E9 205 – Machine Learning For Signal Processing

Practice for Final Exam 2019

Date: Nov. 25, 2019

Instructions

1. This exam is open book. However, computers, mobile phones and other handheld devices are not allowed.
2. Notation - bold symbols are vectors, capital bold symbols are matrices and regular symbols are scalars.
3. Answer all questions.
4. Total Duration - **180 minutes**
5. Total Marks - **100 points**

Name - .....

Dept. - .....

SR Number - .....

1. Aarush is doing a term project on developing a new dimensionality reduction method that respects the Euclidean distance between data points. He has N data points $\mathbf{X} = \{\mathbf{x}_1, ..., \mathbf{x}_N\}$ where $\mathbf{x}_n \in \mathcal{R}^{\mathcal{D}}$. The data is also centered (Sample mean $\frac{1}{N}\sum_n \mathbf{x}_n = \mathbf{0}$). Using this dataset, he computes the pairwise distance matrix $\mathbf{D}_x$ where $[D_x]_{ij} = ||\mathbf{x}_i - \mathbf{x}_j||^2$. Let the Gram matrix be denoted as $\mathbf{G}_x$ where $[G_x]_{ij} = \mathbf{x}_i^T \mathbf{x}_j$.

   (a) The first result of his term project relates the distance matrix $\mathbf{D}_x$ with the Gram matrix $\mathbf{G}_x$. In particular, he shows that Gram entries $[G_x]_{ij}$ can be computed using only the distance matrix $\mathbf{D}_x$. How does he show this ? (**Points 7**)

(b) He attempts to perform dimensionality reduction of data points. Let $\mathbf{Y} = \{\mathbf{y}_1, ..., \mathbf{y}_N\}$ denote the reduced dimensional data point $\mathbf{y}_n \in \mathcal{R}^d$ and $d < D$ derived using a linear transform $\mathbf{y}_n = \mathbf{A}^T\mathbf{x}_n$ where $\mathbf{A}$ is of size $D \times d$. Let $\mathbf{G}_y$ denote the gram matrix in the reduced dimensional space where $[G_y]_{ij} = \mathbf{y}_i^T\mathbf{y}_j$. He proposes to find the dimensionality reducing transform matrix $\mathbf{A}$ using the following criterion

$$\mathbf{A}^* = argmin_{\mathbf{A}}||\mathbf{G}_x - \mathbf{G}_y||_F^2$$

Can you solve the optimization for $\mathbf{A}$. (**Points 9**)

(c) Bhavana, who has taken the MLSP course, looks at Aarush's solution and says that his optimization result is related to PCA problem of reducing D dimensions to d dimensions. Let $\mathbf{B}$ of size $D \times d$ denote the PCA transform matrix ($\mathbf{B}^T\mathbf{x}_n$ is the PCA projection vector). In particular, she establishes the following two connections,

i. The optimal error in Aarush's model i.e., $min_{\mathbf{A}}||\mathbf{G}_x - \mathbf{G}_y||_F^2$ is equal to the PCA residual error.

ii. The solution of previous problem for $\mathbf{A}$ is also related to the PCA matrix $\mathbf{B}$ using the design matrix $\mathbf{X}$ of size $(D \times N)$.

How is she able to prove this to Aarush ?

(**Points 9**)

2. **RBM and Gaussian Models** - A Gaussian-Bernoulli RBM is defined using the energy function over real visible nodes $\mathbf{v}$ and binary hidden nodes $\mathbf{h}$ as,

$$E[\mathbf{v}, \mathbf{h}] = \frac{1}{2}(\mathbf{v} - \mathbf{a})^T(\mathbf{v} - \mathbf{a}) - \mathbf{b}^T\mathbf{h} + \mathbf{h}^T\mathbf{W}\mathbf{v}$$

The associated p.d.f. is given by $p(\mathbf{v}.\mathbf{h}) = \frac{1}{z}exp\{-E(\mathbf{v}, \mathbf{h})\}$, where z is the normalization constant. Let N denote the number of hidden nodes.

