer:

- 1.2 (4 points) Which one of the stumps was used at the previous iteration to obtain the weights on the examples shown in the figure (please answer A, B, or C)?

Briefly justify your answer:

- 1.3 (2 points) In Figure 1.1, circle the training point(s) (possibly none) that the ensemble $h_2(\underline{x}) = h(\underline{x}; \theta_A) + h(\underline{x}; \theta_C)$ cannot classify correctly.

Problem 2

For this problem we are given a training set $D = \{(\underline{x}_1, y_1), \dots, (\underline{x}_n, y_n)\}$ of examples and labels. We have no other data available.

We will use boosting as a feature selection method for an SVM classifier. So, we follow the boosting algorithm for m rounds based on D to get m decision stumps $h(\underline{x}; \hat{\theta}_1), \dots, h(\underline{x}; \hat{\theta}_m)$ (we will drop the “votes” generated by the boosting algorithm). After this we can collect the base classifier predictions into feature vectors

$$\phi(\underline{x}_t) = \frac{1}{\sqrt{m}}[h(\underline{x}_t; \hat{\theta}_1), \dots, h(\underline{x}_t; \hat{\theta}_m)]^T$$

for each training example $t = 1, \dots, n$. Note that $\|\phi(\underline{x}_t)\| = 1$.

To train SVM classifiers based on these feature vectors we will split the dataset D into two equal size sets D_{tr} and D_{te} , and use D_{tr} for training and reserve D_{te} for evaluating the performance of the resulting classifier.

- 2.1 (3 points) Could we use the value of the margin obtained by the hard margin SVM classifiers on D_{tr} as a criterion for selecting between the two kernels below? (Y/N)

$$K_1(\underline{x}, \underline{x}') = \phi(\underline{x})^T \phi(\underline{x}')$$
$$K_2(\underline{x}, \underline{x}') = (1 + \phi(\underline{x})^T \phi(\underline{x}'))^2$$

- 2.2 (6 points) Suppose we train a SVM classifier with kernel $K_1(\underline{x}, \underline{x}')$ based on D_{tr} and evaluate its performance on D_{te} . Does the performance on D_{te} provide a fair measure of how well the classifier is going to work on yet unseen examples (from the same distribution)? Briefly justify your answer.

Problem 3

- 3.1 (6 points) We estimated a mixture of two Gaussians model based on two dimensional data shown in figure 3.1 below. The mixture was initialized randomly in two different

ways and run for three iterations based on each initialization. However, the figures got mixed up (yes, again!). Please draw an arrow from one figure to another to indicate how they follow from each other (you should draw only *four* arrows).

