6.891 Machine learning and neural networks

FALL 2000 – Final exam SOLUTIONS

December 13, 2000

(2 points) Your name and MIT ID #:

John Doe, 00000000

(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 then just write "A".

A

Problem 1

1. **(T/F – 2 points)** Any representation of the inputs that contains (in some form) all the information about the labels would do equally well for classification purposes

   F

   The relevant information should be explicit for it to be useful. The statement implies that, for example, a representation where the relevant information is encrypted would serve equally well.

2. **(T/F – 2 points)** For support vector machines the kernel function is a good place to incorporate prior information about the classification problem

   T

   The input examples appear only in the kernel function. The kernel function is therefore the only place to incorporate prior information that pertains to the examples.

3. **(T/F – 2 points)** Support vector machines, like logistic regression models, give a probability distribution over the possible labels given an input example

   F

   Support vector machines provide only a hard label. The classification margin cannot be directly interpreted as a probability.

4. **(T/F – 2 points)** A wrapper feature selection method can be expected to always outperform a simpler filtering approach

   F

   The key here is the word "always". While wrapper methods tend to give better results as they take into account the form of the classifier, this is by no means guaranteed.

5. **(T/F – 2 points)** In clustering, selecting how small groups should be merged into larger ones is secondary to finding the metric that compares individual examples.

   T

   The key to clustering is the choice of the metric between the examples. There are several ways of merging smaller clusters into larger ones but they all rely on this initial metric.

6. **(T/F – 2 points)** Most clustering algorithms try to minimize or maximize a well-defined objective function

   F

   Most clustering algorithms do not even have an objective function. What is, for example, the objective function in a simple greedy hierarchical clustering? Mixture models, however, do maximize a well-defined objective (likelihood of the examples).

Problem 2

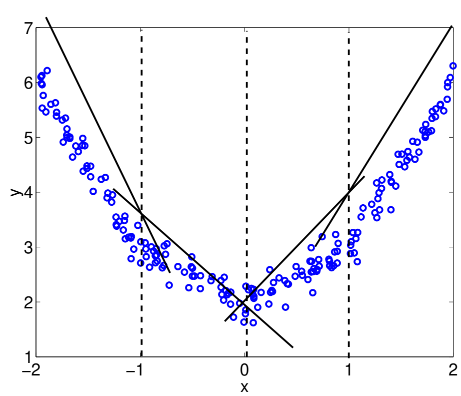


Figure 1: Observed data. The distribution of examples along the x-axis is sampled uniformly in the interval $[-2, 2]$.

1. **(6 points)** Consider a mixture of experts architecture where each expert is a linear regression model (with Gaussian noise) relating a real valued input x to a real valued output y. The choice of the experts is governed by a softmax gating network. How many experts would you need to adequately model the observed data in figure 1? Write *in figure 1* the regions that correspond to a good allocation of experts in this architecture. Briefly explain why the softmax gating network can serve to select the regions you have specified.

> Number of experts = 4
>
> The softmax gating network is given by
>
> $$P(i|x,\eta) = \frac{e^{v_i x + v_{i0}}}{\sum_{j=1}^{4} e^{v_j x + v_{j0}}} \tag{1}$$
>
> So, given x, the expert with the highest probability is the one that has the largest value of $v_i x + v_{i0}$ (a line). By setting the parameters of these lines such that they appear as in figure 1, then within each region we always select the appropriate expert with the highest probability. By scaling these parameter values we move from soft selection of experts to deterministically picking the appropriate one in each region.
>
> Note that the solid lines in the figure have nothing to do with how the experts make their predictions.

2. **(6 points)** Briefly explain why we would prefer such a mixture of experts architecture over estimating a simple mixture of Gaussians model over both x and y, i.e., treating (x, y) as a two-dimensional observation? Note that we can always make predictions from such a mixture model by computing the posterior probability of y given only x. You can assume here that the mixture of Gaussians model would be estimated by maximizing the likelihood of all the observations $\{(x_1, y_1), \ldots, (x_n, y_n)\}$.

> There are at least two reasons for preferring a mixture of experts model in this context: First, since we have indicated that the x-values are distributed uniformly, any Gaussian model over x is a bad choice (a Gaussian tends to concentrate the probability mass close to the mean value and the density decays rapidly as we move away from the mean). We could, of course, increase the number of mixture components to better model the density over x. This leads to higher complexity, problems with overfitting, etc.
> The second reason is that we are not interested in the density over x at all. We merely wish to predict y given x. Fitting a model by maximizing the likelihood of both x and y is simply the wrong criterion. It does not match what we wish to solve.

