STATS 306B Methods for Applied Statistics: Unsupervised Learning
Stanford University, Spring 2014

Problem Set 2

Due: Monday, May 5, 2014 (at the start of class)

Instructions:

- Please submit any code written for this assignment along with your derivations and plots.

Practical advice:

- Problem 1 will likely be the most time consuming.

Problem 1 (EM for Hidden Markov Models)**.**

(a) Implement the EM algorithm for an HMM with hidden states $z_t \in \{1, \ldots, k\}$ (for any $k > 1$) and isotropic Gaussian emission probabilities, $p(x_t|z_t)$, for $x_t \in \mathbb{R}^d$ ($d \geq 1$). That is, $x_t|z_t = j \sim \mathcal{N}(\mu_j, \sigma_j^2 I)$ for unknown parameters (μ_j, σ_j^2). Do not use a pre-existing implementation.

Note: The α and β recursions involve the repeated multiplication of small numbers and hence are susceptible to numerical underflow. Section 12.7 of the assigned HMM chapter presents an effective normalization strategy for countering this numerical underflow.

(b) Use your implementation from part (a) to learn the parameters of an HMM with $k = 4$ states and $d = 2$, using the training observations in `hmm-gauss.dat`. After each EM iteration, evaluate the log likelihood of the learned model parameters on the training data and separately on the test data in `hmm-test.dat`. After EM has converged, run the Viterbi algorithm to assign a state to each datapoint. Plot the training and test data in a way that indicates to which state each point belongs; overlay the learned cluster means on your plot, and distinguish the training and test data points on your plot.

Problem 2 (Hierarchical Clustering of Microarray Data)**.** The file `microarray.txt` contains a microarray dataset with 1000 genes (rows) and 40 samples (columns). The first 20 samples are from healthy patients, while the second 20 are from a diseased group.

(a) Apply one of the forms of agglomerative hierarchial clustering discussed in class to the samples using the negative of the correlation between datapoints as dissimilarity measure, and plot the dendrogram. You may make use of existing hierarchical clustering functionality like 'hclust' in R. Is there a clear indication of the healthy vs. diseased group structure?

(b) Your collaborator wants to know which genes differ the most between the healthy and diseased groups. Suggest a way to answer this and apply it here.

Problem 3 (Spectral Clustering and Image Segmentation)**.** In this problem, you will use spectral clustering to automatically segment an image. The image you will use, `boat32.jpg`, is a heavily downsampled version of `boat.jpg` from the Berkeley Segmentation Dataset and Benchmark (`http://www.eecs.berkeley.edu/Research/Projects/CS/vision/bsds/`).

(a) Treat each pixel x_i of `boat32.jpg` as a datapoint consisting of a row position r_i, a column position c_i, and a pixel intensity value p_i. For any pair of pixels, x_i and x_j, assign the similarity $s_{ij} = 0$ if their row positions or column positions are greater than $\Delta = 3$ pixels apart; otherwise, assign the Gaussian similarity $s_{ij} = \exp\left(-\frac{(p_i - p_j)^2}{M^2\sigma^2}\right)$, where M is the maximum pixel intensity in the image, and $\sigma^2 = 0.015$.

(b) Implement the spectral clustering algorithm for $k = 2$ clusters using the random-walk normalized Laplacian, a weight matrix with entries $w_{ij} = s_{ij}$, and 10 random restarts in the k-means clustering step. Form the two-column matrix U of the bottom 2 eigenvectors of the Laplacian, and produce a scatter plot of the rows of U (recall that each row of U is the spectral clustering representation of a single datapoint). Plot the image and designate which pixels were assigned to each cluster. Comment on the results. Why did we not have you run spectral clustering on the original image `boat.jpg`?

Note: The random-walk normalized Laplacian is not a symmetric matrix. If you use the R function `eigen`, you should include the argument `symmetric = FALSE`.

(c) How do the results change if σ^2 is much larger or smaller? How do they change if Δ is much larger or smaller? (For this final part of the problem, you do not need to include plots of the results.)

Problem 4 (Principal Component Analysis and Olive Oil)**.** The file `olive.txt` contains measurements of eight chemical concentrations on 572 samples of olive oil from 9 different areas of Italy. These areas are further grouped into 3 regions. The data file is formatted as follows

- Column 1: Region
- Column 2: Area
- Columns 3 - 10: Concentrations

(a) Carry out a principal component analysis on the entire dataset for $k = 1, 2, \ldots, 6$ components. How does the fraction of variation explained in each region vary as k varies? How does the fraction of variation explained in each of the following areas vary as k varies: areas 4, 5, and 8?

(b) With analogy to the gap statistic in the clustering setting, formulate and implement a test for the number of principal components that compares the optimal PCA objective on this data to the expected optimal PCA objective for data sampled from an appropriate reference distribution. Apply it to your PC analysis of the entire dataset in this example.

Problem 5 (PCA and Detrending)**.** A researcher has carried out microarray experiments producing a data matrix of real-valued measurements with 1000 rows and 100 columns. Each row represents a gene and each column a patient. Each patient's measurements were taken on a different day, and the days run from earliest on the leftmost to latest on the rightmost. Moreover, the patients fall into two groups: control (C) and treatment (T). The control-treatment assignments are scattered but not distributed uniformly at random over the days. The researcher wishes to test whether each gene measurement is significantly higher or lower in the treatment group vs the control group.

As a pre-analysis (before comparing T vs C) a statistician carried out a principal component analysis of the data and discovered that the leading principal component (a vector of length 100) showed a strong linear trend from left to right and explained 10% of the variation in the data. After showing the principal component to his biological collaborator, the collaborator remembered that each patient sample was run on one of two machines (A and B), and machine A was used more often in the earlier days, while B was used more often on later days. She has a record of which sample was run on which machine. Based on this new information, the statistician decides to replace the data matrix X with $X - \theta uv^\top$ to control for the effect of different machines. Here, u and v are the leading left and right singular vectors of X with singular value θ (v is also the leading principal component).

(a) Critique this detrending idea and suggest a better approach.

(b) Design and run a small simulation experiment to demonstrate the superiority of your idea.

Problem 6 (Feedback)**.** (This "problem" is not graded.)

(a) How much time did you spend on this problem set?

(b) Which problems did you find valuable?