

CS195-5: Introduction to Machine Learning
Fall 2006

Problem Set #7 : Kangaroo

Out: December 6, 2006

Due: never

Instructions

There is no need to submit this problem set. Still, keeping in the class tradition, we assign it a name. It will be called “kangaroo”.

How and what to submit? You do not have to turn in anything. This problem set is for you to check your understanding of material, and to clarify some potentially unclear points.

1 Estimation

Problem 1 [0 points]

We are given a set of observations $X_N = x_1, \dots, x_N \in \mathbb{R}$ and told these are drawn from a uniform distribution,

$$p(x; a, b) = \begin{cases} 1/(b - a), & \text{if } a \leq x \leq b, \\ 0 & \text{otherwise.} \end{cases}$$

Write down the likelihood $p(X_N; a, b)$. What is the maximum likelihood estimate of a and b ?

End of problem 1

2 Clustering

Problem 2 [0 points]

Compare two clustering algorithms: k -means and hierarchical clustering. Specifically, give two advantages for each algorithm over the other.

End of problem 2

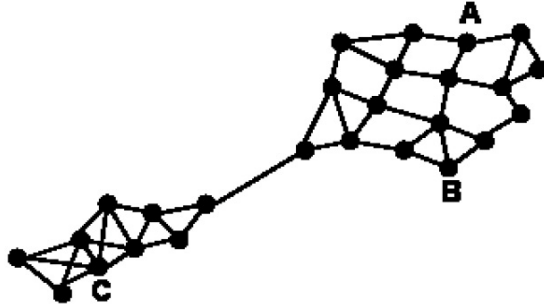


Figure 1: A neighborhood graph.

Problem 3 [0 points]

Figure 1 shows a neighborhood graph constructed for a data set based on thresholding the distances between examples. We will be focusing on the three examples marked A , B and C .

Given the graph we can compute the weight matrix \mathbf{W} according to

$$W_{ij} = \begin{cases} \exp(-\beta\|i\|), & \text{if } i \neq j \text{ and there is an edge btw. } \mathbf{x}_i \text{ and } \mathbf{x}_j, \\ 0 & \text{otherwise.} \end{cases}$$

We can then define a random walk model on this graph, as described in class, with the transition probability matrix \mathbf{P} given by

$$P_{ij} = \frac{W_{ij}}{\sum_l W_{il}}.$$

We will denote by P_{ij}^t the entry on the i -th row and j -th column of the matrix \mathbf{P} raised to the power t .

For each of the expressions below, replace the question mark with either $<$, $>$ or $=$ and briefly justify your decision.

1. P_{AC}^{10} ? P_{AC}^{100}
2. P_{AB}^{10} ? P_{AB}^{100}
3. $\sum_i P_{Ai}^{10}$? $\sum_i P_{Ai}^{100}$
4. P_{BA}^{∞} ? P_{BC}^{∞}

(the last expression refer to the limit when $t \rightarrow \infty$).

End of problem 3

3 Hidden Markov Models

Problem 4 [0 points]

Look *carefully* at the following string of 120 characters A, B and C.

```
AAABBBACAAAABACAABACAABACAABBBBBACAAAABBBBACAABBBBACAAA  
BBACAAABBACABBBBBBACAABACABACAABBACAABBBBACAAAABACABBB
```

Describe a discrete HMM model for this string. How many states are there in the model? (Try to use as few states as possible that allow to capture the regularities you notice). Write down the estimated output probabilities for each state (a 1×3 vector), the transition probability matrix, and the initial state probabilities.

End of problem 4

Advice: Note: you should be able to provide a good qualitative answer without writing any code or doing experiments in Matlab. A more careful estimate will require some counting, although it can still be done without a computer!

Problem 5 [0 points]

Suppose we are fitting an HMM model with Gaussian mixture emission probabilities to a set of observation sequences. We are allowed to set any of the parameters of the model as we please (including the number of states and the number of Gaussian components in the mixtures). Our objective is to maximize the log-likelihood of the observed data under the model.

Could we overfit? If yes, explain how it could happen (i.e., what are the “sources” of overfitting). If not, explain why not.

End of problem 5

4 Mixtures of Experts

We wish to estimate a mixture of experts model for regression. Each expert we use here is a linear regression model of the form

$$p(y|x, \mathbf{w}) = \mathcal{N}(y; w_1x + w_0, \sigma^2)$$

where $\mathcal{N}(y; \mu, \sigma^2)$ denotes a Gaussian distribution over y with mean μ and variance σ^2 . Each expert j has an independent set of parameters $\mathbf{w}_j =$

$[w_{j0}w_{j1}]^T$ and its own σ_j^2 . Note that the first index of w_{jd} refers to the expert.

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

$$p(j = 1 | x, \mathbf{v}) = \sigma(v_1x + v_0)$$

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

Problem 6 [0 points]

Suppose we estimate a mixture of two experts as described above on the data in Figure 2. You can assume that the estimation is successful in the sense that we find a setting of the parameters that maximizes the log-likelihood of the data. In the empty figure below, sketch the mean predictions from the two experts as well as the decision boundary for the gating network, if it had to make “hard decisions” about expert identity.