(a) If $N = 0$, show that the marginal distribution of $\mathbf{v}$ becomes a Gaussian distribution with identity covariance. (**Points** 3)

(b) If $p_n(\mathbf{v})$ denotes the marginal distribution of visible nodes obtained for a RBM with n hidden nodes, find a recursive relation between $p_{n+1}(\mathbf{v})$ and $p_n(\mathbf{v})$. (**Points** 9)

(c) Using the recursive relation check whether the marginal distribution of RBM with N nodes resembles a GMM. How many mixture components are present in the marginal distribution of visible nodes for $N = 3$. (**Points** 8)

3. **Mixture Regression Model** Let $\mathbf{x}[n]$, $\mathbf{y}[n]$ denote a vector time series of input observations and target outputs where each $\mathbf{x}[n], \mathbf{y}[n] \in \mathcal{R}^D$ and $n = 0, .., M$. A class of regression models indexed by k are defined as

$$y_j[n] = g_k(x_j[n]) + \epsilon_k[n]$$

where $y_j[n]$ are the predicted time series, g_k denotes a fixed transformation, $\epsilon_k[n]$ denotes Gaussian noise with zero mean and variance σ_k and j is the index of dimension $j = 1, ..., D$. Let $\boldsymbol{\theta}_k$ denote the set of parameters for the k th regression model containing the parameters of g_k and σ_k A conditional probability model for this regression is defined as follows,

$$p(\mathbf{y}_j|\mathbf{x}_j, \boldsymbol{\theta}_k) = \prod_{n=1}^{M} f_k(y_j[n]|x_j[n], \boldsymbol{\theta}_k)$$

where f_k is the appropriate density function. The mixture regression model is then defined as,

$$p(\mathbf{y}_j|\mathbf{x}_j, \boldsymbol{\theta}) = \sum_{k=1}^{K} w_k p(\mathbf{y}_j|\mathbf{x}_j, \boldsymbol{\theta}_k)$$

where $\boldsymbol{\theta}$ consists of the parameters of all K regression models. Further, assume that each dimension $j = 1, ..., D$ is conditionally independent of each other ($p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}) = \prod_{j=1}^{D} p(\mathbf{y}_j|\mathbf{x}_j, \boldsymbol{\theta})$). If $\mathbf{X}, \mathbf{Y}$ denotes the $D \times M$ matrix of all the input and output vector time series respectively, prove the following,

(a) Find the expression for the joint conditional probablility $p(\mathbf{Y}|\mathbf{X}, \boldsymbol{\theta})$ and log likelihood function $L(\boldsymbol{\theta})$ for this model. (**Points** 5)

(b) Make your choice of latent variables (indicated by $\mathbf{Z}$) to solve the EM algorithm for this problem. Find the joint log likelihood ($log[p(\mathbf{Y}, \mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})]$) and the expression for the EM Q function in terms of current estimates of the model parameters $\boldsymbol{\theta}^t$. (**Points** 5)

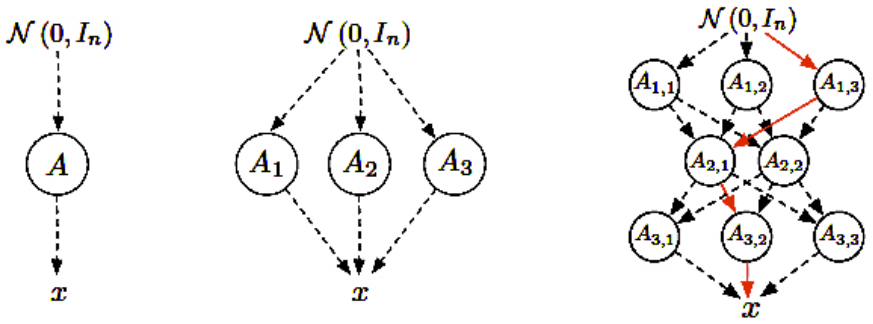
(c) For the p order polynomial regression, the function g_k is defined as,

$$y_j[n] = \sum_{q=0}^{p} \beta_{kq}(x_j[n])^q + \epsilon_k$$

Let $\boldsymbol{\beta}_k = [\beta_{k0}, ..., \beta_{kp}]^T$ denote the parameters of the regression model. Find the EM algorithm for iteratively updating all the parameters of the model $\boldsymbol{\theta}_k = \{\boldsymbol{\beta}_k, \sigma_k\}$ and w_k for $k = 1, ..., K$.

