Statistical Learning: Homework 3

Due 6 Ordibehesht 96 (26 April 2017)

15 April 2017

Instructions: Append your Matlab code to the end of your homework. In your solutions, you should just present your Matlab output (e.g. numbers, table, figure) or snippets of Matlab code as you deem it appropriate. Make sure to present your results (i.e., your Matlab output) in a clear and readable fashion. Careless or confusing presentations will be penalized.

1. The unbiased estimates for the covariance of a d-dimensional Gaussian based on n samples is given by

$$\widehat{\Sigma} = C_n = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \mathbf{m}_n)(X_i - \mathbf{m}_n)^T \tag{1}$$

Where $\mathbf{m}_n = \sum_{i=1}^{n} X_i$. It is clear that it takes $O(nd^2)$ time to compute C_n. If the data points arrive one at a time, it is more efficient to incrementally update these estimates than to recompute from scratch.

(i) Show that the covariance can be sequentially updated as follows,

$$C_{n+1} = \frac{n-1}{n}C_n + \frac{1}{n+1}(X_{n+1} - \mathbf{m}_n)(X_{n+1} - \mathbf{m}_n)^T \tag{2}$$

(ii) How much time does it take per sequential update? (Use big-O notation.)

(iii) Show that we can sequentially update the precision matrix(inverse of covariance matrix) using,

$$C_{n+1}^{-1} = \frac{n}{n-1}\left[C_n^{-1} - \frac{C_n^{-1}(X_{n+1} - \mathbf{m}_n)(X_{n+1} - \mathbf{m}_n)^T C_n^{-1}}{\frac{n^2-1}{n} + (X_{n+1} - \mathbf{m}_n)^T C_n^{-1}(X_{n+1} - \mathbf{m}_n)}\right] \tag{3}$$

(iv) What is the time complexity per update?

2. LDA and QDA

   (i) Write a computer program to perform a quadratic discriminant analysis by fitting a separate Gaussian model per class. Try it out on the vowel data, and compute the misclassification error for the test data. The data can be found in the book website, Vowel test data Vowel training data Vowel dataset information

   (ii) Compare your results with linear discriminant analysis and k-nearest neighbor classifiers. You should write your own code for LDA.

   (iii) Perform Fisher Discriminant Analysis (FDA) to extract two best features and plot the samples.