E9 205 – Machine Learning For Signal Processing

Take Home Exam # 2

Date: Nov. 30, 2016

Instructions

1. This exam is take home open book. However, the use of computers, mobile phones and other handheld devices are not allowed.
2. Academic integrity is expected and all problems are required to be solved individually without any consultation.
3. Notation - bold symbols are vectors, capital bold symbols are matrices and regular symbols are scalars.
4. Take the examination once you have a reasonable level of preparation.
5. Answer all questions.
6. The questions are labelled easy, medium and hard as requested by some students. If the solutions are proving difficult, you are highly encouraged to discuss the solutions with TA/Instructor. Also check the correctness of your solutions with TA.
7. Max Points = 100. Do a self evaluation once you are done. You may also request the TA to note down your marks.

Name - ....

Dept. - ....

SR Number - ....

1. **Bayesian estimation**

Ajay is a stock market consultant where he advises clients on buying, holding and selling of stocks. He deals with various categories of stocks like IT, oil, power, electronics etc. In order to improve his prospects at his job, he decides on using data models and collects the market data containing value of stock, volume of shares sold, number of shares bought etc for the past 30 days. He also hires Diya who just completed her masters in signal processing where she has learnt about maximum likelihood (ML) estimation and Gaussian mixture models (GMMs). At the outset of her job, she suggests modeling the market data for all the categories denoted as $\mathbf{X} = \mathbf{x}_i, i = 1, ..., N$ using a GMM $\lambda = \{\alpha_c, \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c\}_{c=1}^{C}$. Given only a small number of samples ($N = 30$), she decides on using a simple covariance matrix $\boldsymbol{\Sigma}_c = \frac{1}{\tau_c}\mathbf{I}$. While Diya was initially excited about the application of the theory, she finds that many predictions from her model were far away from the actual value. Ajay tries to understand the issue and learns from Google that Bayesian estimation methods are more stable than ML methods. He asks Diya to look into Bayesian techniques. Diya does some reading and comes up with a prior distribution for mixture weights $\boldsymbol{\alpha} = \{\alpha_c\}_{c=1}^{C}$ given by the Dirichlet density,

$$p(\boldsymbol{\alpha}) \propto \prod_{c=1}^{C} \alpha_c^{\zeta_c - 1}$$

where ζ_c are parameters of the Dirichlet density and $\zeta_c > 0$. Also, for mixture component means she prefers to use a Gaussian density given by,

$$p(\boldsymbol{\mu}_c) \propto \exp\{-\frac{\rho_c}{2}(\boldsymbol{\mu}_c - \boldsymbol{m}_c)^*(\boldsymbol{\mu}_c - \boldsymbol{m}_c)\}$$

where $\rho_c > 0$ and $\boldsymbol{m}_c$) are the parameters of the Gaussian distribution. She assumes the mixture component means to be independent of the mixture weights. Further, she believes that reestimating the means and weights would be good enough to improve her predictions and therefore keeps the covariances obtained previously from the ML method. Let $\boldsymbol{\Theta} = \{\alpha_c, \boldsymbol{\mu}_c\}_{c=1}^{C}$ denote the parameters of interest and she prefers to estimate the parameters using the MAP rule $\arg\max_{\boldsymbol{\Theta}} \; p(\boldsymbol{\Theta}|\mathbf{X})$.

(a) Diya sits down to formulate the parameter estimation problem. Since she likes the iterative EM algorithm, she wants to convert the Bayesian estimation problem to an equivalent EM formulation. What would be your suggestion and how do you justify this ? (**Points 5**)

(b) Diya managed to find the equivalent forumlation. Now she does some mathematical analysis and to her delight finds that her choice of prior distributions for $\boldsymbol{\Theta}$ obeys the conjugate density property (the EM style lower bound for the posterior and the prior distribution belong to the same class of densities). How did she achieve this ? (**Points 10**)

(c) Diya proceeds to find the solution to the parameter estimation problem. While she knows that the result should contain terms from the prior density and the ML estimation, she is stuck at the maximization part. How would you help her solve this and keep her happy at her job ? (**Points 5**)

