Semester-end Examination (CS-419)

25-Apr-2014 (2:00pm-5:00pm)

Important Instructions

- Fill the blanks in the questions, in place[1], with as concise and legible answers as possible.
- The blanks in the questions are of sufficient size to accommodate the expected answer. Hence, answers that go well beyond the blank, and/or those that are not legible, and/or those that written in a very tiny font-size, will not be evaluated.
- While answering please use the notation/terminology/named-algorithms/theorem-results as mentioned or referred-to in the lectures notes.
- For the sake of preciseness in the questions, next to every blank a keyword in [[...]] format is written that indicates the type of answer I am expecting:
  - if I am expecting a (real/integer/natural) number as the answer, I will write [[NUMBER]]. In case the number is a rational you may write it as a fraction, in its most simplified form, or you may write it in decimal notation. No marks will be awarded if expressions/formulae are provided instead of numbers.
  - if I am expecting an answer that is a technical term defined or used in our course, I will write [[TERMINOLOGY]]. Further, if the term is a name of an algorithm or a theorem or a model etc., then I would write [[NAME]].
  - if I am expecting mathematical or a logical expression(s)/statement(s) as an answer, I will write [[EXPR]].
  - if I am expecting you to pick the most appropriate choice from, say Choice1, Choice2, Choice3, then I will write [[Choice1/Choice2/Choice3]].
- Marks for questions/blanks are mentioned. Note that these marks are atomic. For e.g., if I mention 4 marks, then you will get 4 if all the answers for the corresponding blanks are absolutely correct and 0 otherwise. So please be very careful in writing your final answers. Sometimes I may mention for e.g., 2+2 Marks after a question consisting of 2 blanks. This means each blank is of 2 (atomic) marks.
- You should NOT carry anything with you other than pens/pencils. If you are caught copying or showing your answers to others or using any other unfair means, then you will get an FR in the course and your case will be reported to appropriate disciplinary committee.

[1] There is no separate answer sheet. You will only be given this question paper and a rough sheet. You should return the question paper containing your answers and keep the rough sheet with you.

Fill in the blanks

1. Consider a machine learning application for which the following background knowledge is available from the domain experts:

   B1 "The output variable is definitely a linear function of the two input variables x_1, x_2."

   B2 "Moreover, it is more likely that it is a linear function of x_1 alone."

   - Now suppose you were to do probabilistic modeling of this problem. Then you would use a linear regression model, so that the information **B1** is utilized. And further employ a suitable prior (over the parameters) so that the information **B2** is utilized.

     [1+1 marks]

   - Now suppose you were to do deterministic modeling for the same problem, which leads to a prediction function that is dependent on a few training examples. Then you would use the support vector regression formalism such that the information **B1** is utilized. And further employ a suitable hierarchy (over models) so that the information **B2** is utilized.

     [1+2 marks]

2. Consider the Gaussian mixture model with n components, denoted by GMMn. Assume that parameter selection is done using the EM algorithm discussed in the lecture. Let us denote the distribution in GMM2 selected using the EM algorithm by gmm2 and that in GMM3 by gmm3. Then, the likelihood of the training data computed using the gmm2 distribution is not comparable to that computed with gmm3.

   [2 marks]

   Explanation: This is because EM algorithm need not necessarily maximize the likelihood.

3. In context of the above problem, now assume that it so happens that the likelihood of the training data is exactly same for both gmm2 and gmm3. Given this, if you are forced to choose one of gmm2 or gmm3 as the predictive distribution, then you will pick gmm2.

   [2 marks]

4. Consider the following binary classification[2] training data

$$\mathcal{D} = \left\{ \left(\begin{bmatrix} 1 \\ 0 \end{bmatrix}, +1 \right), \left(\begin{bmatrix} -1 \\ 0 \end{bmatrix}, +1 \right), \left(\begin{bmatrix} 0 \\ 1 \end{bmatrix}, -1 \right), \left(\begin{bmatrix} 0 \\ -1 \end{bmatrix}, -1 \right) \right\}$$

   and the homogeneous quadratic model, which is the set of all functions of the form: $g(x) = w^\top \phi(x)$, where $w = \begin{bmatrix} w_1 \\ w_2 \\ w_3 \end{bmatrix}$ is the model parameter, and $\phi(x) = \begin{bmatrix} x_1^2 \\ x_2^2 \\ x_1 x_2 \end{bmatrix}$. Then, optimization problem corresponding to the hard-margin SVM[3], discussed in the lecture, for choosing the optimal (homogeneous) quadratic discriminator is:

$$\begin{aligned} \min_{w_1,w_2,w_3} \quad & \underline{\frac{1}{2}\|w\|^2}, \\ \text{s.t.} \quad & \underline{w_1 \geq 1, w_3 \leq -1}. \end{aligned}$$

   Note that you need to fill the above two blanks with expressions involving w alone[4].