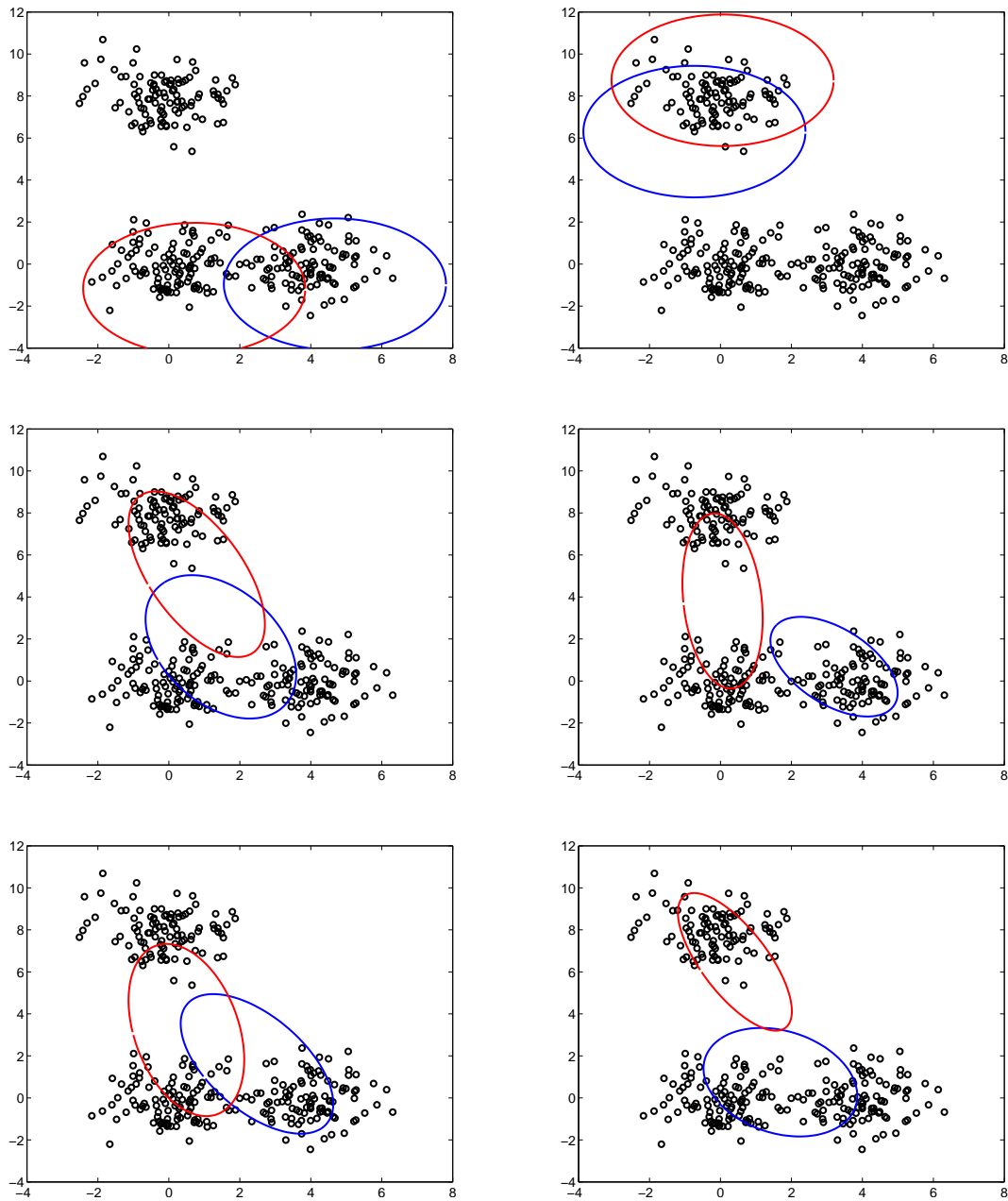


Figure 3.1: mixture model with EM, two initializations, three iterations for each

3.2 (3 points) We also wanted to try another two models based on the same n observations as in 3.1:

$$\text{Model 1} \quad P(\underline{x}; \theta) = N(\underline{x}; \underline{\mu}, \Sigma)$$

$$\text{Model 2} \quad P(\underline{x}; \theta') = P(1)N(\underline{x}; \underline{\mu}_1, \sigma^2 \cdot I) + P(2)N(\underline{x}; \underline{\mu}_2, \sigma^2 \cdot I)$$

You can assume that the parameters are unconstrained to the extent possible (e.g., Σ). How much higher log-likelihood would Model 2 have to assign to the training data for us to select this model with the Bayesian Information Criterion (BIC)?

Problem 4

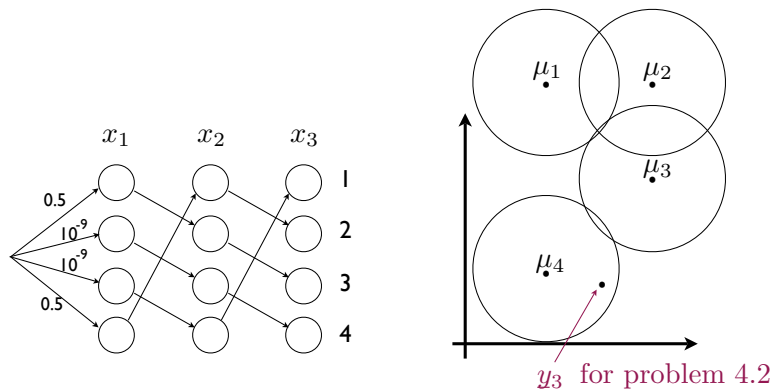


Figure 4.1: HMM with initial state distribution, transitions, and the output distributions.

Consider a homogeneous HMM with four underlying states as illustrated in Figure 4.1. The initial state distribution, permitted state transitions, and the Gaussian output distributions are also shown in the figure. The output distributions $N(\mathbf{y}; \mu_x, \sigma^2 \cdot I)$, $x = 1, 2, 3, 4$, all share the same overall variance parameter σ^2 .

