Probabilistic Machine Learning
(CS772A, Spring 2024)
Practice Set 1

Problem 1

(Simple exercise on computing MLE, MAP, and Posterior) Consider N count-valued observations $\{x_1, x_2, \ldots, x_N\}$ drawn i.i.d. from a Poisson distribution $p(x|\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}$ where λ is the rate parameter of the Poisson. Assume a gamma prior on λ, i.e., $p(\lambda) = \text{Gamma}(\lambda; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda}$, where $\alpha > 0$ is the *shape parameter* and $\beta > 0$ is the *rate parameter*, respectively, of the gamma.[1]. Note that, for this parameterization of gamma distribution, the prior's *mode* is $\frac{\alpha-1}{\beta}$ and mean is $\frac{\alpha}{\beta}$.

- Derive the MLE and MAP estimates for λ.
- Derive the expression for the full posterior distribution for λ.
- Show that the MAP estimate (i.e., mode of the posterior) can be written as weighted combination of the MLE estimate and the prior's mode. Likewise, show that the posterior's *mean* can be written as a weighted combination of the MLE estimate and the prior's *mean*.

Problem 2

(Distribution of Empirical Mean of Gaussian Observations) Consider N scalar-valued observations $x_1, \ldots, x_N$ drawn i.i.d. from $\mathcal{N}(\mu, \sigma^2)$. Consider their empirical mean $\bar{x} = \frac{1}{N}\sum_{n=1}^{N} x_n$. Representing the empirical mean as a linear transformation of a random variable, derive the probability distribution of $\bar{x}$. You may refer to the results on linear transformation of random variables from the prob-stats refresher slides under lecture 1 readings. Briefly explain why the result makes intuitive sense.

Problem 3

(MLE as KL Minimization) Suppose you are given N observations $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N\}$ from some true underlying data distribution $p_{data}(\boldsymbol{x})$ (may assume N to be very large, e.g., infinity). To learn it, you assume a parametrized distribution $p(\boldsymbol{x}|\theta)$ and estimate the parameters θ using MLE. Show that doing MLE is equivalent to finding θ that minimizes the KL divergence between the true distribution $p_{data}(\boldsymbol{x})$ and the assumed distribution $p(\boldsymbol{x}|\theta)$. Note that KL divergence between two probability distributions p and q is asymmetric and can be defined in two different ways: $KL(p||q)$ or $KL(q||p)$. For this problem, minimizing only one of these two will be equivalent to MLE. Why not the other one?

Problem 4

(Prior Hierarchy) Consider a model m with parameters θ and hyperparameters λ. Assume priors $p(\theta|\lambda, m)$, $p(\lambda|m)$ and $p(m)$. Assume we have observed some data $\mathbf{X}$ and the likelihood is of the form $p(\mathbf{X}|\theta, \lambda, m)$. For this problem setup, write down the expressions for computing: (1) $p(\theta|\mathbf{X}, \lambda, m)$, (2) $p(\lambda|\mathbf{X}, m)$, and (3) $p(m|\mathbf{X})$. Also rank these three quantities in terms of the difficulty of computing them (easiest to hardest) and briefly justify your ranking. Note: Your answers should not assume any conjugacy. Also, all quantities in each of these 3 expression should clearly and explicitly show everything you need to condition on.

[1] There is an alternate parameterization of gamma in terms of shape α and scale θ, for which $p(\lambda) \propto \lambda^{\alpha-1} e^{-\frac{\lambda}{\theta}}$

Problem 5

(It Gets Better..) Recall that, for a Bayesian linear regression model with likelihood $p(y|\boldsymbol{x}, \boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}^\top \boldsymbol{x}, \beta^{-1})$ and prior $p(\boldsymbol{w}) = \mathcal{N}or(0, \lambda^{-1}\mathbf{I})$, the *predictive posterior* is $p(y_*|\boldsymbol{x}_*) = \mathcal{N}(\mu_N^\top \boldsymbol{x}_*, \beta^{-1} + \boldsymbol{x}_*^\top \Sigma_N \boldsymbol{x}_*) = \mathcal{N}(\mu_N^\top \boldsymbol{x}_*, \sigma_N^2(\boldsymbol{x}_*))$, where we have defined $\sigma_N^2(\boldsymbol{x}_*) = \beta^{-1} + \boldsymbol{x}_*^\top \Sigma_N \boldsymbol{x}_*$ and μ_N and Σ_N are the mean and covariance matrix of the Gaussian posterior on $\boldsymbol{w}$, s.t., $\mu_N = \Sigma(\beta \sum_{n=1}^N y_n \boldsymbol{x}_n)$ and $\Sigma_N = (\beta \sum_{n=1}^N \boldsymbol{x}_n \boldsymbol{x}_n^\top + \lambda \mathbf{I})^{-1}$. Here, we have used the subscript N to denote that the model is learned using N training examples. As the training set size N increases, what happens to the variance of the predictive posterior? Does it increase or decrease or remain the same? You must also prove your answer formally. You might find the following identity useful: You may make use the following matrix identity:

