6.867 Machine learning

Final exam

December 5, 2002

(2 points) Your name and MIT ID:

J Doe, #000

(4 points) The grade you would give to yourself + a brief justification:

A or perhaps A- if there are any typos or other errors in the solutions...

Problem 1

We wish to estimate a mixture of two experts model for the data displayed in Figure 1. The experts we can use here are linear regression models of the form

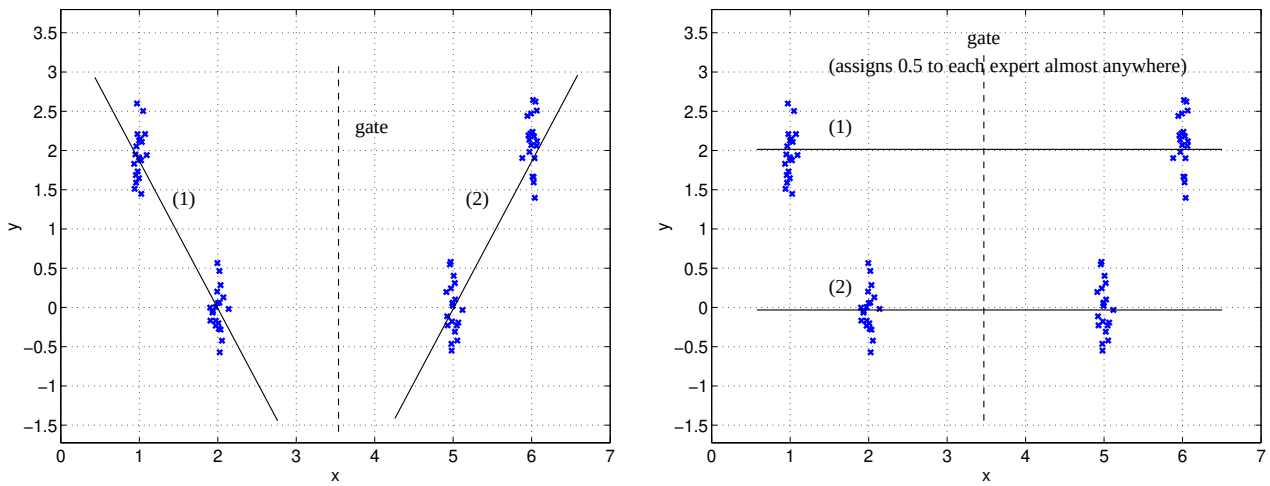
$$p(y|x, \mathbf{w}) = N(\, y;\, w_1 x + w_0,\, \sigma^2\,)$$

where $N(y; \mu, \sigma^2)$ denotes a Gaussian distribution over y with mean μ and variance σ^2. Each expert i can choose its parameters $\mathbf{w}_i = [w_{i0}, w_{i1}]^T$ and σ_i^2 independently from other experts. Note that the first subindex i in w_{ij} refers to the expert.

The gating network in the case of two experts is given by a logistic regression model

$$P(\text{expert} = 1|x, \mathbf{v}) = g(\, v_1 x + v_0\,)$$

where $g(z) = (1 + \exp(-z))^{-1}$ and $\mathbf{v} = [v_0, v_1]^T$.



a) unregularized case b) regularized case

Figure 1: Data for mixtures of experts

1. **(4 points)** Suppose we estimate a mixture of two experts model based on the data in Figure 1. You can assume that the estimation is successful in the sense that we will find a setting of the parameters that maximizes the log-likelihood of the data. Please indicate (approximately) in Figure 1a) the mean predictions from the two experts as well as the decision boundary for the gating network. Label the mean predictions – functions of x – with "(1)" and "(2)" corresponding to the two experts, and the decision boundary with "gate".

2. **(4 points)** We now switch to a regularized maximum likelihood objective by incorporating the following regularization penalty

$$-\frac{c}{2}(w_{11}^2 + w_{21}^2)$$

into the log-likelihood objective. Note that the penalty includes only one parameter from each of the experts. By increasing c, we impose a stronger penalty. Similarly to the previous question, please indicate in Figure 1b) the optimal regularized solution for the mixture of two experts model when the regularization parameter c is set to a very large value.