If we were to sample a sequence of two dimensional outputs \mathbf{y} from this HMM model, we would get $\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3, \dots$ (first time point is $t = 1$).

4.1 (6 points) Suppose someone repeatedly sampled such output sequences from this HMM but only revealed to us the output at time $t = 2$ from each sequence. Could we infer the correct number of the underlying states from these observations for this HMM if we didn't know the HMM beforehand? Briefly explain how or why not.

4.2 (4 points) Suppose we only observe y_3 in the figure (at time $t = 3$). What is the most likely hidden state sequence given y_3 ? Briefly justify your answer:

4.3 (4 points) Would the most likely initial state in question 4.2 change if we were to decrease σ^2 ? Briefly justify your answer:

Problem 5

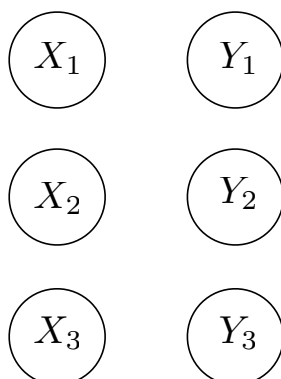


Figure 5.1: Variables (without edges) for drawing the Bayesian network

We would like to extend the HMM model a bit by modifying how the variables are tied together. Figure 5.1 shows a skeleton Bayesian network for this purpose, without any edges.

5.1 **(6 points)** Draw a Bayesian network in Figure 5.1 so that it corresponds to assuming *only* the following two properties:

- (1) Y_1 , Y_2 , and Y_3 are all independent of each other given X_1 , X_2 , and X_3
- (2) X_3 is independent of X_1 given X_2

5.2 **(3 points)** Write down the form of the probability distribution associated with your Bayesian network in Figure 5.1.

5.3 **(4 points)** What counts would we need to compute from *complete observations* in order to estimate the output distribution associated with Y_3 ? Use, e.g., $n(x_1, x_2)$ to describe the number of times we observe $X_1 = x_1$ together with $X_2 = x_2$.

