Probabilistic Machine Learning
(CS772A, Fall 2022)
Homework 1
Due Date: August 31, 2022 (11:59pm)

Instructions:

- Only electronic submissions will be accepted. Your main PDF writeup must be typeset in LaTeX (please also refer to the "Additional Instructions" below).
- Your submission will have two parts: The main PDF writeup (to be submitted via Gradescope `https://www.gradescope.com/`) and the code for the programming part (to be submitted via this Dropbox link: `https://tinyurl.com/5n6zracf`). Both parts must be submitted by the deadline to receive full credit (**delay in submitting either part would incur late penalty for both parts**). We will be accepting late submissions upto 72 hours after the deadline (with every 24 hours delay incurring a 10% late penalty, applied on per-hour delay basis). We won't be able to accept submissions after that.
- We have created your Gradescope account (you should have received the notification). Please use your IITK CC ID (not any other email ID) to login. Use the "Forgot Password" option to set your password.

Additional Instructions

- We have provided a LaTeX template file `hw1sol.tex` to help typeset your PDF writeup. There is also a style file `pmi.sty` that contain shortcuts to many of the useful LaTeX commends for doing things such as boldfaced/calligraphic fonts for letters, various mathematical/greek symbols, etc., and others. Use of these shortcuts is recommended (but not necessary).
- Your answer to every question should begin on a new page. The provided template is designed to do this automatically. However, if it fails to do so, use the `\clearpage` option in LaTeX before starting the answer to a new question, to *enforce* this.
- While submitting your assignment on the Gradescope website, you will have to specify on which page(s) is question 1 answered, on which page(s) is question 2 answered etc. To do this properly, first ensure that the answer to each question starts on a different page.
- Be careful to flush all your floats (figures, tables) corresponding to question n before starting the answer to question $n + 1$ otherwise, while grading, we might miss your important parts of your answers.
- Your solutions must appear in proper order in the PDF file i.e. solution to question n must be complete in the PDF file (including all plots, tables, proofs etc) before you present a solution to question $n + 1$.
- For the programming part, all the code and README should be zipped together and submitted as a single file named `yourrollnumber.zip`. Please DO NOT submit the data provided.

Problem 1 (15 marks)

(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 2 (10 marks)

(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 $\bar{x}$. Briefly explain why the result makes intuitive sense.

Problem 3 (20 marks)

(Benefits of Probabilistic Joint Modeling) Consider a dataset of test-scores of students from M schools in a district: $\boldsymbol{x} = \{\boldsymbol{x}^{(m)}\}_{m=1}^{M} = \{x_1^m, \ldots, x_{N_m}^{(m)}\}_{m=1}^{M}$, where N_m denotes the number of students in school m. Assume the scores of students in school m are drawn independently as $x_n^{(m)} \sim \mathcal{N}(\mu_m, \sigma^2)$ where the Gaussian's mean μ_m is unknown and the variance σ^2 is same for all schools and known (for simplicity). Assume the means $\mu_1, \ldots, \mu_M$ of the M Gaussians to also be Gaussian distributed $\mu_m \sim \mathcal{N}(\mu_0, \sigma_0^2)$ where μ_0 and σ_0^2 are hyperparameters.

1. Assume the hyperparameters μ_0 and σ_0^2 to be known. Derive the posterior distribution of μ_m and write down the mean and variance of this posterior distribution. **Note:** While you can derive it the usual way, the derivation will be much more compact if you use the result of Problem 2 and think of each school's data as a *single* observation (the empirical mean of observations) having the distribution derived in Problem 3.

2. Assume the hyperparameter μ_0 to be unknown (but still keep σ_0^2 as fixed for simplicity). Derive the marginal likelihood $p(\boldsymbol{x}|\mu_0, \sigma^2, \sigma_0^2)$ and use MLE-II (i.e., maximization of the marginal likelihood) to estimate μ_0 (note again that σ^2 and σ_0^2 are known here). Note: Looking at the form/expression of the marginal likelihood, if the MLE-II result looks obvious to you, you may skip the derivation and directly write the result.

3. Consider using this MLE-II estimate of μ_0 from part (2) in the posteriors of each μ_m you derived in part (1). Do you see any benefit in using the MLE-II estimate of μ_0 as opposed to using a known value of μ_0?

Problem 4 (25 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 $x=3$.

Do not submit the code for this part. All of the answers/derivations for this part (including the plots) should be in the PDF writeup.

Problem 5 (30 marks): Programming Assignment

(Bayesian Linear Regression) Consider a toy data set consisting of 10 training examples $\{x_n,y_n\}_{n=1}^{10}$ with each input x_n as well as the output y_n being scalars. The data is given below.

$$\begin{aligned}\boldsymbol{x} &= [-2.23,-1.30,-0.42,0.30,0.33,0.52,0.87,1.80,2.74,3.62];\\ \boldsymbol{y} &= [1.01,0.69,-0.66,-1.34,-1.75,-0.98,0.25,1.57,1.65,1.51]\end{aligned}$$

We would like to learn a Bayesian linear regression model using this data, assuming a Gaussian likelihood model for the outputs with fixed noise precision $\beta=4$. However, instead of working with the original scalar-valued inputs, we will map each input x using a degree-k polynomial as $\phi_k(x)=[1,x,x^2,\ldots,x^k]^\top$. Note that, when using the mapping ϕ_k, each original input becomes $k+1$ dimensional. Denote the entire set of mapped inputs as $\phi_k(\boldsymbol{x})$, a $10\times(k+1)$ matrix. Consider $k=1,2,3$ and 4, and learn a Bayesian linear regression model for each case. Assume the following prior on the regression weights: $p(\boldsymbol{w})=\mathcal{N}(\boldsymbol{w}|\mathbf{0},\mathbf{I})$ with $\boldsymbol{w}\in\mathbb{R}^{k+1}$.

1. For each k, compute the posterior of $\boldsymbol{w}$ and show a plot with 10 random functions drawn from the inferred posterior (show the functions for the input range $x\in[-4,4]$). Also show the original training examples on the same plot to illustrate how well the functions fit the training data.

2. For each k, compute and plot the **mean** of the posterior predictive $p(y_*|\phi_k(x_*), \phi_k(\boldsymbol{x}), \boldsymbol{y}, \beta)$ on the interval $x_* \in [-4, 4]$. On the same plot, also show the predictive posterior mean plus-and-minus two times the predictive posterior standard deviation.

3. Compute the log marginal likelihood $\log p(\boldsymbol{y} \mid \phi_k(\boldsymbol{x}), \beta)$ of the training data for each of the 4 mappings $k = 1, 2, 3, 4$. Which of these 4 "models" seems to explain the data the best?

4. Using the MAP estimate $\boldsymbol{w}_{MAP}$, Compute the log likelihood $\log p(\boldsymbol{y}|\boldsymbol{w}_{MAP}, \phi_k(\boldsymbol{x}), \beta)$ for each k. Which of these 4 models seems to have the highest log likelihood? Is your answer the same as that based on the log marginal likelihood (part 3)? Which of these two criteria (highest log likelihood or highest log marginal likelihood) do you think is more reasonable to select the best model and why?

5. For your best model, suppose you could include an additional training input x' (along with its output y') to "improve" your learned model using this additional example. Where in the region $x \in [-4, 4]$ would you like the chosen x' to be? Explain your answer briefly,

Your implementation should be in Python notebook (and should not use an existing implementation of Bayesian linear regression from any library).

Submit the plots as well as the code in a single zip file (named `yourrollnumber.zip`).