6.867 Machine learning

Final exam (Fall 2003)

December 10, 2003

Problem 1: your information

1.1. Your name and MIT ID:

1.2. The grade you would give to yourself + brief justification (if you feel that there's no question your grade should be an A, then just say A):

Problem 2

2.1. **(3 points)** Let $\mathcal{F}$ be a set of classifiers whose VC-dimension is 5. Suppose we have four training examples and labels, $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_4, y_4)\}$, and select a classifier $\hat{f}$ from $\mathcal{F}$ by minimizing classification error on the training set. In the absence of any other information about the set of classifiers $\mathcal{F}$, can we say that the prediction $\hat{f}(\mathbf{x}_5)$ for a new example $\mathbf{x}_5$ has any relation to the training set? Briefly justify your answer.

□

2.2. **(T/F – 2 points)** Consider a set of classifiers that includes all linear classifiers that use different choices of strict subsets of the components of the input vectors $\mathbf{x} \in \mathcal{R}^d$. Claim: the VC-dimension of this combined set cannot be more than $d + 1$. □

2.3. **(T/F – 2 points)** Structural risk minimization is based on comparing upper bounds on the generalization error, where the bounds hold with probability $1 - \delta$ over the choice of the training set. Claim: the value of the confidence parameter δ cannot affect model selection decisions. □

2.4. **(6 points)** Suppose we use class-conditional Gaussians to solve a binary classification task. The covariance matrices of the two Gaussians are constrained to be $\sigma^2 I$, where the value of σ^2 is fixed and I is the identity matrix. The only adjustable parameters are therefore the means of the class conditional Gaussians, and the prior class frequencies. We use the maximum likelihood criterion to train the model. Check all that apply.

( ) For any three distinct training points and sufficiently small σ^2, the classifier would have zero classification error on the training set

( ) For any three training points and sufficiently large σ^2, the classifier would always make one classification error on the training set

( ) The classification error of this classifier on the training set is always at least that of a linear SVM, whether the points are linearly separable or not

Problem 3

3.1. (T/F – 2 points) In the AdaBoost algorithm, the weights on all the misclassified points will go up by the same multiplicative factor. □

3.2. (3 points) Provide a brief rationale for the following observation about AdaBoost. The weighted error of the k^{th} weak classifier (measured relative to the weights at the beginning of the k^{th} iteration) tends to increase as a function of the iteration k.

Consider a text classification problem, where documents are represented by binary (0/1) feature vectors $\phi = [\phi_1, \ldots, \phi_m]^T$; here ϕ_i indicates whether word i appears in the document. We define a set of weak classifiers, $h(\phi; \theta) = y\phi_i$, parameterized by $\theta = \{i, y\}$ (the choice of the component, $i \in \{1, \ldots, m\}$, and the class label, $y \in \{-1, 1\}$, that the component should be associated with). There are exactly $2m$ possible weak learners of this type.

We use this boosting algorithm for feature selection. The idea is to simply run the boosting algorithm and select the features or components in the order in which they were identified by the weak learners. We assume that the boosting algorithm finds the best available weak classifier at each iteration.

3.3. (T/F – 2 points) The boosting algorithm described here can select the exact same weak classifier more than once. □

3.4. (4 points) Is the ranking of features generated by the boosting algorithm likely to be more useful for a linear classifier than the ranking from simple mutual information calculations (estimates $\hat{I}(y; \phi_i)$). Briefly justify your answer.