$$(\mathbf{M} + \boldsymbol{v}\boldsymbol{v}^\top)^{-1} = \mathbf{M}^{-1} - \frac{(\mathbf{M}^{-1}\boldsymbol{v})(\boldsymbol{v}^\top \mathbf{M}^{-1})}{1 + \boldsymbol{v}^\top \mathbf{M}^{-1} \boldsymbol{v}}$$

Where $\mathbf{M}$ denotes a square matrix and $\boldsymbol{v}$ denotes a column vector.

Problem 6

(When You Integrate Out..) Suppose x is a scalar random variable drawn from a univariate Gaussian $p(x|\eta) = \mathcal{N}(x|0, \eta)$. The variance η itself is drawn from an exponential distribution: $p(\eta|\gamma) = \text{Exp}(\eta|\gamma^2/2)$, where $\gamma > 0$. Note that the exponential distribution is defined as $\text{Exp}(x|\lambda) = \lambda \exp(-\lambda x)$. Derive the expression of the marginal distribution of x, i.e., $p(x|\gamma) = \int p(x|\eta)p(\eta|\gamma)d\eta$ after integrating out η. What does the marginal distribution $p(x|\gamma)$ mean?

Plot both $p(x|\eta)$ and $p(x|\gamma)$ and include in the writeup PDF itself. What difference do you see between the shapes of these two distributions? **Note:** You don't need to submit the code used to generate the plots. Just the plots (appropriately labeled) are fine.

Hint: You will notice that $\int p(x|\eta)p(\eta|\gamma)d\eta$ is a hard to compute integral. However, the solution does have a closed form expression. One way to get the result is to compute the **moment generating function (MGF)**[2] of $\int p(x|\eta)p(\eta|\gamma)d\eta$ (note that this is a p.d.f.) and compare the obtained MGF expression with the MGFs of various p.d.f.s given in the table on the following Wikipedia page: `https://en.wikipedia.org/wiki/Moment-generating_function`, and identify which p.d.f.'s MGF it matches with. That will give you the form of distribution $p(x|\gamma)$. Specifically, name this distribution and identify its parameters.

Problem 7

(Hierarchical Modeling) Suppose we have student data from M schools where N_m denotes the number of students in school m. The data for each school $m = 1, \ldots, M$ is in the following form: For student n in school m, there is a response variable (e.g., score in some exam) $y_n^{(m)} \in \mathbb{R}$ and a feature vector $\boldsymbol{x}_n^{(m)} \in \mathbb{R}^D$.

Assume a linear regression model for these scores, i.e., $p(y_n^{(m)}|\boldsymbol{x}_n^{(m)}, \boldsymbol{w}_m) = \mathcal{N}(y_n^{(m)}|\boldsymbol{w}_m^\top \boldsymbol{x}_n^{(m)}, \beta^{-1})$, where $\boldsymbol{w}_m \in \mathbb{R}^D$ denotes the regression weight vector for school m, and β is known. Note that this can also be denoted as $p(\boldsymbol{y}^{(m)}|\mathbf{X}^{(m)}, \boldsymbol{w}_m) = \mathcal{N}(\boldsymbol{y}^{(m)}|\mathbf{X}^{(m)}\boldsymbol{w}_m, \beta^{-1}\mathbf{I}_{N_m})$, where $\boldsymbol{y}^{(m)}$ is $N_m \times 1$ and $\mathbf{X}^{(m)}$ is $N_m \times D$. Assume a prior $p(\boldsymbol{w}_m) = \mathcal{N}(\boldsymbol{w}_m|\boldsymbol{w}_0, \lambda^{-1}\mathbf{I}_D)$, λ to be known and $\boldsymbol{w}_0$ to be unknown.

Derive the expression for the **log** of the MLE-II objective for estimating $\boldsymbol{w}_0$. **You do not need to optimize this objective w.r.t.** $\boldsymbol{w}_0$; just writing down the final expression of objective function is fine. Also state what is the benefit of this approach as opposed to fixing $\boldsymbol{w}_0$ to some value, if our goal is to learn the school-specific weight vectors $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_M$? (Feel free to make direct use of properties of Gaussian distributions; you may refer to the provided results in the prob-stats refresher slides, or in books).