3. **(T/F – 2 points)** A mixture of experts model is resistant to overfitting in the sense that if we have too many experts most of them will never be selected by the gating network

> F

Like any mixture model, a mixture of experts model is susceptible to overfitting.

4. **(T/F – 2 points)** When estimating a mixture model with several components it can make a difference whether we use a flat mixture model or the corresponding hierarchical mixture model

   T

   The question here is whether it makes any difference in estimation if we use a flat 5-component mixture model or a hierarchical version of this. If data possesses any hierarchical structure, then the hierarchical mixture model is more likely to find this structure. The hierarchical organization "ties" mixture components together during estimation.

Problem 3

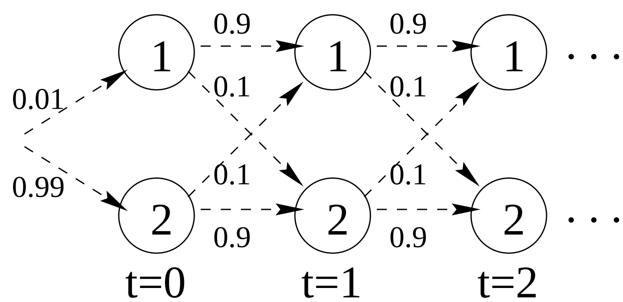


Figure 2: A hidden Markov model. The figure shows the hidden state transitions and the associated probabilities along with the initial state distribution.

Consider a hidden Markov model illustrated in figure 2. We assume that the state dependent outputs (coin flips) are governed by the following distributions

$$\begin{array}{ll} P(x = heads|s = 1) = 0.51 & P(x = heads|s = 2) = 0.49 \\ P(x = tails|s = 1) = 0.49 & P(x = tails|s = 2) = 0.51 \end{array}$$

In other words, our coin is slightly biased towards *heads* in state 1 whereas in state 2 *tails* is a bit more probable outcome.

1. **(6 points)** Now, suppose we observe three coin flips all resulting in *heads*. The sequence of observations is therefore $\{heads, heads, heads\}$. What is the most likely state sequence corresponding to these three observations? Briefly explain your reasoning. (you shouldn't need a calculator for this)

The most likely hidden state sequence is 2,2,2. This is because the output probabilities are nearly identical and we are very likely to start from 2 and stay there. We loose a factor of 9 in probability if we ever switch to state 1.

2. **(6 points)** What happens to the most likely state sequence if we observe a long sequence of all *heads* (e.g., 1000)? Briefly explain your reasoning. (you shouldn't need a calculator for this)

The hidden state sequence becomes 21111...

When we increase the number of observations (all *heads*), we increase the pressure for the system to switch to 1 since state 1 has a slight advantage per observation. Eventually the switch will take place (note that there's no benefit from ever switching back to state 2). When does this happen? The cost of the 2–¿1 transition is the same regardless of when it takes place. The more we postpone the transition, however, the more *heads* we would have to generate from state 2. However, it is somewhat better to go via 2 initially and switch right after (0.99*0.49*0.1 $\cdots$) rather than start from 1 to begin with (0.01*0.51*0.9 $\cdots$).

3. **(T/F – 2 points)** The only limitation of first order homogeneous Markov chains is that they can capture only one time step dependencies among the observations

F

A first order Markov model can only capture one time step dependencies. A homogeneous first order Markov chain assumes, in addition, that this dependence does not change over time.

4. **(T/F – 2 points)** By increasing the number of hidden states in an HMM, we can model well (in the maximum likelihood sense) practically any finite sequence of observations

T

To model any finite length sequence, we can increase the number of hidden states in an HMM to be the number of observations in the sequence and therefore (with appropriate parameter choices) generate the observed sequence with probability one. Given a fixed number of finite sequences (say n), we would still be able to assign probability 1/n for generating each sequence. This is not useful, of course, but highlights the fact that the complexity of HMMs is not limited.

5. **(T/F – 2 points)** HMMs can model observation sequences arising from multiple underlying processes and the description of such models in terms of state transition diagrams is concise

   F

   The key here is the second part of the sentence. As discussed in the lecture, it is quite awkward to write transition diagrams describing multiple processes.

Problem 4

Your task here is to identify the relevant variables and the graph structure that captures the following (imaginary) setting. There may be multiple "correct" answers.