Category - Hard

2. **Restricted Boltzmann Machine** - The Gaussian Bernoulli RBM is defined using visible units $\mathbf{v}$, hidden units $\mathbf{h}$. The energy function and the joint probability density function are given by,

$$\begin{aligned} E(\mathbf{v},\mathbf{h}) &= 0.5(\mathbf{v}-\mathbf{a})^T(\mathbf{v}-\mathbf{a}) - \mathbf{b}^T\mathbf{h} - \mathbf{h}^T\mathbf{W}\mathbf{v} \\ P(\mathbf{v},\mathbf{h}) &= \frac{e^{-E(\mathbf{v},\mathbf{h})}}{Z} \end{aligned}$$

where Z is a normalization constant. Show that conditional probability of visible units given the hidden units is Gaussian - $P(\mathbf{v}|\mathbf{h}) \sim \mathcal{N}(\mathbf{v}, \mathbf{W}^T\mathbf{h} + \mathbf{a}, \mathbf{I})$. (**Points** 10)

Category - Easy

3. **Ensemble of DNNs** Tarun and Asha are two students doing a term project with deep neural networks (DNNs). Their task is to use DNNs for approximating a scalar function $h(\mathbf{x})$. In order to understand the impact of different initializations, number of hidden units etc, they train N different DNNs and compare the errors obtained on a held out set. Assume that every DNN makes a random error $\epsilon_i(\mathbf{x}) = (v_i(\mathbf{x}) - h(\mathbf{x}))$. Let $J_i, i = 1..., N$ denote the expected error of i th DNN,

$$J_i = \mathbb{E}\big[(v_i(\mathbf{x}) - h(\mathbf{x}))^2\big]$$

where $v_i(\mathbf{x})$ denotes the output of the i th DNN. Tarun proposes an idea of combining the outputs from N DNNs using a simple averaging.

$$v_A(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^{N} v_i(\mathbf{x})$$

He claims that expected error J_A by using the average output v_A is better than the average expected error , i.e., $J_A \le J_{avg}$, where,

$$J_{avg} = \frac{1}{N}\sum_{i=1}^{N} J_i$$

(a) Asha is not convinced. She takes the simple case when networks are all unbiased and uncorrelated, i.e., $\mathbb{E}[\epsilon_i(\mathbf{x})] = 0$ and $\mathbb{E}[\epsilon_i(\mathbf{x})\epsilon_j(\mathbf{x})] = 0 \ \forall i \neq j = 1, ..., N$. She is able to prove Tarun's claim for the simple case. Would you also be able to ? (**Points 5**)

(b) Asha now considers a generic and more realistic case when the errors from individual DNNs are correlated. She is unable to prove the result for the generic case. But Tarun still argues that it is indeed true that J_A is better than J_{avg}. Do you think Tarun is right ? Justify your answer.
Remark - Cauchy's Inequality states $\left(\sum_k a_k b_k\right)^2 \le \sum_k a_k^2 \sum_k b_k^2$ (**Points 5**)

(c) When they take their discussion and findings to Prof. Raj, he modifies their solution to a weighted average,

$$v_W(\mathbf{x}) = \sum_{i=1}^{N} \alpha_i v_i(\mathbf{x})$$

where α_i are positive constants and $\sum_{i=1}^{N} \alpha_i = 1$. He claims that the optimal value of these constants can be found using the correlation matrix of errors $\boldsymbol{C} = [C_{ij}]$ where $C_{ij} = \mathbb{E}[\epsilon_i \epsilon_j]$. Tarun and Asha are asked to find these optimal values for weights α_i which minimize the expected error J_W for the output $v_W(\mathbf{x})$ assuming that the covariance matrix $\boldsymbol{C}$ is full rank. What would be your solution for these weights ? (**Points 5**)