3. **(3 points)** Are the variances in the predictive Gaussian distributions of the experts

( ) larger,

( ) smaller,

(x) about the same

after the regularization?

s	1	2
P(x=1)	0	0.1
P(x=2)	0.199	0
P(x=3)	0.8	0.7
P(x=4)	0.001	0.2

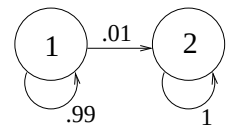


Figure 2: A two-state HMM for Problem 2

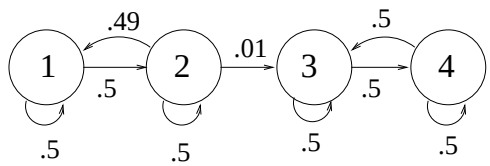


Figure 3: An alternative, four-state HMM for Problem 2

Problem 2

Figure 2 shows a two-state HMM. The transition probabilities of the Markov chain are given in the transition diagram. The output distribution corresponding to each state is defined over $\{1, 2, 3, 4\}$ and is given in the table next to the diagram. The HMM is equally likely to start from either of the two states.

1. **(3 points)** Give an example of an output sequence of length 2 which can not be generated by the HMM in Figure 2. | 1,2 |

2. **(2 points)** We generated a sequence of $6,867^{2002}$ observations from the HMM, and found that the last observation in the sequence was 3. What is the most likely hidden state corresponding to that last observation? | 2 |

3. **(2 points)** Consider an output sequence 3 3. What is the most likely sequence of hidden states corresponding to these observations? | 1,1 |

4. **(2 points)** Now, consider an output sequence 3 3 4. What are *the first two states* of the most likely hidden state sequence? | 2,2 |

5. **(4 points)** We can try to increase the modeling capacity of the HMM a bit by breaking each state into two states. Following this idea, we created the diagram in Figure 3. Can we set the initial state distribution and the output distributions so that this 4-state model, with the transition probabilities indicated in the diagram, would be equivalent to the original 2-state model? If yes, how? If no, why not?

No we cannot. First note that we have to associate the first two states in the 4-state model with state 1 of the 2-state model. The probability of leaving the first two states in the 4-state model, however, depends on time (whether the chain happens to be in state 1 or 2). In contrast, in the 2-state model the probability of transitioning to 2 is always 0.01.

6. **(T/F – 2 points)** The Markov chain in Figure 3 is ergodic

F

Problem 3

Figure 4 shows a graphical model over four binary valued variables, $x_1, \ldots, x_4$. We do not know the parameters of the probability distribution associated with the graph.

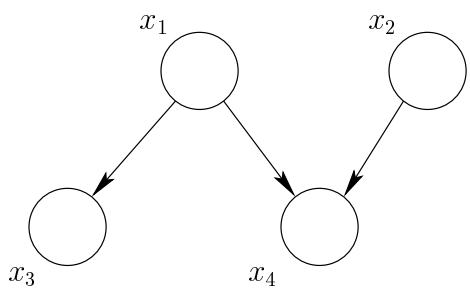


Figure 4: A graphical model

1. **(2 points)** Would it typically help to know the value of x_3 so as to gain more information about x_2? (please answer **yes** or **no**) | no |

2. **(2 points)** Assume we already know the value of x_4. Would it help in this case to know the value of x_3 to gain more information about x_2? (please answer **yes** or **no**) | yes |

3. **(3 points)** List three different conditional independence statements between the four variables that can be inferred from the graph. You can include marginal independence by saying "given nothing".

   (a) (x_1) is independent of (x_2) given (nothing)

   (b) (x_3) is independent of (x_2) given (nothing)

   (c) (x_3) is independent of (x_4) given (x_1)

4. **(2 points)** The following table gives a possible partial specification of the conditional probability $P(x_4|x_1, x_2)$ associated with the graph. Fill in the missing values so that we could omit the arrow $x_1 \rightarrow x_4$ in the graph and the graph would still adequately represent the probability distribution.

