E9 205 – Machine Learning for Signal Processing

Homework # 1
Due date: Sept. 10, 2018 (in class).

Analytical in person and coding part via email to mlsp18.iisc@gmail.com
Assignment should be solved individually without consent.

August 29, 2018

1. **Induction in PCA** - We have proved that in order to maximize the variance of 1 dimensional projection $y = \boldsymbol{w}^T\boldsymbol{x}$ of D dimensional data $\boldsymbol{x}$, the solution is given by $\boldsymbol{w} = \boldsymbol{u}_1$, where $\boldsymbol{u}_1$ is the eigen vector corresponding to the largest eigen value of sample covariance matrix $\boldsymbol{S} = \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^T$ and $\boldsymbol{\mu}$ denotes the sample mean.

   Let us suppose that the variance of M dimensional projection $\boldsymbol{y}_M = \boldsymbol{W}_M^T\boldsymbol{x}$ is maximized by $\boldsymbol{W} = [\boldsymbol{u}_1\ \boldsymbol{u}_2\ ..\ \boldsymbol{u}_M]$ where $\boldsymbol{u}_1 ... \boldsymbol{u}_M$ are the orthonormal eigen vectors of S corresponding to the M largest eigen values. Prove the induction that variance of $M+1$ dimensional projection $\boldsymbol{y}_{M+1} = \boldsymbol{W}_{M+1}^T\boldsymbol{x}$ is maximized by choosing $\boldsymbol{W}_{M+1} = [\boldsymbol{W}_M\quad \boldsymbol{u}_{M+1}]$. With this proof, given it is true for $M = 1$, we have PCA solution for any M. (**Points** 10)

2. Prove the following two matrix derivative properties for square symmetric matrices $\mathbf{A}, \mathbf{B}$,

$$\frac{\partial}{\partial \mathbf{A}} \log(|\mathbf{A}|) = 2\mathbf{A}^{-1} - diag(\mathbf{A}^{-1})$$

$$\frac{\partial}{\partial \mathbf{A}} tr(\mathbf{AB}) = 2\mathbf{B} - diag(\mathbf{B})$$

   (**Points** 15)

3. **Fisherfaces -** Sagar is a data scientist who analyzes face images for detecting emotions. In his course, he has leant about LDA and wants to use it to reduce the dimensionality before training a classifier. However, he is faced with a situation where he has N face images each of dimension D with $N << D$. As he knows to apply PCA for high dimensional data, he uses whitening to reduce the dimensionality to $d < N$. The whitenning process can be described as,

$$\boldsymbol{y} = \boldsymbol{\Lambda}^{-\frac{1}{2}}\boldsymbol{W}^T(\boldsymbol{x}-\boldsymbol{\mu})$$

   where $\boldsymbol{x}$ is the input D dimensional image, $\boldsymbol{\mu}$ is the sample mean of input images, $\boldsymbol{W}$ is the PCA projection matrix of dimension $D \times d$, $\boldsymbol{\Lambda}$ is $d \times d$ diagonal matrix containing d largest eigenvalues of sample covariance and $\boldsymbol{y}$ is the whitenned output of dimension d. Given a set of N data points, $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N$ and the corresponding class labels $t_1, t_2, ..., t_N$, (where $t_n = \{1, 2, ..., K\}$, is one of the K-class labels), he tries to learn Fisher LDA projection on the whittened outputs, $\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_N$. Here, let $\boldsymbol{\mu}$ denote the sample mean for the N samples.

(a) As a first step, Sagar tries to find the total covariance (sample covarince) of whitenned outputs y, given as,

$$\boldsymbol{S}_T^y = \frac{1}{N}\sum_{n=1}^{N} \mathbf{y}_n \mathbf{y}_n^T$$

Show that for this case, $\boldsymbol{S}_T^y = \boldsymbol{I}$ where $\boldsymbol{I}$ is the $d \times d$ identity matrix. (**Points** 10)

(b) Assuming that $\boldsymbol{S}_w^y$ is invertible, show that the first LDA projection vector $\boldsymbol{w}$ is given by the eigenvector of $\boldsymbol{S}_w^y$ with minimum magnitude of eigen value. (**Points** 15)

4. **Fischer faces** - Data is posted here
http://leap.ee.iisc.ac.in/sriram/teaching/MLSP_18/assignments/data/Data.tar.gz

   15 subject faces with happy/sad emotion are provided in the data. Each image is of 100x100 matrix. Perform PCA on to reduce the dimension from 10000 to K (using PCA for high dimensional data) and then perform LDA to one dimension. Plot the one dimension features for each image. Select the optimum threshold to classify the emotion and report the classification accuracy on the test data. What is the best choice of K which gives the maximum separability ? (**Points** 25)

5. **Speech spectrogram** - We have clean and noisy speech files here
http://leap.ee.iisc.ac.in/sriram/teaching/MLSP_18/assignments/data/speech.zip
The files are in wav format sampled at $16kHz$. Write a function to compute the spectrogram of clean and noisy files (use 25 ms triangular window with a shift of 10 ms, compute 256 point magnitude FFT and retain the first 128 dimensions in each window, apply log of the magnitude of the FFT.). For example, a speech file of 3s will have a spectrogram of size 128×298 (without any padding at the end, where 128 is the dimension of the feature vector and 298 is the number of frames).

   (a) Assume that each speech frame (256 dimensional vector) is independent of each other. From the clean files, compute the whitening transform. Apply the transform on the noisy speech features. Find the average of the absolute value of the non-diagonal entries of the sampled covariance matrix of the "whitenned" clean and noisy speech features. Comment on use of whitenning when the data distribution has changed.

   (b) Repeat the above procedure by reversing the roles of clean and noisy files.

(**Points** 25)