(d) After a month, Tarun and Asha not only find their solution for weights, but also analytically show that expected error J_W using the outputs $v_W(\mathbf{x})$ is better than the expected error J_{avg}. Further, they also show that J_W is also better than the weighted average $J_{wavg} = \sum_{i=1}^{N} \alpha_i J_i$, i.e.,

$$J_W \leq J_{wavg}$$

They take their analysis to Prof. Raj who is also excited and suggests that they write a paper. How did they arrive at this delightful situation ? (**Points 5**)

Category - Medium

4. **Reverse linear prediction** - A modification of linear prediction is to predict the present sample from future samples. Let $x[n]$ be a discrete sequence. The reverse linear prediction is the process of predicting $x[n-L]$ from $x[n], x[n-1], .., x[n-L+1]$. The error in reverse linear prediction is given by

$$e_{n,L} = x[n-L] - \hat{x}[n-L]$$
$$\hat{x}[n-L] = \sum_{l=1}^{L} a_{L,l} x[n-l+1]$$

Find the normal equations which minimize the expected squared prediction error $\mathbb{E}[|e_{n,L}|^2]$. How is this different from the forward linear prediction discussed in the class ? (**Points** 10)

Category - Easy

5. **Modified HMM** - In the conventional HMM, the association between the state sequence and observation sequence is not taken into account. In a modified formulation, we would like to add this component in the modeling. Let $H(\mathbf{O}|\mathbf{q})$ denote the conditional entropy of the observation sequence $\mathbf{O} = \{\mathbf{o}_t\}_{t=1}^{T}$ given the state sequence $\mathbf{q} = \{q_t\}_{t=1}^{T}$. The conditional entropy is defined as,

$$H(\mathbf{O}|\mathbf{q}) = -\sum_{t}\sum_{q_t}\int_{\mathbf{o}_t} P(\mathbf{o}_t, q_t) \log\big(P(\mathbf{o}_t|q_t)\big)\partial\mathbf{o}_t$$

In the HMM formulation, we would like to minimize the conditional entropy (maximizing the mutual information). Thus, we modify the ML method in the following manner,

$$J = -(1-\epsilon)\ H(\mathbf{O}|\mathbf{q}) + \epsilon\ \log P(\mathbf{O})$$

where ϵ is a positive constant whose maximum value is 1 and J is the function to be maximized to determine the parameters. We assume continuous density HMM with state emission probabilities as single Gaussian distribution ($b_j(\mathbf{o}_t) \sim \mathcal{N}(\mathbf{o}_t, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$). Derive the update equations for mean and covariance of the emission probabilties for the modified HMM. (**Points** 20)

Category - Hard

6. **2-norm margin support vector** - A modification of the standard support vector machine formulation involves the 2-norm of the slack variables ζ_n. The 2-norm margin SVM formulation is defined by the objective function

$$min\ \frac{1}{2}\mathbf{w}^T\mathbf{w} + \frac{C}{2}\sum_n \zeta_n^2$$

Using this modification, derive the primal and dual objective functions (the variables are defined according the formulation given in class) and the KKT conditions. (**Points** 10)

Category - Easy

7. **Discriminant Analysis Revisited** - Vikas has been working with restricted Boltzmann machines for his masters thesis where he finds probablity of node in the hidden layer being active is given by a sigmoidal function. Arathy who took MLSP course suggests that sigmoidal activations exist even in discriminant analysis. To prove her case, she denotes $\mathbf{x}$ as the input data with binary class label $y = \{0, 1\}$. Further she assumes a Bernoulli distribution for y given by $p(y) = \phi^y(1-\phi)^{1-y}$ and Gaussian class conditional distribution $p(\mathbf{x}|y = 0) = \mathcal{N}(\mathbf{x}; \mu_0, \Sigma)$ and $p(\mathbf{x}|y = 1) = \mathcal{N}(\mathbf{x}; \mu_1, \Sigma)$. Using these assumptions, Arathy claims that $p(y = 1/\mathbf{x})$ is a sigmoidal function. Is she right in her claim ? If so, find the exact sigmoidal function. Also, given a set of data points $\{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N\}$ and their corresponding labels $\{y_1, y_2, .., y_N\}$, find the maximum likelihood estimates of the parameters of Arathy's model. (**Points** 10)

Category - Medium