[2] MGF of a p.d.f. $p(x)$ is defined as $M_X(t) = \int_{-\infty}^{\infty} e^{tx} p(x) dx$

Problem 8

(Peeking into the neighborhood) Consider a regression model where the joint distribution of any input $\boldsymbol{x} \in \mathbb{R}^D$ and its output $y \in \mathbb{R}$ is $p(\boldsymbol{x}, y) = \frac{1}{N}\sum_{n=1}^{N} f(\boldsymbol{x} - \boldsymbol{x}_n, y - y_n)$ where $\{(\boldsymbol{x}_n, y_n)\}_{n=1}^{N}$ denotes the training examples. Further assume $f(\boldsymbol{x} - \boldsymbol{x}_n, y - y_n) = \mathcal{N}([\boldsymbol{x} - \boldsymbol{x}_n, y - y_n]^\top | \mathbf{0}, \sigma^2 \mathbf{I}_{D+1})$. For this model, derive the conditional distribution of the output y given the input, i.e., $p(y|\boldsymbol{x})$, as well as the expectation $\mathbb{E}[y|\boldsymbol{x}]$. Also give a brief justification as to why the expressions $p(y|\boldsymbol{x})$ and $\mathbb{E}[y|\boldsymbol{x}]$ make intuitive sense.

Problem 9 (30 marks)

(Spike-and-Slab Model for Sparsity) Suppose w is a real-valued r.v. that can either be close to zero with probability π, or take a wide range of real values with probability $(1 - \pi)$. An example of this could be in a regression problem where w is the weight of some feature. The feature could be irrelevant for predicting the output (in which case we would expect w to be close to zero) or be useful (in which case we would expect w to be non-zero with a wide range of possible values). We want to infer w from data taking a Bayesian approach. Note that, in practice, $\boldsymbol{w}$ is a vector (with each entry modeled this way) but here we will consider the scalar w case.

A popular approach to solve such problems is to impose a *spike and slab prior* on w. Let $b \in \{0, 1\}$ be a binary random variable and define the following *conditional* prior on w:

$$p(w|b, \sigma^2_{\text{spike}}, \sigma^2_{\text{slab}}) = \begin{cases} \mathcal{N}(w|0, \sigma^2_{\text{spike}}) & b = 0 \\ \mathcal{N}(w|0, \sigma^2_{\text{slab}}) & b = 1, \end{cases}$$

Depending on the value of b (which itself is unknown), w is assumed drawn from one of the two distributions: a "peaky" one $\mathcal{N}(w|0, \sigma^2_{\text{spike}})$ with variance σ^2_{spike} being very small, and a "flat" one $\mathcal{N}(w|0, \sigma^2_{\text{slab}})$, with $\sigma_{\text{slab}} \gg \sigma_{\text{spike}}$. So, basically, the value of the binary "mask" b decides whether the feature is relevant or not.

We usually don't know b, so we must either infer it with w, or marginalize it if we care about the value of w.

- Assume a prior $p(b = 1) = \pi = 1/2$, which means both Gaussians are equally likely for w. What is the *marginal* prior $p(w|\sigma^2_{\text{spike}}, \sigma^2_{\text{slab}})$, i.e., the prior over w after integrating out b?
- Plot this marginal prior distribution for $(\sigma^2_{\text{spike}}, \sigma^2_{\text{slab}}) = (1, 100)$. Briefly comment on how the shape of this distribution compares with that of a typical Gaussian distribution?
- Suppose someone gave us a "noisy" version of w defined as $x = w + \epsilon$ where $\epsilon \sim \mathcal{N}(\epsilon|0, \rho^2)$. This is equivalent to writing $p(x|w, \rho^2) = \mathcal{N}(x|w, \rho^2)$. Assume the variance ρ^2 to be known. Given x, what is the posterior distribution of b, $p(b = 1|x, \sigma^2_{\text{spike}}, \sigma^2_{\text{slab}}, \rho^2)$? Note that w must NOT appear in this expression (has to be integrated out first). Plot the resulting posterior $p(b = 1|x, \sigma^2_{\text{spike}}, \sigma^2_{\text{slab}}, \rho^2)$ as a function of x.
- Given the noisy observation $x = w + \epsilon$ as defined above, what is the posterior distribution of w, i.e., $p(w|x, \sigma^2_{\text{spike}}, \sigma^2_{\text{slab}}, \rho^2)$? Note that b must NOT appear in this expression (has to be integrated out or summed over since b is discrete).
- Assume $(\sigma^2_{\text{spike}}, \sigma^2_{\text{slab}}) = (1, 100)$, the noise variance $\rho^2 = 0.01$. For these settings of the hyperparameters, plot the posterior distribution of w given a noisy observation $x = 3$.