$P(x_4 = 1\|x_1 = 0, x_2 = 0)$	0.8
$P(x_4 = 1\|x_1 = 0, x_2 = 1)$	0.4
$P(x_4 = 1\|x_1 = 1, x_2 = 0)$	**0.8**
$P(x_4 = 1\|x_1 = 1, x_2 = 1)$	**0.4**

5. **(4 points)** Let's again focus on the original graph in Figure 4. Since we don't know the underlying probability distribution, we need to estimate it from observed data. Unfortunately, the dataset we have is incomplete and contains only observations for x_2 and x_3. In other words, the dataset is $D = \{(x_2^t, x_3^t), t = 1, \ldots, n\}$. In the joint distribution

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

we have a number of components (smaller probability tables) that need to be estimated. Please indicate which components we can hope to estimate (adjust) on the basis of the available data?

(x) $P(x_1)$

(x) $P(x_2)$

(x) $P(x_3|x_1)$

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

6. **(4 points)** If we use the EM algorithm to carry out the estimation task, what posterior probabilities do we have to evaluate in the E-step? Please provide the necessary posterior probabilities in the form $P(\cdots|x_2^t, x_3^t)$.

We only need to evaluate $P(x_1|x_3^t)$.
Since x_4 is never observed, x_1 and x_2 are always independent. We thus have to estimate only two independent parts $x_1 \to x_3$ and x_2, where x_1 is unobserved. In the EM-algorithm we need to fill-in the missing values for x_1, and thus evaluate $P(x_1|x_3^t)$

Problem 4

We try to select here between two models. Both models are logistic regression models but differ in terms of the type of features that are used in making the predictions. More specifically, the models have the common squashed additive form

$$P(y = 1|x, \mathbf{w}) = g(\, w_0 + w_1\phi_1(x) + \ldots + \phi_m(x)\,)$$

where the input is a real number $x \in \mathcal{R}$. The models differ in terms of the number and the type of basis functions used:

$$\begin{aligned} \text{model 1} \quad &: \quad m = 1,\ \phi_1(x) = x \\ \text{model 2} \quad &: \quad m = 2,\ \phi_1(x) = x,\ \phi_2(x) = \sin(x) \end{aligned}$$

1. **(4 points)** Suppose we have n training examples (x^t, y^t), $t = 1, \ldots, n$, and we evaluate structural risk minimization scores (bounds on the generalization error) for the two classifiers. Which of the following statements are valid in general for our two models:

   ( ) Score for model 1 $\geq$ Score for model 2

   ( ) Score for model 1 $\leq$ Score for model 2

   ( ) Score for model 1 $=$ Score for model 2

   (x) Each of the above three cases may be correct depending on the data

   ( ) None of the above

2. **(4 points)** We will now switch to the Bayesian information criterion (BIC) for selecting among the two models. Let $L_1(n)$ be the log-probability of the labels that model 1 assigns to n training labels, where the probabilities are evaluated at the maximum likelihood setting of the parameters. Let $L_2(n)$ be the corresponding log-probability for model 2. We imagine here that $L_1(n)$ and $L_2(n)$ are evaluated on the basis of the first n training examples from a much larger set.

   Now, in our empirical studies, we found that these log-probabilities are related in a simple way:

$$L_2(n) - L_1(n) \approx 0.01 \cdot n$$

   How will we end up selecting between the two models as a function of the number of training examples? Please choose one of the following cases.

   ( ) Always select 1

   ( ) Always select 2

   (x) First select 1, then 2 for larger n

   ( ) First select 2, then 1 for larger n

3. **(4 points)** Provide a brief justification for your answer to the previous question.

For large n we would select model 2 since it has a consistent (albeit small) advantage. Initially, however, we would choose model 1 due to smaller complexity penalty.