Another set of figures

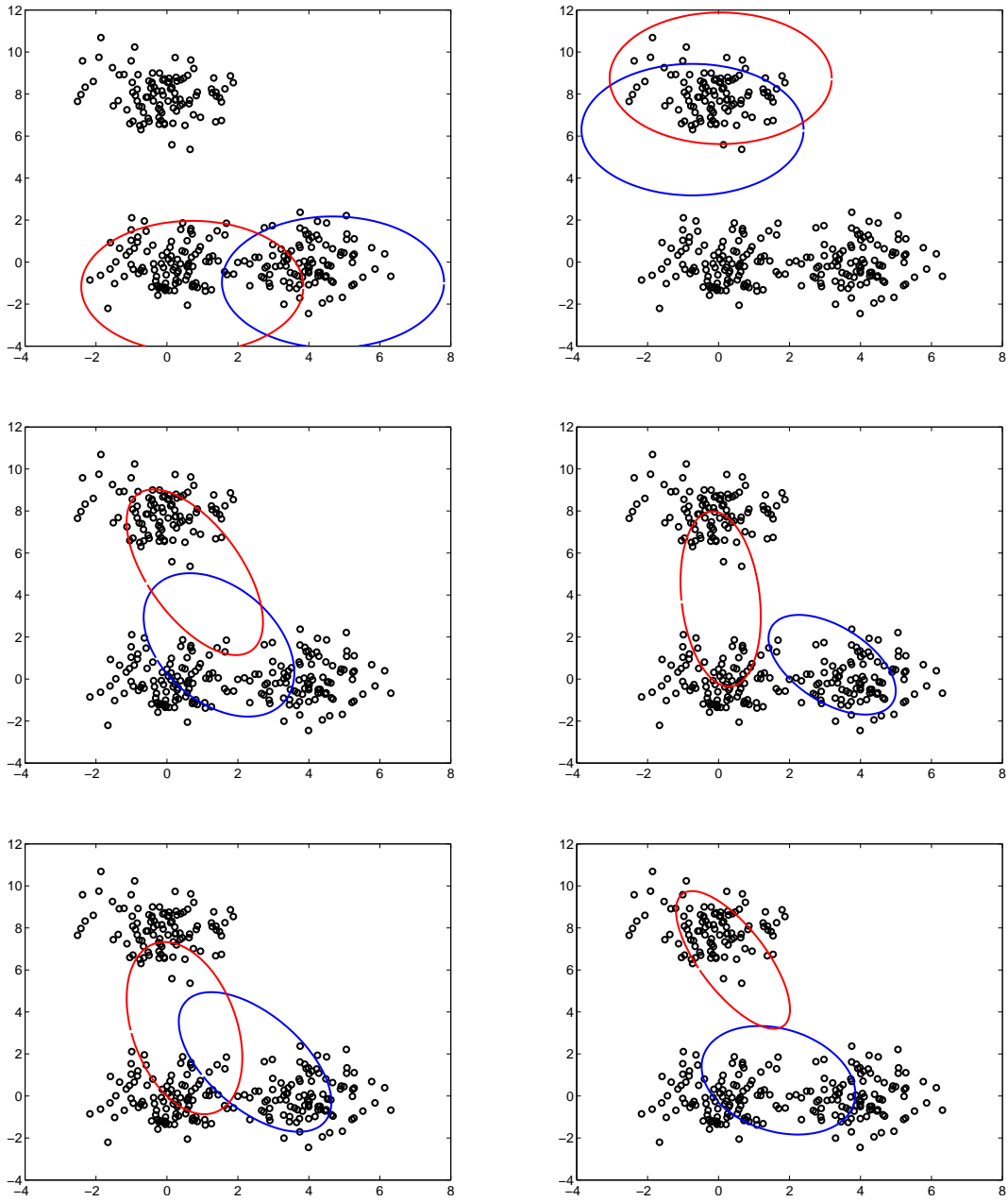


Figure 3.1: mixture model with EM, two initializations, three iterations for each

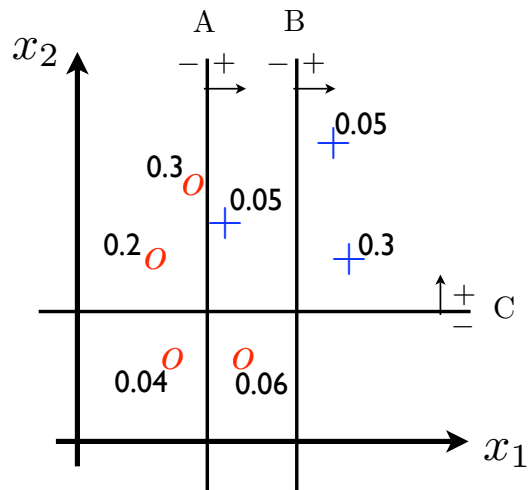


Figure 1.1: Labeled examples, weights on the examples, and three possible stumps.

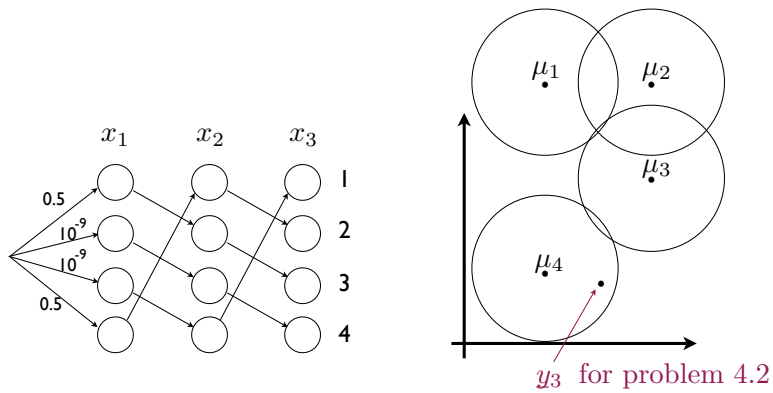


Figure 4.1: HMM with initial state distribution, transitions, and the output distributions.

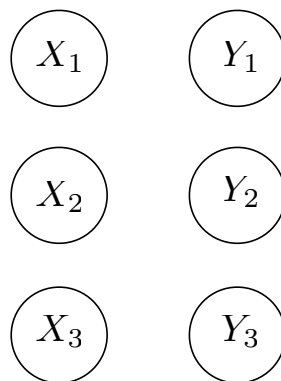


Figure 5.1: Variables (without edges) for drawing the Bayesian network