6.867 Machine learning and neural networks

FALL 2001 – Final exam

December 11, 2001

(2 points) Your name and MIT ID #:

(4 points) The grade you would give to yourself + brief justification. If you feel that there's no question that your grade should be A (and you feel we agree with you) then just write "A".

Problem 1

1. **(T/F – 2 points)** The sequence of output symbols sampled from a hidden Markov model satisfies the first order Markov property

2. **(T/F – 2 points)** Increasing the number of values for the the hidden states in an HMM has much greater effect on the computational cost of forward-backward algorithm than increasing the length of the observation sequence.

3. **(T/F – 2 points)** In HMMs, if there are at least two distinct most likely hidden state sequences and the two state sequences cross in the middle (share a single state at an intermediate time point), then there are at least four most likely state sequences.

4. **(T/F – 2 points)** One advantage of Boosting is that it does not overfit.

5. **(T/F – 2 points)** Support vector machines are resistant to outliers, i.e., very noisy examples drawn from a different distribution.

6. **(T/F – 2 points)** Active learning can substantially reduce the number of training examples that we need.

Problem 2

Consider two classifiers: 1) an SVM with a quadratic (second order polynomial) kernel function and 2) an unconstrained mixture of two Gaussians model, one Gaussian per class label. These classifiers try to map examples in $\mathcal{R}^2$ to binary labels. We assume that the problem is separable, no slack penalties are added to the SVM classifier, and that we have sufficiently many training examples to estimate the covariance matrices of the two Gaussian components.

1. **(T/F – 2 points)** The two classifiers have the same VC-dimension.

2. **(4 points)** Suppose we evaluated the structural risk minimization score for the two classifiers. The score is the bound on the expected loss of the classifier, when the classifier is estimated on the basis of n training examples. Which of the two classifiers might yield the better (lower) score? Provide a brief justification.

3. **(4 points)** Suppose now that we regularize the estimation of the covariance matrices for the mixture of two Gaussians. In other words, we would estimate each class conditional covariance matrix according to

$$\hat{\Sigma}_{reg} = \frac{n}{n+n'}\hat{\Sigma} + \frac{n'}{n+n'}S \tag{1}$$

where n is the number of training examples, $\hat{\Sigma}$ is the unregularized estimate of the covariance matrix (sample covariance matrix of the examples in one class), S is our prior covariance matrix (same for both classes), and n' the equivalent sample size that we can use to balance between the prior and the data.

In computing the VC-dimension of a classifier, we can choose the set of points that we try to "shatter". In particular, we can scale any k points by a large factor and use the resulting set of points for shattering. In light of this, would you expect our regularization to change the VC-dimension? Why or why not?

4. **(T/F – 2 points)** Regularization in the above sense would improve the structural risk minimization score for the mixture of two Gaussians.

Problem 3

The problem here is to predict the identity of a person based on a single handwritten character. The observed characters (one of 'a', 'b', or 'c') are transformed into binary 8 by 8 pixel images. There are four different people we need to identify on the basis of such characters. To do this, we have a training set of about 200 examples, where each example consists of a binary 8x8 image and the identity of the person it belongs to. You can assume that the overall number of occurences of each person and each character in the training set is roughly balanced.

We would like to use a mixture of experts architecture to solve this problem.

1. **(2 points)** How might the experts be useful ? Suggest what task each expert might solve.

2. **(4 points)** Draw a graphical model that describes the mixture of experts architecture in this context. Indicate what the variables are and the number of values that they can take. Shade any nodes corresponding to variables that are always observed.

3. **(4 points)** Before implementing the mixture of experts architecture, we need to know the parametric form of the conditional probabilities in your graphical model. Provide a *reasonable* specification of the relevant conditional probabilities to the extent that you could ask your class-mate to implement the classifier.

4. **(4 points)** So we implemented your method, ran the estimation algorithm once, and measured the test performance. The method was unfortunately performing at a chance level. Provide *two* possible explanations for this. (there may be multiple correct answers here)

5. **(3 points)** Would we get anything reasonable out of the estimation if, initially, all the experts were identical while the parameters of the gating network would be chosen randomly? By reasonable we mean training performance. Provide a brief justification.

6. **(3 points)** Would we get anything reasonable out the estimation if now the gating network is initially set so that it assigns each training example uniformly to all experts but the experts themselves are initialized with random parameter values? Again, reasonable refers to the training error. Provide a brief justification.