To see this a bit more precisely, let's recall the form of the BIC score: for model 1 it is defined as $BIC_1 = L_1(n) - \frac{d_1}{2}\log(n)$, where d_1 is the number of parameters in model 1. The difference between the BIC scores for the two models is therefore

$$\begin{aligned} BIC_2 - BIC_1 &= \overbrace{L_2(n) - L_1(n)}^{0.01 \cdot n} - \frac{\overbrace{d_2 - d_1}^{3-2}}{2}\log(n) \\ &= 0.01 \cdot n - \frac{1}{2}\log(n) \end{aligned}$$

When n is small the complexity term dominates and $BIC_2 < BIC_1$ (the difference is negative). For large n the linear increase of the log-likelihood difference overcomes the logarithmic penalty and $BIC_2 > BIC_1$.

Problem 5

Consider a simple two-class document (text) classification problem. Each document is represented by a binary feature vector $[\phi_1^i, \ldots, \phi_N^i]$, where $\phi_k^i = 1$ if keyword k is present in the document, and zero otherwise. N is the number of keywords we have chosen to include.

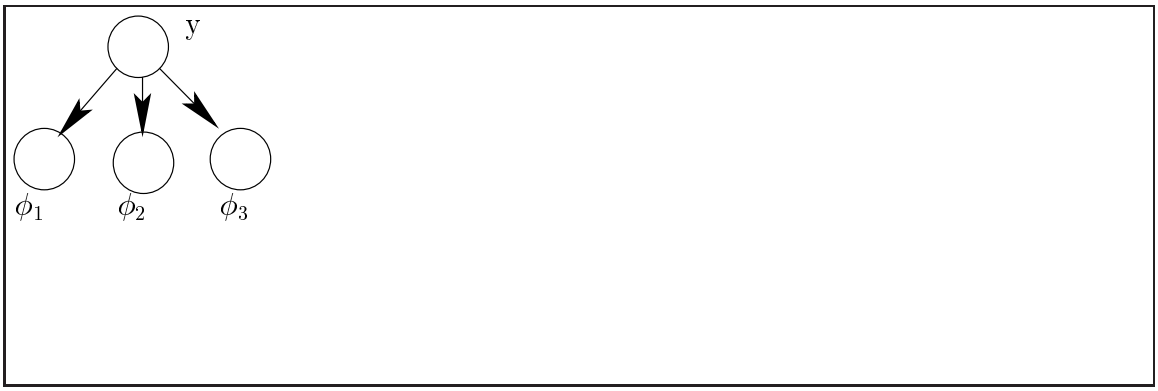
We use a Naive Bayes model for this classification task. The joint distribution of the features and the binary labels $y \in \{0, 1\}$ is in this case given by

$$P(\phi_1, \ldots, \phi_N, y) = P(y) \prod_{k=1}^{N} P(\phi_k | y)$$

where, for example,

$$P(\phi_k = 1 | y = 0) = \theta_{k,0}, \quad P(\phi_k = 1 | y = 1) = \theta_{k,1}$$

1. **(2 points)** In the space below, draw the graphical model corresponding to Naive Bayes generative model described above. Assume that $N = 3$ (three keywords).



2. **(4 points)** To be able to make use of training examples with possibly missing labels, we will have to resort to the EM algorithm. In the EM algorithm we need to evaluate the posterior probability of the label y given the document. We will use a message passing algorithm (belief propagation) to get this posterior probability. The problem here is that we relied on a rather careless friend to evaluate whether a document contains any of the keywords. In other words, we do not fully trust the "observed" values of the features. Let $\hat{\phi}_k$ be the "observed" value for the kth feature in a given document. The evidence we now have about the actual value of ϕ_k is given by $P(\hat{\phi}_k | \phi_k)$, which is a table that models how we expect the friend to respond.

   Given that we observe $\hat{\phi}_k = 1$, what is the message that ϕ_k needs to send to y?

