CSED515 Machine Learning - Assignment 2

Dongwoo Kim

Due date: 23:59pm, April 1st, 2023

Remark

Assignment Submission. All students must submit their homework via PLMS. Submit your answer on PLMS in a single PDF file named with `your_student_id.pdf`. You can scan your hand-written answers or write your answer with a tablet (If you are ambitious enough, try to use LaTeX to write your answers).

Late Homework Policy. We do not allow late submission. If you have any question regarding this, please send an email to the lecturer.

Honor Code. We strongly encourage students to form study groups. Students may discuss and work on homework problems in groups. However, each student must write down their solutions independently, i.e., each student must understand the solution well enough in order to reconstruct it by him/herself. Using code or solutions obtained from the web (GitHub/Google/ etc.) is considered an honor code violation. We check all the submissions for plagiarism. We take the honor code very seriously and expect students to do the same.

Importance Notice Note that the first assignment aims to identify your background on linear algebra, probability and statistics, and vector calculus. Show your work with your answers. You will get zero score if there is no proper explanation. You can use any external resources to solve the problems, however, please make sure that you have understood the details of what you have written. If you think you are not ready to solve these questions, please consider to take preliminary classes before taking this class. The list of preliminary courses are provided during the second lecture.

1. **Maximum Likelihood Estimator [15pt]**

Let $x \in \{0, 1\}$ denote the result of a coin toss ($x = 0$ for tails, $x = 1$ for heads). The coin is potentially biased, so that heads occurs with probability θ_1. Suppose that someone else observes the coin flip and reports to you the outcome, y. But this person is unreliable and only reports the result correctly with probability θ_2; i.e., $p(y \mid x, \theta_2)$ is given by

	$y = 0$	$y = 1$
$x = 0$	θ_2	$1 - \theta_2$
$x = 1$	$1 - \theta_2$	θ_2

Assume that θ_2 is independent of x and θ_1.

(a) Write down the joint probability distribution $p(x, y \mid \boldsymbol{\theta})$ as a 2×2 table, in terms of $\boldsymbol{\theta} = (\theta_1, \theta_2)$.

(b) Suppose have the following dataset: $\boldsymbol{x} = (1, 1, 0, 1, 1, 0, 0), \boldsymbol{y} = (1, 0, 0, 0, 1, 0, 1)$. What are the MLEs for θ_1 and θ_2 ? Justify your answer. Hint: note that the likelihood function factorizes, $p(x, y \mid \boldsymbol{\theta}) = p(y \mid x, \theta_2)\, p(x \mid \theta_1)$ What is $p(\mathcal{D} \mid \hat{\boldsymbol{\theta}}, M_2)$ where M_2 denotes this 2-parameter model? (You may leave your answer in fractional form if you wish.)

(c) Given the same dataset used in the previous question, now consider a model with 4 parameters, $\boldsymbol{\theta} = (\theta_{0,0}, \theta_{0,1}, \theta_{1,0}, \theta_{1,1})$, representing $p(x, y \mid \boldsymbol{\theta}) = \theta_{x,y}$. (Only 3 of these parameters are free to vary, since they must sum to one.) What is the MLE of $\boldsymbol{\theta}$? What is $p(\mathcal{D} \mid \hat{\boldsymbol{\theta}}, M_4)$ where M_4 denotes this 4-parameter model?

2. **Biased Estimator [15pt]**

During the lecture, I said MLE of variance is biased. Show that $\hat{\sigma}^2_{\text{MLE}} = \frac{1}{N}\sum_{n=1}^{N}(x_n - \hat{\mu})^2$ is a biased estimator of σ^2, i.e., show

$$\mathbf{E}_{X_1,\ldots,X_n\sim\mathcal{N}(\mu,\sigma)}\left[\hat{\sigma}^2(X_1,\ldots,X_n)\right] \neq \sigma^2$$

where $\hat{\mu}$ is the empirical mean of the samples.

Hint: note that $X_1,\ldots,X_N$ are independent, and use the fact that the expectation of a product of independent random variables is the product of the expectations.