(**Points** 10)

4. **Deep GMM** A Gaussian distribution, a GMM and a deep GMM architecture is graphically shown above. All the layers realize linear connections and the tranformation is given by the matrix indexed in the node. There is an additional bias term at each node that is not shown in the graphical illustration.



(a) Gaussian (b) GMM (c) Deep GMM

For example, for figure (a), the marginal distribution is obtained by linear transformation of a standard Gaussian, i.e., $(p(\mathbf{x}) = \mathcal{N}(\mathbf{x}, \mathbf{b}, \mathbf{A}\mathbf{A}^T)$. For the figure in (b), the marginal distribution is given by,

$$p(\mathbf{x}) = \sum_{i=1}^{3} \pi_i \, \mathcal{N}(\mathbf{x}, \mathbf{b}_i, \mathbf{A}_i \mathbf{A}_i^T)$$

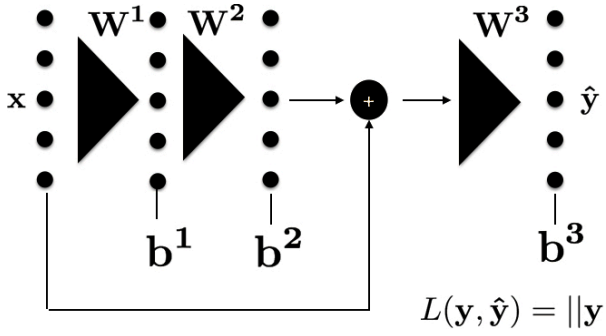
For figure (c), a deep GMM architecture is depicted and the probablity density is based on the path chosen. Let $\boldsymbol{\phi}$ denote the set of paths through the network with each path p containing 3 steps (3 layer deep GMM). For example, $\{(1,3),(2,1),(3,2)\}$ represents one possible path. Also, let $\mathbf{b}_{i,j}$ denote the bias vector of layer $i = 1, 2, 3$ and node j (The corresponding transform matrices $\mathbf{A}_{i,j}$ are shown in the figure. Let π_p denote the path probablity, then the marginal distribution in this case is given by,

$$p(\mathbf{x}) = \sum_{p \in \boldsymbol{\phi}} \pi_p \, \mathcal{N}(\mathbf{x}, \boldsymbol{\mu}_p, \Sigma_p \Sigma_p^T)$$

(a) Find the expression for $\boldsymbol{\mu}_p, \Sigma_p$ for any choice of path. (**Points** 3)

(b) Formulate an iterative algorithm using backpropagation to update the model parameters that maximize the log-likelihood. Does it have similarities to EM algorithm. (**Points** 17)

5. **Skip connection DNN** An image enhancement DNN is shown below for $\mathbf{x}, \mathbf{y} \in \mathcal{R}^D$.



The output units have a linear activation function and hidden units realize a ReLU activation function. All the weight matrices are square matrices as well. For this DNN architecture, derive the backpropagation update rule for all the parameters of the model $\mathbf{W}^1, \mathbf{W}^2, \mathbf{W}^3, \mathbf{b}^1, \mathbf{b}^2, \mathbf{b}^3$. (**Points 10**)

6. If θ denotes the angle between two vectors $\phi(\mathbf{x})$ and $\phi(\mathbf{z})$, check whether $\cos(\theta)$ is a valid kernel. Justify your answer using either definition of valid kernels or using the kernel rules.
(**Points 5**)