Assignment 2: MLE, EM, Regression

10-701/15-781: Machine Learning (Fall 2004)

Out: Sept. 30th, 2004
Due: Oct. 14th 2004, Thursday, In class,

a *This assignment has four problems to test your understanding about MLE, EM and regression.*

b *For the questions requiring programming, please use matlab. You need to submit your code to TAs. Please also submit a printing version of your code.*

c *For questions and clarifications, contact Max (maxim+@cs.cmu.edu) or Yanjun (qyj@cs.cmu.edu).*

d *Policy on collaboration:*

Homeworks will be done individually: each student must hand in their own answers. It is acceptable, however, for students to collaborate in figuring out answers and helping each other solve the problems. We will be assuming that, as participants in a graduate course, you will be taking the responsibility to make sure you personally understand the solution to any work arising from such collaboration.

d *Policy on late homework:*

Homework is worth full credit at the beginning of class on the due date. It is worth half credit for the next 48 hours. It is worth zero credit after that. You must turn in all of the 4 homeworks, even if for zero credit, in order to pass the course.

e *To help the graders (including yourself...), please be neat, answer the questions in the order they are stated. Staple each "Problem" separately, and be sure to write your Andrew ID and name on the top of every page.*

Question 1. Maximum Likelihood Estimation

Suppose X is a binary random variable that takes value 0 with probability p and value 1 with probability $1-p$. Let $X_1, \ldots, X_n$ be iid samples of X.

1.1 Compute an MLE estimate of p (denote it by $\hat{p}$).

1.2 Is $\hat{p}$ an unbiased estimate of p? Prove the answer.

1.3 Compute the expected square error of $\hat{p}$ in terms of p.

1.4 Prove that if you know that p lies in the interval $[\frac{1}{4}; \frac{3}{4}]$ and you are given only $n = 3$ samples of X, then $\hat{p}$ is an inadmissible estimator of p when minimizing the expected square error of estimation. (An estimator δ of a parameter θ is said to be *inadmissible* when there exists a different estimator δ' such that $R(\theta, \delta') \leq R(\theta, \delta)$ for all θ and $R(\theta, \delta') < R(\theta, \delta)$ for some θ, where $R(\theta, \delta)$ is a risk function and in this problem it is the expected square error of the estimator).

Question 2. EM

For the following questions, please give clear step by step derivation.

2.1 Suppose that the p.d.f. of a random variable X has a 2-component mixture form:

$$p_\alpha(x) = \alpha * p_1(x) + (1-\alpha) * p_2(x) \tag{1}$$

One component is the density model $p_1(x)$ and the other component is the density model $p_2(x)$. We know both $p_1(x)$ and $p_2(x)$. We do not know α. Given that $\{ x_1, x_2, ..., x_n \}$ are iid samples from the distribution of X, please give an EM algorithm for estimating α. (Describe the E-step and M-step clearly in your answer).

2.2 Suppose that $Y_1 \sim exp(1/\theta_1)$ and $Y_2 \sim exp(1/\theta_2)$, and $\theta_1 \neq \theta_2$. Y_1 and Y_2 are independent. Let $X = Y_1 + Y_2$ denote the sum of Y_1 and Y_2, Given that $\{ x_1, x_2, ..., x_n \}$ are iid samples from the distribution of X.

- Derive an expression for the density of X in terms of θ_1 and θ_2
(Hint1: The density of Y_1 is $f_{\theta_1}(y) = \theta_1 e^{-\theta_1 y}$, similarly for Y_2)
(Hint2: You could first derive CDF of X, $F(x) = P(Y_1 + Y_2 < x) = \int_0^x \int_0^{x-y_1} f_{\theta_1}(y_1) f_{\theta_2}(y_2) dy_2 dy_1$)

- Derive the E-step and M-step, and give explicit expressions for the parameter updates in the EM process for computing the MLE of θ_1 and θ_2.

Question 3. Gaussian mixtures

In this problem you will implement a Gaussian mixture model algorithm and will apply it to the problem of clustering gene expression data. Gene expression measures the levels of messenger RNA (mRNA) in the cell. The data you will be working with is from a model organism called yeast, and the measurements were taken to study the cell cycle system in that organism. The cell cycle system is one of the most important biological systems playing a major role in development and cancer.

All implementation should be done in Matlab. At the end of each sub-problem where you need to implement a new function we specify the prototype of the function.