   [3 marks]

   Solve the above optimization problem for the optimal w. The equation[5] for the discriminating

[2] Labels are +1 or -1.

[3] Hard-margin SVM is same as the SVM presented in Murphy's book where all slack variables are set to zero, i.e., $\xi_i = 0 \ \forall \ i$.

[4] The expression should not involve ϕ or x etc.

[5] Note that your expression should not involve ϕ or x or w etc.

quadratic surface with this optimal w is $\underline{x_1^2 - x_2^2 = 0}$.

[2marks]

5. A coin, with unknown probability of heads, was tossed 5 times and it was head only twice. Assume two Beta-Bernoulli[6] models are available: one with hyperparameters $a = 3, b = 3$, denoted by $\mathcal{M}_1$, and the other with hyperparameters $a = 1, b = 1$, denoted by $\mathcal{M}_2$. Let $\hat{m}_i$ denote that distribution in $\mathcal{M}_i$, which is chosen according to maximum likelihood principle. Then the likelihood of the training data with $\hat{m}_1$ is $\underline{0.03456}$ and that with $\hat{m}_2$ is $\underline{0.03456}$.

[1mark]

Let m_i denote that distribution in $\mathcal{M}_i$, which is chosen according to MAP principle. Then the likelihood of the training data with m_1 is $\underline{0.033870176}$ and that with m_2 is $\underline{0.03456}$. Hence the likelihood with m_1 is $\leq$ that with m_2.

[2marks]

The likelihood of the training data with the BAM corresponding to $\mathcal{M}_1$ is $\underline{0.033529751}$ and that corresponding to $\mathcal{M}_2$ is $\underline{0.034271435}$. Among these two numbers, the former is $\leq$ the latter.

[2marks]

The marginal likelihood of $\mathcal{M}_1$ is $\underline{0.002164502}$ and that of $\mathcal{M}_2$ is $\underline{0.002380952}$. Hence, the maximum marginal likelihood principle will select $\underline{\mathcal{M}_2}$.

[2marks]

6. Let $\mathcal{M}$ denote a model consisting of all distributions with pdf/pmf given by f_ψ for the various values of the parameters $\psi \in \Psi$. Now consider this definition: the model $\mathcal{M}$ is said to belong to the exponential family iff there exist the following:

   - a, perhaps modified, parameterization of the pdf/pmf in terms of parameters $\theta \in \Theta \subset \mathbb{R}^d$. In other words, consider $g : \Psi \mapsto \Theta$ and $\theta \equiv g(\psi)$. Then the new parameterized pdf/pmf is given by $\hat{f}_\theta(x) \equiv f_\psi(x)\ \forall\ x \in \mathcal{X} \subset \mathbb{R}^n$.
   - a function $h : \mathbb{R}^n \mapsto \mathbb{R}_+$,
   - a function $\phi : \mathbb{R}^n \mapsto \mathbb{R}^d$,

   such that $f_\psi(x) \equiv \hat{f}_\theta(x) = \frac{1}{Z(\theta)} h(x) \exp\{\theta^\top \phi(x)\}$, where $Z(\theta)$ is simply the normalization factor[7]. It turns out that many models familiar to you belong to this family:

   Multinoulli model: Let ψ_i denote the probability that X takes value i, for all $i = 1, \ldots, 3$. Let $I(x, i)$ denote 0 if $x \neq i$ and denote 1 if $x = i$. Once this multinoulli's pmf is written in the exponential form,

$$\theta = \underline{\left[\log\left(\frac{\psi_1}{\psi_3}\right)\ \log\left(\frac{\psi_2}{\psi 3}\right)\right]^\top},\ \Theta = \underline{\left\{\theta = g(\psi) \mid \psi_1 + \psi_2 + \psi_3 = 1,\ \psi_i \geq 0\ \forall\ i = 1, 2, 3\right\}}$$

$$\phi(x) = \underline{\left[\text{I(x,1) I(x,2)}\right]^\top},\ Z(\theta) = \underline{1 + \exp(\theta_1) + \exp(\theta_2)},\ h(x) = \underline{1}.$$

   Alternative answers are possible with 3-size vectors etc., which some of you have written correctly..