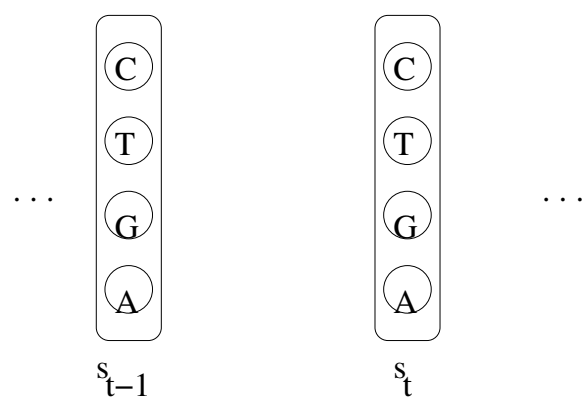
Problem 4

Consider the following pair of observed sequences:

Sequence 1 (s_t):	A	A	T	T	G	G	C	C	A	A	T	T	G	G	C	C	...
Sequence 2 (x_t):	1	1	2	2	1	1	2	2	1	1	2	2	1	1	2	2	...
Position t:	0	1	2	3	4	...											

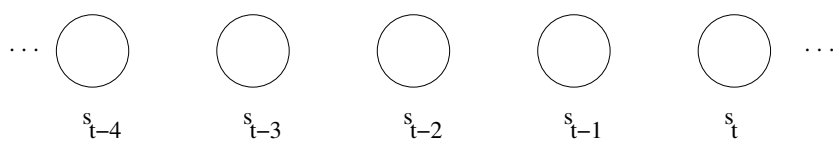
where we assume that the pattern (highlighted with the spaces) will continue forever. Let $s_t \in \{A, G, T, C\}$, $t = 0, 1, 2, \ldots$ denote the variables associated with the first sequence, and $x_t \in \{1, 2\}$, $t = 0, 1, 2, \ldots$ the variables characterizing the second sequence. So, for example, given the sequences above, the observed values for these variables are $s_0 = A, s_1 = A, s_2 = C, \ldots$, and, similarly, $x_0 = 1, x_1 = 1, x_2 = 2, \ldots$.

1. **(4 points)** If we use a simple first order homogeneous markov model to predict the first sequence (values for s_t only), what is the maximum likelihood solution that we would find? In the *transition diagram* below, please draw the relevant transitions and the associated probabilities (this should not require much calculation)

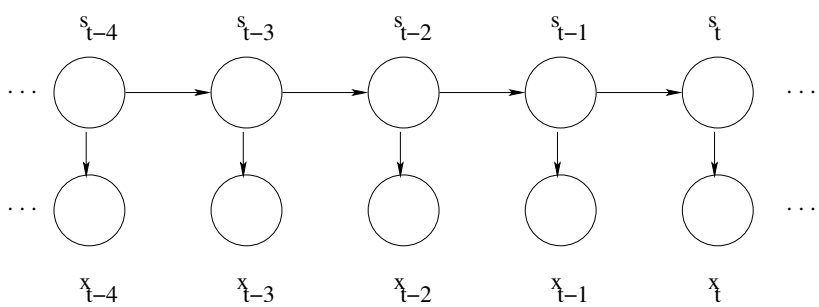


2. **(T/F – 2 points)** The resulting first order Markov model is *ergodic*

3. **(4 points)** To improve the Markov model a bit we would like to define a graphical model that predicts the value of s_t on the basis of the previous observed values $s_{t-1}, s_{t-2}, \ldots$ (looking as far back as needed). The model parameters/structure are assumed to remain the same if we shift the model one step. In other words, it is the same graphical model that predicts s_t on the basis of $s_{t-1}, s_{t-2}, \ldots$ as the model that predicts s_{t-1} on the basis of $s_{t-2}, s_{t-3}, \ldots$. In the graph below, draw the *minimum number of arrows* that are needed to predict the first observed sequence perfectly (disregarding the first few symbols in the sequence). Since we slide the model along the sequence, you can draw the arrows only for s_t.



4. Now, to incorporate the second observation sequence, we will use a standard hidden Markov model:



where again $s_t \in \{A, G, T, C\}$ and $x_t \in \{1, 2\}$. We will estimate the parameters of this HMM in two different ways.

(I) Treat the pair of observed sequences (s_t, x_t) (given above) as complete observations of the variables in the model and estimate the parameters in the maximum likelihood sense. The initial state distribution $P_0(s_0)$ is set according to the overall frequency of symbols in the first observed sequence (uniform).

(II) Use only the second observed sequence (x_t) in estimating the parameters, again in the maximum likelihood sense. The initial state distribution is again uniform across the four symbols.

We assume that both estimation processes will be successfull relative to their criteria.

a) **(3 points)** What are the observation probabilities $P(x|s)$ ($x \in \{1, 2\}$, $s \in \{A, G, T, C\}$) resulting from the first estimation approach? (should not require much calculation)

b) **(3 points)** Which estimation approach is likely to yield a more accurate model over the second observed sequence (x_t)? Briefly explain why.

