Probabilistic Machine Learning (CS772A)
Homework 2 (Due date: Sept 16, 2017, 11:59pm)

Instructions

- We will only accept electronic submissions and the main writeup must be as a PDF file. If you are handwriting your solutions, please scan the hard-copy and convert it into PDF. Your name and roll number should be clearly written at the top. In case you are submitting multiple files, all files must to be zipped and **submitted as a single file** (named: your-roll-number.zip). Please do not email us your submissions. Your submissions have to be uploaded at the following link: `https://tinyurl.com/y9mb7m8x`).
- Each late submission will receive a 10% penalty per day for up to 3 days. No submissions will be accepted after the 3rd late day.

Problem 1 (20 marks)

(Constructing Richer Priors) Suppose we have a prior $p(\theta)$ over some parameter θ and suppose we can write $p(\theta)$ as a convex combination of K conjugate priors, i.e., $p(\theta) = \sum_{k=1}^{K} p(z = k)p(\theta|z = k)$, where $\sum_{k=1}^{K} p(z = k) = 1$ and each $p(\theta|z = k)$ is a conjugate prior w.r.t. a likelihood model $p(\mathcal{D}|\theta, z = k)$, where $\mathcal{D}$ denotes the observed data.

Suppose you observe some data $\mathcal{D}$ from a model with parameters θ. Show that the posterior $p(\theta|\mathcal{D})$ can also be written as a convex combination of *conjugate* distributions.

Does this property make $p(\theta)$ a conjugate to the likelihood $p(\mathcal{D}|\theta)$?

Problem 2 (20 marks)

(Regression using Generative Models) The probabilistic linear regression model where we directly model a real-valued response as $p(y|\boldsymbol{x}, \boldsymbol{w}, \beta) = \mathcal{N}(y|\boldsymbol{w}^\top \boldsymbol{x}, \beta^{-1})$ can be termed *discriminative* since only y is modeled, not $\boldsymbol{x}$. Suppose we want to construct a *generative* approach to probabilistic linear *regression*. In this approach, instead of directly modeling $p(y|\boldsymbol{x})$, we would like to first model the joint distribution $p(\boldsymbol{x}, y)$ and then use this distribution to obtain our linear regression model $p(y|\boldsymbol{x})$. Assume $\boldsymbol{x}$ to be real-valued vector, i.e., $\boldsymbol{x} \in \mathbb{R}^D$.

How would you construct this generative regression model? In particular,

- What distribution would be use for $p(\boldsymbol{x}, y)$ and what would be its parameters? Justify your choice.
- How would you get the actual distribution of interest, i.e., $p(y|\boldsymbol{x})$? What would be its parameters?
- Assuming you are given training data in form of N examples $\{\boldsymbol{x}_n, y_n\}_{n=1}^{N}$, give the MLE solution for the parameters involved in this approach.

Problem 3 (10 marks)

(Exp. Family?) Consider a K component GMM for some data with one-dimensional observations $x \in \mathbb{R}$. Assume the mean and variance of component k to be (μ_k, σ_k^2) and its mixing proportion to be π_k. Is the complete data likelihood $p(x, z)$ when z is observed an exponential family distribution? If yes, why? If no, why not?

Problem 4 (25 marks)

(Mixture Model for Binary Vectors) You have N observations $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$, each of which is in form of a binary vector of size D. Assume these N observations are drawn from a mixture model with K components. What would be an appropriate mixture model for this type of data? Derive the EM algorithm to infer the latent variables and the MAP (not MLE) estimate for the parameters of this model. Clearly mention all the steps of the algorithm. You may assume appropriate prior distributions on the latent variables and the parameters. You may assume the hyperparameters to be fixed.

Problem 5 (25 marks)

(EM for Hyperparameter Estimation) In the class, we looked at the MLE-II approach for point estimation of hyperparameter for the probabilistic linear regression model. The MLE-II approach to learning the hyperparameters λ, β for this model reduces to doing MLE on marginal likelihood $p(\boldsymbol{y}|\mathbf{X}, \lambda, \beta) = \int p(\boldsymbol{y}, \boldsymbol{w}|\mathbf{X}, \lambda, \beta)p(\boldsymbol{w}|\lambda)d\boldsymbol{w}$.

Another approach would be to use the Expectation Maximization (EM) algorithm where instead of doing MLE on (log)$p(\boldsymbol{y}|\mathbf{X}, \lambda, \beta)$, we estimate λ, β by doing MLE on the expected complete data log-likelihood defined as $\mathbb{E}[\log p(\boldsymbol{y}, \boldsymbol{w}|\mathbf{X}, \lambda, \beta)]$, where the expectation is w.r.t. $\boldsymbol{w}$ which is treated here as the latent variable.

Assuming $p(y_n|\boldsymbol{w}, \boldsymbol{x}_n) = \mathcal{N}(y|\boldsymbol{w}^\top \boldsymbol{x}_n, \beta^{-1})$ and $p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}|\mathbf{0}, \lambda^{-1}\mathbf{I})$, sketch the EM algorithm, clearly specifying the E and M steps. In particular,

- What would be estimated in the E step (has to be a distribution)? Give the complete expression for that.
- Derive the M step updates for λ and β (note: some calculations may require you to make use of properties of matrix trace). Clearly specify the required expectations you will need to compute for these updates and give their expressions.