[6] The pdf for Beta distribution is given by: $p(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}$, where $a > 0, b > 0$. Recall that the Gamma function satisfies: $\Gamma(a+1) = a\Gamma(a)$.

[7] For conts. rvs., it is given by $Z(\theta) = \int_{\mathcal{X}} h(x) \exp\{\theta^\top \phi(x)\}\text{dx}$ and for discrete it is $\sum_{x \in \mathcal{X}} h(x) \exp\{\theta^\top \phi(x)\}$.

[3marks]

Gaussian model: Let $\mu \in \mathbb{R}$ denote its mean and let σ^2 denote its variance. Once this Gaussian's pdf is written in the exponential form,

$$\theta = \left[\frac{\mu}{\sigma^2} \;\; \frac{-1}{2\sigma^2}\right]^\top, \;\; \Theta = \underline{\mathbb{R} \times \mathbb{R}_-}$$
$$\phi(x) = [\underline{x \; x^2}]^\top, \;\; Z(\theta) = \underline{\sqrt{\frac{-\pi}{\theta_2}} \exp\left(\frac{-\theta_1^2}{4\theta_2}\right)}, \;\; h(x) = \underline{1}.$$

Alternative answers are possible with 3-size vectors etc., which some of you have written correctly..

[3marks]

More commonly, each entry of $\phi(x)$ is called as a sufficient static for x (and hence $\phi(x)$ is the vector of sufficient statistics for x) and $Z : \Theta \mapsto \mathbb{R}$ is called as the partition function. Interestingly, it turns out that $\log(Z(\theta))$ is a convex function in θ and the conjugate prior turns out be exponential again[8]. You may prove these leisurely sometime later after this examination. Infact, owing to these two facts, the expressions related to MLE, MAP, BAM turn out to be extremely elegant.

Now let $\mathcal{F}$ denote a particular model that belongs to the exponential family and $\theta \in \Theta$ represents its model parameters. Consider a binary classification problem, with class labels represented by +1 and -1. Assume that the class-conditionals are modeled using $\mathcal{F}$ and the class prior is modeled using the Bernoulli model. Let θ_{+1} and θ_{-1} be the optimal parameters chosen according to MLE for the class-conditionals of classes $+1$ and -1 respectively. Let α be that selected by MLE for the class-prior and represents the prior probability of class $+1$. Then the equation of the discriminating surface is given by:

$$\underline{(\theta_{+1} - \theta_{-1})^\top \phi(x) + \log\left(\frac{\alpha Z(\theta_{-1})}{(1-\alpha)Z(\theta_{+1})}\right) = 0.}$$

In case $\mathcal{F}$ is the Gaussian model, then this is indeed a quadratic surface.

[2marks]

Observe that there exists a transformation $\zeta : \mathcal{X} \mapsto \mathbb{R}^{d+1}$ such that the form of the distribution $p(y/x)$ with the above described exponential model based generative model is exactly same as that with logistic regression over the transformed data $\zeta(x)$. This transformation is given by $\zeta(x) = \left[\underline{\phi(x)^\top \;\; 1}\right]^\top$. Also, the relation between w, the parameter of logistic regression and $\theta_{+1}, \theta_{-1}, \alpha$ is given by

$$w = \underline{\left[\theta_{+1}^\top - \theta_{-1}^\top \quad \log\left(\frac{\alpha Z(\theta_{-1})}{(1-\alpha)Z(\theta_{+1})}\right)\right]^\top}.$$

Let us refer to this as non-linear logistic regression.

[2marks]

Non-linear logistic regression can be easily generalized to multi-class classification. The only difference is that there will be one w for each class[9]. Now consider the generative model, HMM, where emission distributions are modeled by $\mathcal{F}$ (parameterized by θ) and π, A represent the vector of initial state probabilities and the state transition probabilities matrix respectively. Provide expressions with the corresponding non-linear logistic regression for:

$$\zeta(x) = \zeta(x_1, \ldots, x_T) = \underline{\left[\phi(x_1)^\top \;\; \ldots \;\; \phi(x_T)^\top \;\; 1\right]^\top}$$

[8]This is sometimes called as self-conjugacy.