5. Consider now the two HMMs resulting from using each of the estimation approaches (approaches I and II above). These HMMs are estimated on the basis of the pair of observed sequences given above. We'd like to evaluate the probability that these two models assign to a new (different) observation sequence `1 2 1 2`, i.e., $x_0 = 1, x_1 = 2, x_2 = 1, x_3 = 2$. For the first model, for which we have some idea about what the s_t variables will capture, we also want to know the the associated most likely hidden state sequence. (these should not require much calculation)

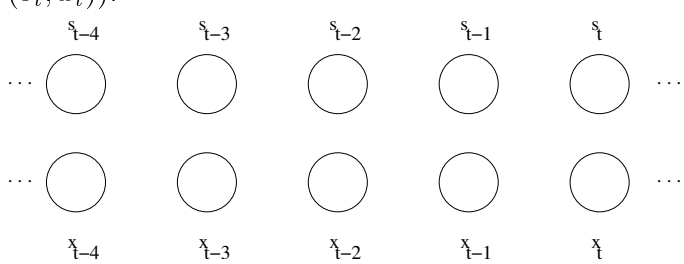
a) **(2 points)** What is the probability that the first model (approach I) assigns to this new sequence of observations?

b) **(2 points)** What is the probability that the second model (approach II) gives to the new sequence of observations?

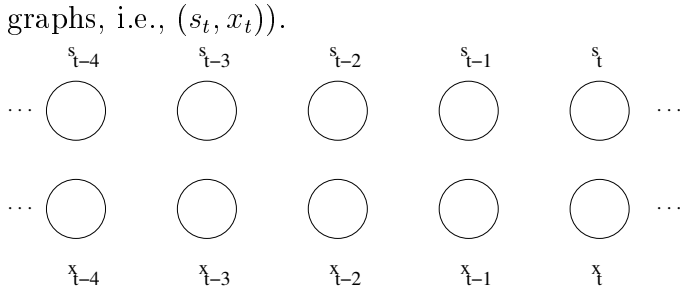
c) **(2 points)** What is the most likely hidden state sequence in the first model (from approach I) associated with the new observed sequence?

6. **(4 points)** Finally, let's assume that we observe only the second sequence (x_t) (the same sequence as given above). In building a graphical model over this sequence we are no longer limiting ourselves to HMMs. However, we only consider models whose structure/parameters remain the same as we slide along the sequence. The variables s_t are included as before as they might come handy as hidden variables in predicting the observed sequence.

a) In the figure below, draw the arrows that any reasonable model selection criterion would find given an unlimited supply of the observed sequence $x_t, x_{t+1}, \ldots$. You only need to draw the arrows for the last pair of variables in the graphs, i.e., (s_t, x_t)).



b) Given only a small number of obervations, the model selection criterion might select a different model. In the figure below, indicate a possible alternate model that any reasonable model selection criterion would find given only a few examples. You only need to draw the arrows for the last pair of variables in the graphs, i.e., (s_t, x_t)).



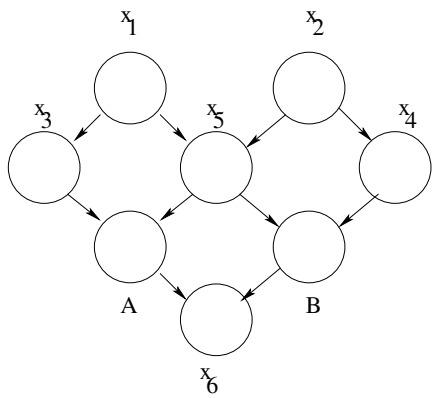


Figure 1: Decision makers A and B and their "context"