3. **MLE for the Poisson Distribution [15pt]**

(a) The Poisson pmf is defined as $\text{Poi}(x \mid \lambda) = e^{-\lambda}\frac{\lambda^x}{x!}$, for $x \in \{0, 1, 2, \ldots\}$ where $\lambda > 0$ is the rate parameter. Derive the MLE of λ given observations $x_1, ..., x_n$ where $x_i \sim \text{Poi}(\lambda)$.

(b) Now we perform a conjugate Bayesian analysis using Gamma distribution. Let $D = \{x_1, x_2, ..., x_n\}$ be a set of observations where $x_i \sim \text{Poi}(\lambda)$. Derive the posterior $p(\lambda \mid D)$ assuming a conjugate prior $p(\lambda) = \text{Ga}(\lambda \mid a, b) \propto \lambda^{a-1}e^{-\lambda b}$.

(c) What does the posterior mean tend to as $a \to 0$ and $b \to 0$? (Recall that the mean of a $\text{Ga}(a, b)$ distribution is a/b.)

4. **Bayesian Analysis of Exponential Distribution [25pt]**

A lifetime X of a machine is modeled by an exponential distribution with unknown parameter θ. The likelihood is $p(x \mid \theta) = \theta e^{-\theta x}$ for $x \geq 0, \theta > 0$.

(a) Derive the MLE of θ given observations $\mathcal{D} = \{x_1, x_2, ..., x_N\}$.

(b) Suppose we observe $X_1 = 5, X_2 = 6, X_3 = 4$ (the lifetimes (in years) of 3 different iid machines). What is the MLE given this data?

(c) Assume that an expert believes θ should have a prior distribution that is also exponential

$$p(\theta) = \text{Expon}(\theta \mid \lambda)$$

Choose the prior parameter, call it $\hat{\lambda}$, such that $\mathbb{E}[\theta] = 1/3$. Hint: recall that the Gamma distribution has the form

$$\text{Ga}(\theta \mid a, b) \propto \theta^{a-1} e^{-\theta b}$$

and its mean is a/b.

(d) Following the previous question, what is the posterior, $p(\theta \mid \mathcal{D}, \hat{\lambda})$?

(e) Is the exponential prior conjugate to the exponential likelihood?

5. **Linear Regression [20pt]**

When we have multiple independent outputs $\boldsymbol{y} \in \mathbb{R}^M$ given input $\boldsymbol{x}$ in linear regression, the model becomes

$$p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{W}) = \prod_{j=1}^{M} \mathcal{N}(y_j|\mathbf{w}_j^\top \boldsymbol{x}, \sigma_j^2),$$

where $\sigma_j \in \mathbb{R}$. Note that we assume an independent Gaussian noise for each output dimension j.

In this question, we apply this result to a model with 2 dimensional output vector $\boldsymbol{y}_i \in \mathbb{R}^2$. Suppose we have some binary input data, $x_i \in \{0, 1\}$. The training data is as follows:

x	y
0	$(-1, -1)^T$
0	$(-1, -2)^T$
0	$(-2, -1)^T$
1	$(1, 1)^T$
1	$(1, 2)^T$
1	$(2, 1)^T$

Let us embed each x_i into $2d$ using the following basis function:

$$\phi(0) = (1, 0)^T, \quad \phi(1) = (0, 1)^T$$

The model becomes

$$\hat{\boldsymbol{y}} = \mathbf{W}^T \boldsymbol{\phi}(x)$$

where $\mathbf{W}$ is a 2×2 matrix. Derive and compute the MLE for $\mathbf{W}$ from the above data.

6. **Offset Term for Linear Regression [10pt]**

In linear regression, it is common to include a column of 1's in the design matrix, so we can solve for the offset term w_0 term and the other parameters $\mathbf{w}$ at the same time. However, it is also possible to solve for $\mathbf{w}$ and w_0 separately. Let's define the linear regression with the offset term explicitly, i.e., $\mathbb{E}[y \mid \mathbf{x}] = w_0 + \mathbf{w}^T\mathbf{x}$. Derive the MLE of w_0.