E9 205 – Machine Learning for Signal Processing

Homework # 2 - Part A
Due date: Oct 3, 2016 (3:30PM)

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** 5)

2. **Maximum Likelihood Classification** Consider a generative classification model with K classes defined by prior probabilities $p(C_k) = \pi_k$ and class-conditional densities $p(\boldsymbol{\phi}|C_k)$ where $\boldsymbol{\phi}$ is the input feature vector. Suppose that a training data is given $\{\boldsymbol{\phi}_n, \boldsymbol{t}_n\}$ for $n = 1, .., N$ and $\boldsymbol{t}_n$ denotes a binary target vector of dimension K with components $t_{nj} = \delta_{j,k}$ if input pattern $\boldsymbol{\phi}_n$ belongs to class k. Assuming that the data points are drawn independently, show that the ML solution for prior probabilities is given by,

   $$\pi_k = \frac{N_k}{N}$$

   where N_k is the number of points belonging to class k. (**Points** 5)

3. **Maximum Likelihood Linear Regression** - Kiran is doing his PhD on acoustic channel estimation. In order to estimate the channel characteristics, he designs an experiment in which he records the speech signal at the source as well as at the output of the acoustic channel. Let $\mathbf{x}_i, i = 1, \ldots, N$ and $\mathbf{y}_i, i = 1, \ldots, N$ denote the feature sequence corresponding to the source and channel outputs.

   (a) He begins with an assumption of a linear model for the channel,

   $$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{b} + \boldsymbol{\epsilon}$$

   where the source features $\mathbf{x}$ are assumed to be non-random and $\boldsymbol{\epsilon}$ represents i.i.d. channel noise $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^{\mathbf{2}}\mathbf{I})$. Given this model, his advisor Mohan recommends the maximum likelihood (ML) method to estimate the parameters of the channel $\big(\mathbf{A}, \mathbf{b}, \sigma\big)$. How will you solve the problem if you were Kiran ? (**Points** 10)

   (b) While Kiran is successful in estimating the parameters of his model, Mohan is unhappy (as usual) with the results when the model is used to approximate a cell phone transmission. Mohan proposes a more complex model where the source speech data $\mathbf{x}_i, i = 1, \ldots, N$ is modeled as a Gaussian mixture model (GMM).

$$\mathbf{x} \sim \sum_{m=1}^{M} \alpha_m \mathcal{N}(\mathbf{x}, \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)$$

Further, the channel is modeled as a linear transformation of the GMM mean components $\hat{\boldsymbol{\mu}}_m = \mathbf{A}_m \boldsymbol{\mu}_m + \mathbf{b}_m$. The covariances are not affected in this model. With the channel ouptuts $\mathbf{y}_i, i = 1, \ldots, N$, how will you help Kiran achieve his PhD faster by solving for the channel parameters $\mathbf{A}_m, \mathbf{b}_m, m = 1, \ldots, M$ assuming that the source signal GMM is already estimated. Simplify your result. (**Points** 15)

4. **Discrete HMM** - Leela is doing a term project on using HMMs as a generative model. She uses a two state discrete HMM. She assumes a simple model with self transition probabilities $a_{11} = 0.8$ and $a_{22} = 0.8$ and initial probability of $\pi_1 = 0.6$. Further, the HMM emits only binary symbols with $b_1(1) = 0$ and $b_2(1) = 1$. Let o_t indicate the symbol emitted at time t. In one of the experiments, she observes $o_3 = 0, o_4 = 0, o_5 = 1$. Find the probability of this observation sequence ? (**Points** 5)