Problem 4

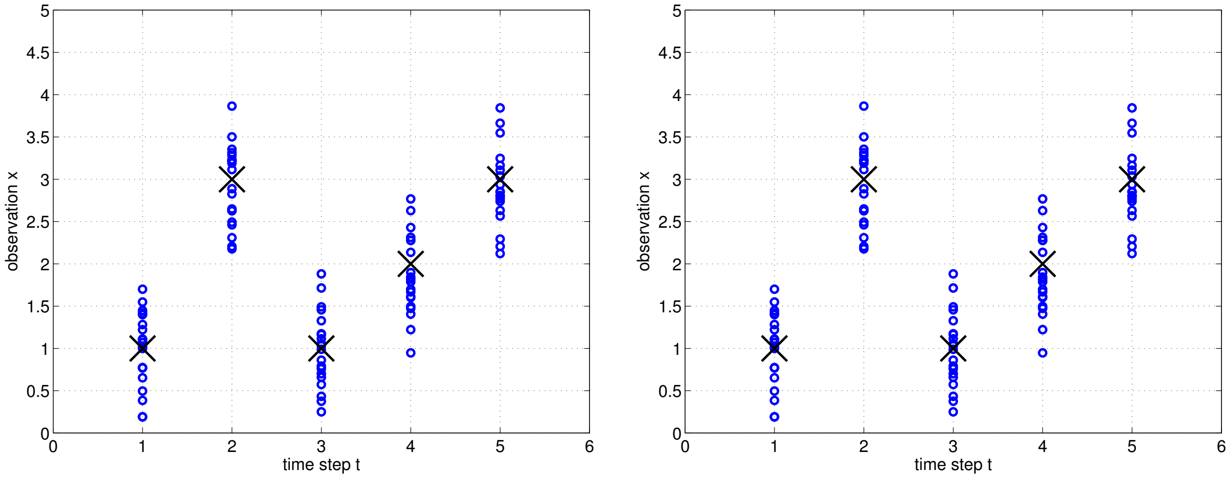


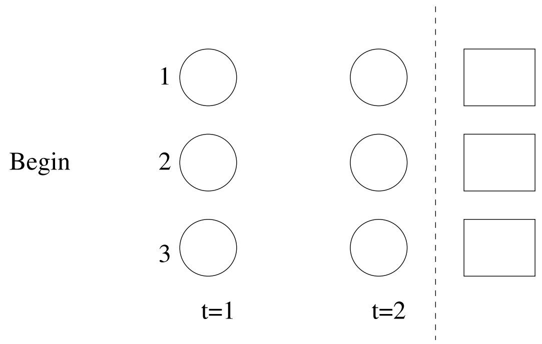
Figure 1a)

Figure 1b)

Figure 1: Time dependent observations. The data points in the figure are generated as sets of five consecutive time dependent observations, $x_1, \ldots, x_5$. The clusters come from repeatedly generating five consecutive samples. Each visible cluster consists of 20 points, and has approximately the same variance. The mean of each cluster is shown with a large X.

Consider the data in Figure 1 (see the caption for details). We begin by modeling this data with a three state HMM, where each state has a Gaussian output distribution with some mean and variance (means and variances can be set independently for each state).

4.1. (4 points) Draw the state transition diagram and the initial state distribution for a three state HMM that models the data in Figure 1 in the maximum likelihood sense. Indicate the possible transitions and their probabilities in the figure below (whether or not the state is reachable after the first two steps). In order words, your drawing should characterize the 1st order homogeneous Markov chain govering the evolution of the states. Also indicate the means of the corresponding Gaussian output distributions (please use the boxes).



4.2. (4 points) In Figure 1a draw as ovals the clusters of outputs that would form if we repeatedly generated samples from your HMM over time steps $t = 1, \ldots, 5$. The height of the ovals should reflect the variance of the clusters.

4.3. (4 points) Suppose at time $t = 2$ we observe $x_2 = 1.5$ but don't see the observations for other time points. What is the most likely state at $t = 2$ according to the marginal posterior probability $\gamma_2(s)$ defined as $P(s_2 = s|x_2 = 1.5)$.

□

4.4. (2 points) What would be the most likely state at $t = 2$ if we also saw $x_3 = 0$ at $t = 3$? In this case $\gamma_2(s) = P(s_2 = s|x_2 = 1.5, x_3 = 0)$.

□

4.5. (4 points) We can also try to model the data with conditional mixtures (mixtures of experts), where the conditioning is based on the time step. Suppose we only use two experts which are linear regression models with additive Gaussian noise, i.e.,

$$P(x|t, \theta_i) = \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\Big\{ -\frac{1}{2\sigma_i^2}(x - \theta_{i0} - \theta_{i1}t)^2 \Big\}$$

for $i = 1, 2$. The gating network is a logistic regression model from t to binary selection of the experts. Assuming your estimation of the conditional mixture model is successfully in the maximum likelihood sense, draw the resulting mean predictions of the two linear regression models as a function of time t in Figure 1b). Also, with a vertical line, indicate where the gating network would change it's preference from one expert to the other.

4.6. (T/F – 2 points) Claim: by repeatedly sampling from your conditional mixture model at successive time points $t = 1, 2, 3, 4, 5$, the resulting samples would resemble the data in Figure 1

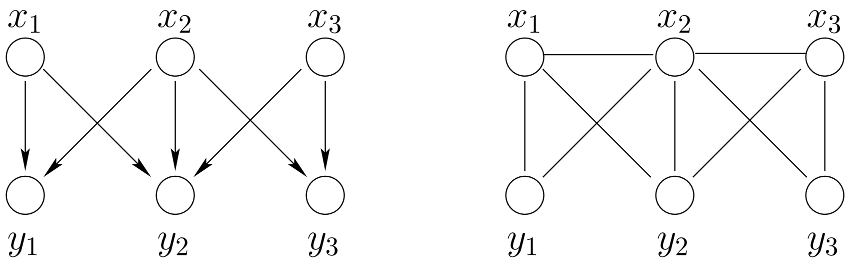
□

.

4.7. (4 points) Having two competing models for the same data, the HMM and the mixture of experts model, we'd like to select the better one. We think that any reasonable model selection criterion would be able to select the better model in this case. Which model would we choose? Provide a brief justification.