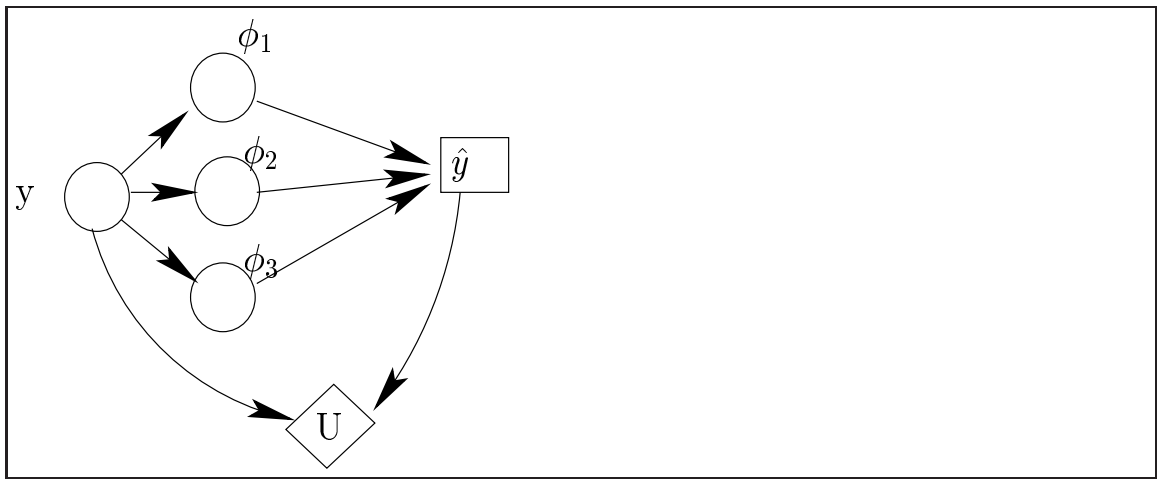
There are a couple of options here depending on how you choose to incorporate the evidence $\hat{\phi}_k = 1$. Perhaps the simplest way is to put it directly in the single node potential $\psi_k(\phi_k)$. In other words, we define $\psi_k(\phi_k) = P(\hat{\phi}_k = 1|\phi_k)$, and $\psi_{ky}(\phi_k, y) = P(\phi_k|y)$. In this notation, the message is

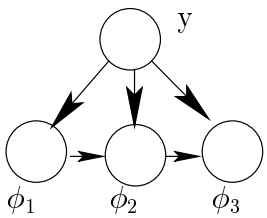
$$m_{k\to y}(y) = \sum_{\phi_k=0,1} \psi_k(\phi_k)\psi_{ky}(\phi_k, y)$$

3. **(3 points)** We know that the EM algorithm is in some sense monotonic. Let $\hat{\theta}^{(t)}_{k,y}$ be the estimate of the parameters $\theta_{k,y}$ in the beginning of iteration t of the EM algorithm, and $\theta^{(t)}$ be the vector of all parameter estimates in that iteration, $[\hat{\theta}^{(t)}_{1,0}, \hat{\theta}^{(t)}_{1,1}, \ldots, \hat{\theta}^{(t)}_{N,y=1}]$. Which of the following quantities increases monotonically with t?

( ) $P(\phi_k = 1|y, \theta^{(t)}_{k,y})$ for all k
(x) $\prod_{i=1}^{N} P(\phi^i_1, \ldots, \phi^i_N|\theta^{(t)})$
( ) $\prod_{i=1}^{N} P(y_i = 1|\phi^i_1, \ldots, \phi^i_N, \theta^{(t)})$

4. **(4 points)** The class labels actually correspond to "relevant" and "irrelevant" documents. In classifying any document as relevant or irrelevant, we have to take into account that we might prefer to miss a few relevant documents if we can avoid misclassifying a large number of irrelevant documents as relevant. To express such a preference we define a utility $U(y, \hat{y})$, where y is the correct label and $\hat{y}$ is how we classify the document. Draw an influence diagram that incorporates the Naive Bayes model, our decisions, and the utility. Mark each node in the graph according to the variables (or utility) that they represent.



5. **(2 points)** Let's modify the Naive Bayes model a bit, to account for some of the possible dependencies between the keywords. For example, suppose we order the

keywords so that it would be useful to model the dependency of ϕ_k on ϕ_{k-1}, $k = 2, \ldots, N$ (the keywords may, for example, represent nested categories). We expect these dependencies to be the same for each class, but the parameters can be affected by the class label. Draw the graphical model for this – call it *Sophisticated Bayes* – model. Please assume again that $N = 3$.