[9]As in linear logistic regression, if there are k classes, then instead of k number of ws, we can use $k-1$. To keep notation simple, let us use k number of ws only.

and

$$w_y = w_{y_1,\ldots,y_T} = \underline{\left[\theta_{y_1}^\top \ \ldots \ \theta_{y_T}^\top \ \log\left(\frac{\pi(y_1)A(y_1,y_2)\ldots A(y_T,yT-1)}{Z(\theta_{y_1})\ldots Z(\theta_{y_T})}\right)\right]^\top}.$$

[3marks]

It may at first appear that there are too many ws, one for each state sequence. But a close observation will reveal that they are related, as given by your expression above (in the blank), and essentially only the (π, A, θ) are the free variables. An alternative to the above, popularly called as Conditional Random Fields (CRFs), is to assume that the $p(y/\zeta(x))$ itself factorizes, say as $p(y/\zeta(x)) = p(y_1/\zeta(x))p(y_2/y_1, \zeta(x))\ldots p(y_T/y_{T-1}, \zeta(x))$. The advantage with CRF is that the number of parameters is itself low (and hence parameter learning is less messy). If the number of states is k, then the number of w parameters with CRF is $\underline{k^2 + k}$. Alternative answer is possible for the last blank as $k^2 - 1$..

[1mark]

7. Consider an unsupervised learning problem where U_X is the unknown distribution and $X \in \mathbb{R}^n$. Suppose you were to model U_X using a multivariate Gaussian with parameters $\theta = (\mu, \Sigma)$. From the lectures, this case is very familiar to you. Recall that in this case the MLE estimates for μ and Σ turn out to be the sample mean and sample covariance.

   Now consider a more practical situation[10] where some feature values are missing for some training examples[11].

   In general, let us denote the observed part of a sample x_i by x_{io} and its missing/hidden part by x_{ih}. Hence the training data is simply the set of $x_{io}, i = 1, \ldots, m$. Let us still assume that the x_is are iid samples of U_X. Also, suppose we still want to model U_X using a multivariate Gaussian. Needless to say, while writing the expression for likelihood you would want to involve the hidden/missing variables. The expression for log-likelihood of the training data is $\log(p_\theta(\mathcal{D})) = \sum_{i=1}^m \log(p_\theta(x_{io})) =$

$$\underline{\sum_{i=1}^m \log(\int_{\mathcal{X}_{ih}} p_\theta(x_{io}, x_{ih}) \ \mathrm{d}x_{ih})}$$

   (write an expression involving x_{ih}s).

   [2marks]

   Now suppose you want to employ the EM algorithm for parameter selection. Let us assume t iterations of it are performed and the parameter after this iteration is θ_t. The q_{t+1} distribution you would choose will then be given by:

$$\underline{q_{t+1}(x_{ih}) = p_{\theta_t}(x_{ih}/x_{io})}.$$

   Hint: Recall that log is a concave function and hence satisfies the so called Jensen's inequality $\log(\mathbb{E}[Z]) \geq \mathbb{E}[\log(Z)]$, where Z is *any* random variable[12] such that the involved expectations are finite.

[10] You would have observed that some real-world datasets in the UCI repository (the online repository you downloaded the datasets for your practical assignments) do have missing feature values.

[11] Here is an example of such a 3-dimensional data: $\mathcal{D} = \left\{ \begin{bmatrix} 2.5 \\ ? \\ 3.4 \end{bmatrix}, \begin{bmatrix} 5 \\ 3.3 \\ 8 \end{bmatrix}, \begin{bmatrix} 0.1 \\ ? \\ ? \end{bmatrix}, \begin{bmatrix} ? \\ 3 \\ 1 \end{bmatrix} \right\}$. '?' represents missing datum.

[12] In lectures we used a special case of Jensen's inequality where Z is discrete. Note that when Z is Bernoulli, then the Jensen's inequality provides the definition of a concave function.

[3marks]

After this examination, at leisure, write down the entire EM algorithm for this missing value problem.