End of problem 6

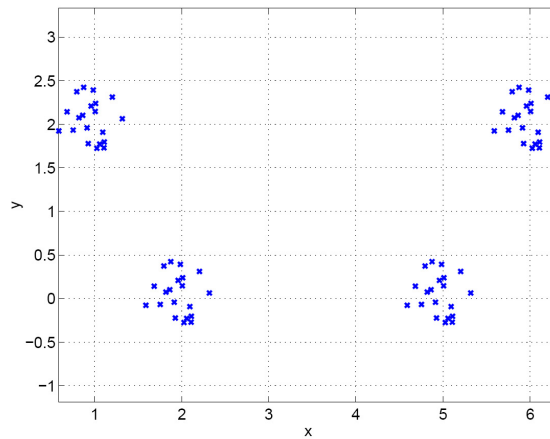


Figure 2: Data/graph for mixtures of experts

Problem 7 [0 points]

We now switch to a *regularized* maximum likelihood objective by incorporating the following regularization penalty :

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

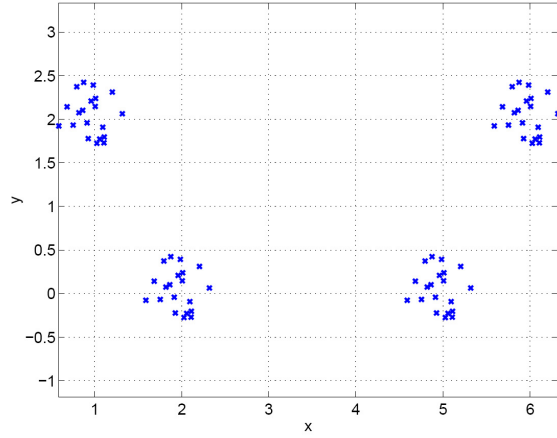


Figure 3: Data/graph for regularized mixtures of experts

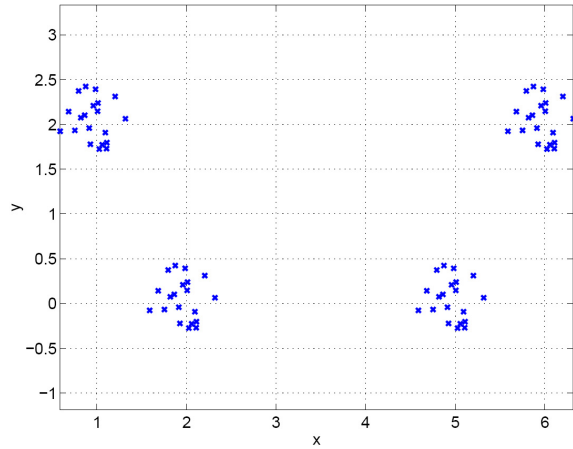


Figure 4: Another data/graph for mixtures of experts

into the log-likelihood objective. Note that the penalty includes only one parameter from each of the experts. Increasing λ means we impose a stronger penalty. Similar to the previous problem, sketch in figure 3 the optimal regularized solution for the mixture of two experts model when the regularization parameter λ is set to a very large value.

Are the variances in the predictive Gaussian distributions of the experts larger, smaller, or about the same after the regularization? Justify your answer.

End of problem 7

Problem 8 [0 points]

Consider again the *unregularized* mixture of two experts. If we tried to estimate this model on the basis of the data in figure 4, what would the solution look like in this case? As before, sketch the solution in the figure and justify your answer.

End of problem 8

5 Graphical models

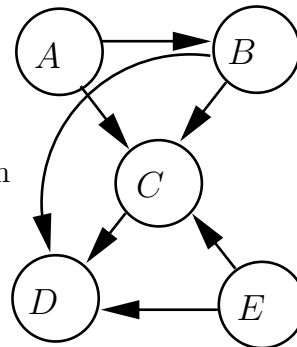


Figure 5: A graphical model over five random variables.

Problem 9 [0 points]

For the graphical model in Figure 5, write down in the factorized form the joint probability distribution over the five variables A, B, C, D and E , defined by the model.

Suppose now that all the variables in the model are discrete and can take integer values between 1 and 5. How many parameters do we need (in other words, how many numbers do we need to store) to fully specify the joint probability distribution over A, B, C, D and E ? How many parameters would we need if we did not make any independence assumptions, and had a fully connected graph?

End of problem 9

Problem 10 [0 points]

Consider again the model in Figure 5. For each of the following statements, write down the *minimal* set of variables X for which the statement is true.

1. A is independent of D given X .
2. B is independent of E given X .

Confirm your answers by running the Bayes Ball algorithm on the model with the nodes in the appropriate X shaded.

End of problem 10

Problem 11 [0 points]

Consider a classification task over a domain represented by a four-dimensional feature vector $\mathbf{x} = [x_1, x_2, x_3, x_4]^T$. We can represent various assumptions models for this task using a directed graphical model with five variables: y (class label) and x_1, \dots, x_4 .

Sketch the graphical model for each of the following assumptions:

1. Naive Bayes classifier with Gaussian model $p(x_j | y) = \mathcal{N}(x_j; \mu_j, \sigma_j^2)$.
2. A jointly Gaussian model for $p(\mathbf{x} | y) = \mathcal{N}(\mathbf{x}; \mu, \Sigma)$ where the covariance is restricted to the be of the form

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & 0 & 0 \\ \sigma_{12} & \sigma_2^2 & 0 & 0 \\ 0 & 0 & \sigma_3^2 & \sigma_{34} \\ 0 & 0 & \sigma_{34} & \sigma_4^2 \end{bmatrix}.$$

3. Completely unconstrained jointly Gaussian model.

For each model, do not include any edges that are not necessary.

End of problem 11