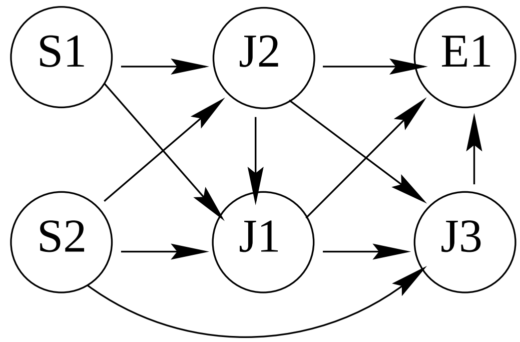
"A panel of three judges determines the outcome of presidential elections. Each judge can vote for one of the two possible candidates and the outcome is obtained by a majority rule. Two of the judges are impartial in the sense that they will listen to arguments from two spokespersons each working for one of the candidates while the remaining judge consistently pays attention to only one of the spokespersons. Each spokesperson will ask a judge to vote for a specific candidate. The spokespersons never talk nor listen to each other directly."

1. **(6 points)** Identify the relevant variables based on the above description. For each variable state the possible values that it can take. If you use abbreviations to identify the variables make sure they are not ambiguous.

   The variables are:

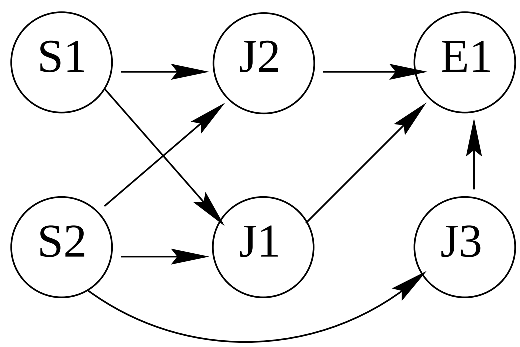
Election outcome	E1	$\{1,2\}$ (candidate 1 or 2)
Spokesperson 1	S1	$\{1,2\}$ (argument has the effect of supporting 1 or 2)
Spokesperson 2	S2	$\{1,2\}$ (argument has the effect of supporting 1 or 2)
Judge 1 (impartial)	J1	$\{1,2\}$ (votes for 1 or 2)
Judge 2 (impartial)	J2	$\{1,2\}$ (votes for 1 or 2)
Judge 3 (partial)	J3	$\{1,2\}$ (votes for 1 or 2; listens to S2 only, say)

2. **(6 points)** Draw a Bayesian network that captures the interactions between the variables. *Avoid any assumptions that you cannot make on the basis of the above description.* Please indicate which variables correspond to which nodes.



The description does not tell us whether the judges discuss the case amongst themselves. We therefore cannot make any independence assumptions between the judges conditionally on the spokespersons. We therefore draw all possible arrows between the judges so long as the resulting graph is acyclic. *The directions of the arrows that connect the judges are irrelevant; they are all equivalent. You could also draw an undirected edge between each pair of judges but not bi-directional edges.*

3. **(4 points)** The graph might change if the above description had started with "A panel of three *independent* judges...". If the graph would change, please draw the new graph. Otherwise state that there are no changes.



If the judges make their decisions independently of each other (but still contingent on the spokespersons), we simply remove all the arrows between the judges.

4. **(4 points)** Explain under what circumstances (setting of some of the variables etc.) we might observe "explaining away" in the graph you just drew. If none exists, briefly explain why not.

   (For your convience, here's a brief description of "explaining away": *When we have multiple possible causes for a single known effect, explaining away refers to the phenomenon where acquiring further evidence about the presence of one of the causes makes the other ones less likely.*)

There is a natural explaining away effect here. Suppose we know the outcome of the election (E1 = 1). This increases the probability that each judge individually voted for candidate 1. If we now learn that the first two judges voted for candidate 1 (i.e., J1 = 1 and J2 = 1), then there's no remaining evidence supporting that J3 voted for candidate 1 (since the election outcome is a majority vote). The additional evidence (J1 = 1 and J2 = 1) now fully explains the initial observation (E1 = 1).

5. **(T/F – 2 points)** The graph structure is useful *only if* it captures all the independence properties present in the underlying probability distribution

   F

   Graph structure is useful if the properties that we can derive from the graph are true for the underlying probability distribution. It is often the case that we cannot capture all the independence properties with a graph.

6. **(T/F – 2 points)** Given any probability distribution, we can find a Bayesian network as well as a Markov random field that is consistent with the distribution

   T

   A fully connected undirected graph (or its directed quivalent) is consistent with any distribution as it makes no independence assumptions whatsoever.

7. **(T/F – 2 points)** A Boltzmann machine where *all* the variables are observable can *only* capture second order statistics (means and covariances) between the variables

   T

   When all the variables are observed, Boltzmann machines care only about the second order statistics (recall the estimation equations in the lecture notes). This is no longer true if there are unobserved variables as such variables can correlate more than two observed variables (this is analogous to the case of one underlying but unknown cause and multiple effects).