□

Problem 5



a) Bayesian network (directed) b) Markov random field (undirected)

Figure 2: Graphical models

5.1. (2 points) List two different types of independence properties satisfied by the Bayesian network model in Figure 2a.

5.2. (2 points) Write the factorization of the joint distribution implied by the directed graph in Figure 2a.

5.3. (2 points) Provide an alternative factorization of the joint distribution, different from the previous one. Your factorization should be consistent with all the properties of the directed graph in Figure 2a. Consistency here means: whatever is implied by the graph should hold for the associated distribution.

5.4. (4 points) Provide an independence statement that holds for the undirected model in Figure 2b but does NOT hold for the Bayesian network. Which edge(s) should we add to the undirected model so that it would be consistent with (wouldn't imply anything that is not true for) the Bayesian network?

5.5. (2 points) Is your resulting undirected graph triangulated (Y/N)?

5.6. (4 points) Provide two directed graphs representing 1) a mixture of two experts model for classification, and 2) a mixture of Gaussians classifiers with two mixture components per class. Please use the following notation: $\mathbf{x}$ for the input observation, y for the class, and i for any selection of components.

Additional set of figures

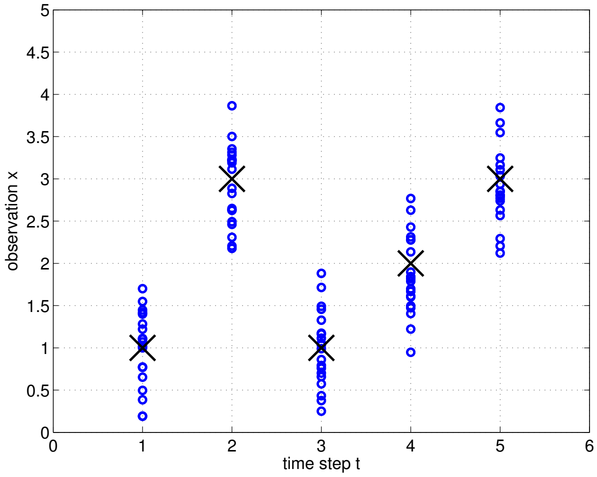


Figure 1a)

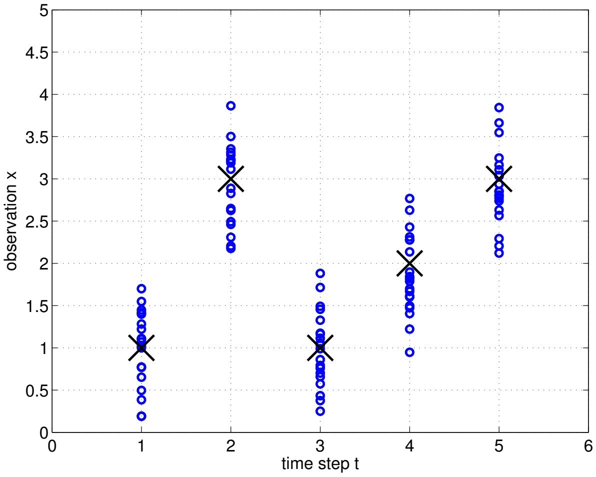
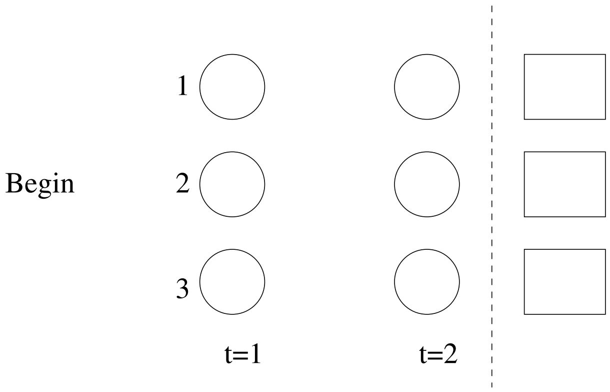
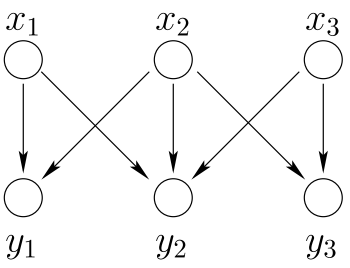
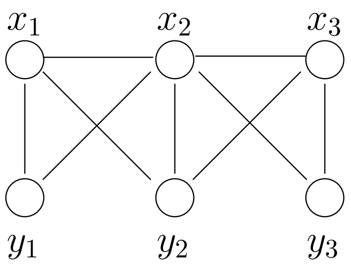


Figure 1b)





a) Bayesian network (directed)



b) Markov random field (undirected)