Problem 5

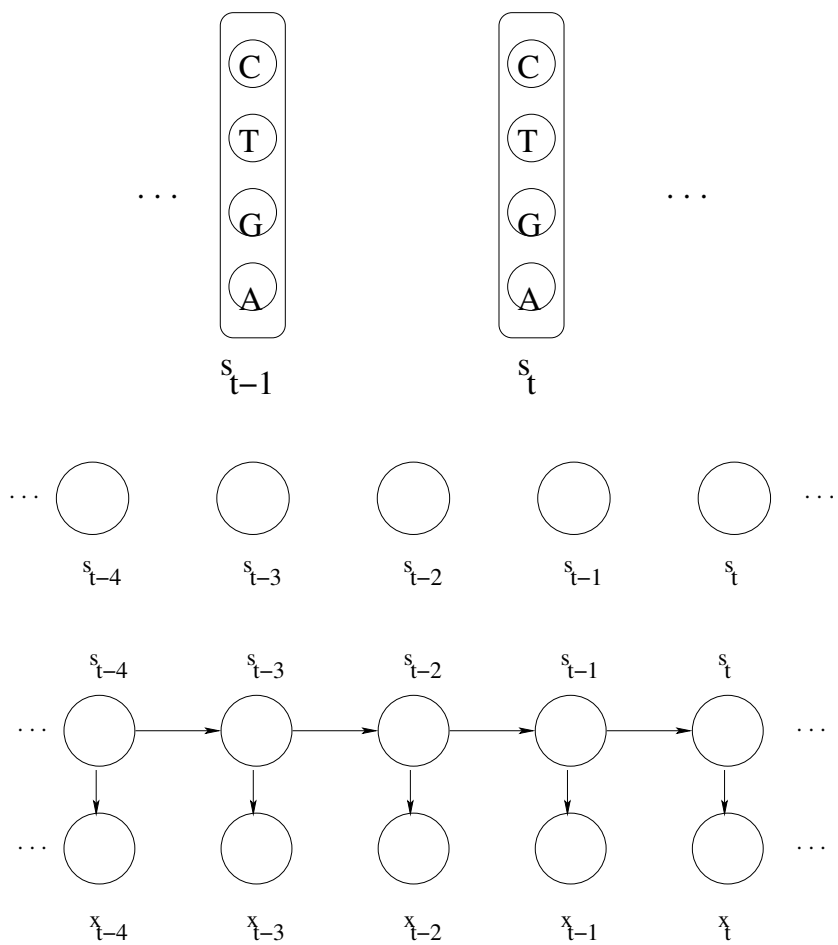
The Bayesian network in figure 1 claims to model how two people, call them A and B, make decisions in different contexts. The context is specified by the setting of the binary context variables $x_1, x_2, \ldots, x_6$. The values of these variables are not known unless we specifically ask for such values.

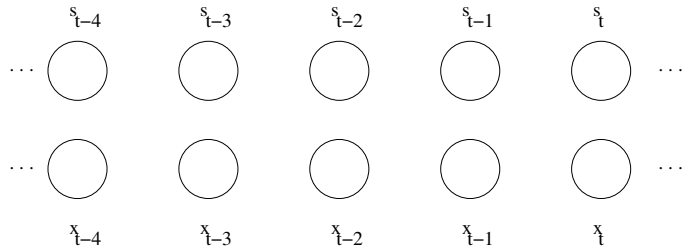
1. **(6 points)** We are interested in finding out what information we'd have to acquire to ensure that A and B will make their decisions independently from one another. Specify the *smallest set of context variables* whose instantiation would render A and B independent. Briefly explain your reasoning (there may be more than one strategy for arriving at the same decision)

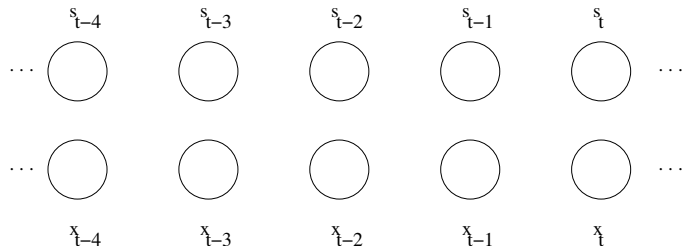
2. **(T/F – 2 points)** We can in general achieve independence with less information, i.e., we don't have to fully instantiate the selected context variables but provide some evidence about their values

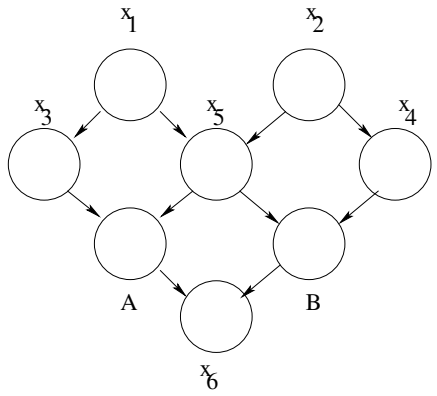
3. **(4 points)** Could your choice of the minimal set of context variables change if we also provided you with the actual probability values associated with the dependencies in the graph? Provide a brief justification.

Additional set of figures




s_{t-4}
s_{t-3}
s_{t-2}
s_{t-1}
s_t
x_{t-4}
x_{t-3}
x_{t-2}
x_{t-1}
x_t


s_{t-4}
s_{t-3}
s_{t-2}
s_{t-1}
s_t
x_{t-4}
x_{t-3}
x_{t-2}
x_{t-1}
x_t


x_1
x_2
x_3
x_5
x_4
A
B
x_6