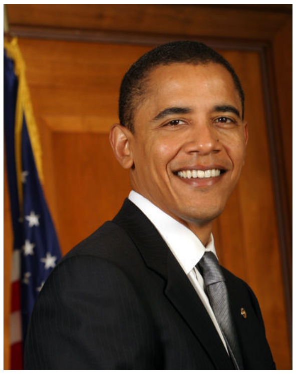
COMP 652: Machine Learning - Assignment 4

Posted Monday November 16, 2009
Due Wednesday, November 25, 2009

Please e-mail your assignment (in pdf) as well as your Matlab source code to cs652@cs.mcgill.ca by midnight.

1. [30 points] **K-means**



Consider the data in the file "hw4q1.txt". This file contains a large number (210,012) of length 3 vectors, each on one line. Each vector represents the red, green, and blue intensity values of one of the pixels in the image shown at the right. The image has 516 rows and 407 columns. The pixels in the file are listed row by row from top to bottom, and within each row from left to right. For example, the first pixel in the file is the uppermost left pixel in the image. The second line of the file contains the pixel to the right of that one, and so on. In this assignment, we will explore clustering methods, applying them in particular to the problem of dividing the pixels of the image into a small number of similar clusters.

Consider the K-means clustering algorithm, as described in class. In particular, consider a version in which the inputs to the algorithm are:

(a) The set of data to be clustered. (I.e., the vectors $\mathbf{x_1}, \mathbf{x_2}, \mathbf{x_3}, \ldots$)

(b) The desired number of clusters, K.

(c) Initial centroids for the K clusters.

Then the algorithm proceeds by alternating: (1) assigning each instance to the class with the nearest centroid, and (2) recomputing the centroids of each class—until the assignments and centroids stop changing.

There are many implementations of K-means publicly available. However, please implement K-means clustering in MATLAB by yourself. Then, use your implementation to cluster the data in the file mentioned above, using $K = 8$, and the initial centroids:

R	G	B
255	255	255
255	0	0
128	0	0
0	255	0
0	128	0
0	0	255
0	0	128
0	0	0

Turn in your code, as well as a report on all of the following:

- How many clusters there are in the end. (Recall that a cluster can "disappear" in one iteration of the algorithm if no vectors or closest to its centroid.)
- The final centroids of each cluster.
- The number of pixels associated to each cluster.
- The sum of squared distances from each pixel to the nearest centroid after every iteration of the algorithm.

Visualize your result by replacing each pixel with the centroid to which it is closest, and displaying the resulting image.

2. [20 points] **Understanding correlations**

   In this problem, we will show that the independence of two random variables is a sufficient but not necessary condition for the correlation matrix to be diagonal.

   (a) [10 points] Consider two random variables X and Z which are independent, i.e. $p(X, Z) = p(X)p(Z)$. Show that any off-diagonal elements in the correlation matrix must be 0.

   (b) [10 points] Now suppose that $Z = X^2$, and $X\tilde{U}nif[-1. + 1]$. Write down $p(Z|X)$. Show that all off-diagonal elements in the correlation matrix must be 0, by using the fact that $p(X, Z) = p(Z|X)p(X)$

3. [40 points] **Playing with PCA**

   (a) [5 points] Generate 200 examples from a Gaussian with mean (5, 20) and covariance matrix:

   $$\begin{bmatrix} 10 & 2 \\ 2 & 5 \end{bmatrix}$$

   Plot the data you generated.

   (b) [5 points] Make a prediction about what directions the principal components should have, based on the class notes.

   (c) [10 points] Run PCA on this data and describe what happens.

   (d) [5 points] Now subtract the mean from all the data points and run PCA. Describe again what happens. Is there any difference in the principal components found? Explain the result.

   (e) [5 points] Now subtract the mean from all the data points and divide the result by the standard deviation (so as to normalize the data). Run PCA again and explain what happens to the result.

   (f) [5 points] Multiply the second coordinate of every point in the data set by 1000 and run PCA again. What happens to the principle components, and why?

   (g) [5 points] Comment on the robustness of PCA wrt the scaling of the data, and on what needs to be done to make sure that good results are obtained.