3.1 Download the file 'alphaVals.txt'. This file contains 18 time points (every 7 minutes from 0 to 119) measuring the log expression ratios of 745 cycling genes. Each row in this file corresponds to one of the genes. Also, download the file 'geneNames.txt' which contains the names of these genes. For some of the genes, we are missing some of their values due to problems with the microarray technology (the tools used to measure gene expression). These cases are represented by values greater than 100.

3.2 Implement (in matlab) an EM algorithm for learning a mixture of five (18-dimensional) Gaussians. It should learn means, convariance matrices and weights for each of the Gaussian. You can assume, however, independence between the different data points, resulting in a diagonal covariance matrix. How can you deal with the missing data? Why is this correct? Plot the centers identified for each of the five classes. Each center should be plotted as a time-series of 18 time points. Hand this plot with your solutions.

Here is the prototype of the matlab function you need to implement:

$$function[mu, s, w] = emcluster(x, k, ploton); \tag{2}$$

x is input data, where each row is an 18-dimensional sample. Values above 100 represent missing values. k is the number of desired clusters. *ploton* is either 1 or 0. If 1, then before returning the function plots log-likelihood of the data after each EM iteration (the function will have to store the log-likelihood of the data after each iteration, and then plot these values as a function of iteration number at the end). If 0, the function does not plot anything. The function outputs mu, a matrix

with k rows and 18 columns (each row is a center of a cluster), s is also k by 18, with each row being diagonal elements of the corresponding covariance matrix, and w is a column vector of size k, where $w(i)$ is a weight for ith cluster.

3.3 How many more parameters would you have had to assign if we remove the independence assumption above? Explain.

3.4 Suggest and implement a method for determining the number of Gaussians (or classes) that are the most appropriate for this data. Please confine the set of choices to values in between 2 and 7. (Hint: the method can use an empirical evaluation of clustering results for each possible number of classes). Explain the method.

Here is the prototype of the matlab function you need to implement:

$$function[k, mu, s, w] = clust(x); \tag{3}$$

x is input data, where each row is an 18-dimensional sample. Once again values above 100 represent missing values. k is the number of classes selected by the function. mu, s and w are defined as in 3.2.

3.5 Use the Gaussians determined in (d) to perform hard clustering of your data by finding, for each gene i the Gaussian j that maximizes the likelihood: $p(i|j)$. Use the function 'printSelectedGenes.m' to write the names of the genes in each of the clusters to a separate file.

Here is the prototype of the matlab function you need to implement:

$$function[c] = hardclust(x, k, mu, s, w); \tag{4}$$

x is defined as before. k, mu, s, w are the output variables from the function written in 3.4 and are therefore defined there. c is a column vector of the same length as the number of rows in x. For each row, it should indicate the cluster the corresponding gene belongs to. The function should also write out files as specified above. The filenames should be: clust1, clust2, ..., clustk.

One way to determine the function of the genes in the different clusters is to query a biological database. These databases contain hundreds of functional categories. For each of these categories they list the set of genes that are determined to be associated with this category (note that a gene can be associated with more than one category). By computing the significance of the intersection between each of the categories and the cluster we can identify significantly enriched categories which help us define the role of the genes in the cluster.

Here, we will use functional assignment for genes from the MIPS database. You should first download the files 'cin.txt', 'catSize.txt' and 'catNames.txt' which will be used by the program we have supplied (see next). We have provided the file 'compSigClust.m' which uses these files to compute the intersection of your clusters with each of the database categories (whose names are listed in 'catNames.txt'). For each of the files you generated in 3.5 (containing names for genes in the clusters you identified) you will use 'compSigClust.m' to find the three most significant categories associated with this cluster.

3.6 Use compSigClust.m to perform the statistical significance test (everything is already implemented here, so just use the function). Hand in a printout with the top three categories for each cluster (this is the output of compSigClust.m).

Question 4. Regression

Linear regression models a real-valued output Y given an input vector X as

$$Y|X \sim Normal(\mu(X), \sigma^2)$$

where the mean is a linear function of the input: $\mu(X) = \beta^T X = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p$

Logistic regression models a binary output Y by

$$Y|X \sim Bernoulli(\theta(X))$$

where the Bernoulli parameter is related to $\beta^T X$ by the logit transformation

$$logit(\theta(X)) \equiv \log\left(\frac{\theta(X)}{1-\theta(x)}\right) = \beta^T X$$

Given data $\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$, for each of the two regression models above , show that at the MLE $\hat{\beta}$

$$\sum_{i=1}^{n} x_i * y_i = \sum_{i=1}^{n} x_i * E[\text{ Y } |X = x_i, \beta = \hat{\beta}]$$