


Homework #No. 1 (Probability, Bayesian Linear Regression)

For questions, please refer to Moodle.
Released on 25/09/2023

GENERAL INSTRUCTIONS

- Submission of solutions is not mandatory but solving the exercises are highly recommended. The master solution will be released after the exercise deadline.
- Part of the exercises are available on Moodle as a quiz. These problems are marked with .

Exercise 0: Warm-Up

Four points are chosen uniformly at random on the surface of a sphere. What is the probability that the center of the sphere lies inside the tetrahedron whose vertices are at the four points?

Exercise 1: Bayes Rule

Imagine you have installed a sophisticated alarm system in your home. In case of a thief, it reacts with a 100% probability and sends you a notification. However, an alarm can also be triggered by an earthquake in 10% of the cases. Let's introduce the following random variables: Alarm (A) to represent whether the alarm at your home is triggered (1 for activated, 0 for not activated); Theft (T) to indicate whether theft is occurring (1 for theft, 0 for no theft); Earthquake (E) to represent whether an earthquake is happening (1 for earthquake, 0 for no earthquake). For your neighborhood, the probability of theft is $p(T = 1) = 2 \times 10^{-4}$ and of earthquake is $p(T = 1) = 10^{-2}$. Conditional probabilities of alarm $p(A = 1|T, E)$ can be summarised by

	E=0	E=1
T=0	0	0.1
T=1	1	1

So the full probabilistic model is $p(A, T, E) = P(A|T, E)P(T)P(E)$

- You are busy at university and receive a notification from the alarm system. How would you react? Compute the probability of $p(T = 1|A = 1)$.
- How much would the $p(T = 1|A = 1)$ be changed if prior probability $p(T = 1)$ is 10 times higher in your neighborhood?
- Your trusty radio (variable R) sometimes broadcasts information about earthquakes in the area (1 for radio reporting, 0 for no radio reporting). Let's assume $P(R = 1|E = 1) = 0.5$, while $P(R = 1|E = 0) = 0$. Assume the following joint probability factorization $p(A, T, E, R) = P(A|T, E)P(R|E)P(T)P(E)$. How would your posterior estimate of the probability of theft change if you heard a radio broadcast about an earthquake together with receiving a notification? Compute the probability of $p(T = 1|A = 1, R = 1)$.

Exercise 2: True/False Question

- Let $X \sim \mathcal{N}(\mu, \Sigma)$ be a Gaussian random vector.

- ☒ Any affine transformation of X , such as $Y = M^\top X + b$, is also Gaussian. ☐ True ☐ False
- ☒ All non-affine transformations of X are *not* Gaussian. ☐ True ☐ False

(b) Let X, Y, Z be independent standard normal random variables.

- ☒ The random variable $\frac{X+YZ}{\sqrt{1+Z^2}}$ is Gaussian. ☐ True ☐ False
- ☒ Let $\alpha \sim \text{Unif}(0, 1)$ be independent of X and Y . Then $X \cos(\alpha) + Y \sin(\alpha) \sim \mathcal{N}(0, 1)$. ☐ True ☐ False

(c) Let $X \sim \mathcal{N}(0, 1)$ and Z be ± 1 with equal probability and independent of X .

- ☒ The random variable $Y = ZX$ is standard Gaussian. ☐ True ☐ False
- ☒ The vector $(X, Y) \in \mathbb{R}^2$ is a Gaussian random vector. ☐ True ☐ False

Exercise 3: Bayesian Inference

A simple example of Bayesian inference is mean estimation. Assume we have a set of random variables (our data) X_1, \dots, X_N , each of which is distributed according to $\mathcal{N}(\mu, \sigma^2)$. Suppose that the variance σ^2 is a known constant available to us, and our goal is to estimate μ . To do things in a Bayesian way, we start off with a prior $p(\mu)$, which we assume for convenience to be Gaussian $\mathcal{N}(\mu_0, \sigma_0^2)$. Each of μ_0 and σ_0^2 can be endowed with a prior probability distribution again and this process can go on (this way of modelling things is called Hierarchical Bayesian Modeling). However, in our case, suppose that we have an idea about μ_0 and σ_0^2 and treat them as constants (this approach is called Empirical Bayes). Hence, the posterior distribution of μ can be written as

$$p(\mu \mid \{X_1, \dots, X_N\}, \sigma^2, \mu_0, \sigma_0^2).$$

- (a) Write down the posterior density in terms of $\mu, X_1, \dots, X_N, \sigma^2, \mu_0, \sigma_0^2$.
- (b) Prove that this density is still a Gaussian. Compute the mean and variance of this distribution.
- (c) Compare the MAP and MLE for this problem. Can you explain what is the role of the prior? Specifically, can you observe the dependence on μ_0 and σ_0^2 and how they effect the MAP?

Exercise 4: Multivariate Gaussian Distribution

A vector-valued random variable $x \in \mathbb{R}^d$ is said to have a multivariate normal distribution with mean $\mu \in \mathbb{R}^d$ and covariance matrix $\Sigma \in \mathbf{S}_{++}^d$ if its pdf is:

$$p(x; \mu, \Sigma) = \frac{1}{(2\pi)^{n/2} \det(\Sigma)^{1/2}} \exp \left(-\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right)$$

Prove the following facts, stated in the first lecture:

- (a) Every marginal of a Gaussian vector is Gaussian (slide [49]),
- (b) Conditioning on a subset of variables of a joint Gaussian is Gaussian (slide [50]).

Recall the following fact about characteristic functions, which will help you to solve the next questions:
For a random vector $X \in \mathbb{R}^d$, define its characteristic function φ_X as

$$\varphi_X(t) = \mathbb{E}[\exp(i t^\top X)], \quad \text{for all } t \in \mathbb{R}^d.$$

The characteristic function completely identifies a distribution. For a multivariate Normal distribution $\mathcal{N}(\mu, \Sigma)$, its characteristic function can be computed explicitly:

$$\varphi(t) = \exp(i t^\top \mu - \frac{1}{2} t^\top \Sigma t).$$

- (c) [📌] Let $X = (X_1, \dots, X_d)$ be a d -dimensional standard Gaussian random vector, that is, $X \sim \mathcal{N}_d(0, I)$. Define $Y = AX + \mu$, where A is a $d \times d$ matrix and $\mu \in \mathbb{R}^d$. What is the distribution of Y ?
1. $\mathcal{N}(\mu, A)$ 2. $\mathcal{N}(\mu, A^\top A)$ 3. $\mathcal{N}(\mu, A^2)$ 4. $\mathcal{N}(\mu, AA^\top)$
- (d) [📌] If B is an $r \times d$ matrix, what is the distribution of BY ?
1. $\mathcal{N}(B\mu, BAA^\top B^\top)$ 2. $\mathcal{N}(B\mu, BAA^\top)$ 3. $\mathcal{N}(\mu, BAA^\top B^\top)$ 4. $\mathcal{N}(\mu, BAA^\top)$
- (e) [📌] Let $X = (X_1, X_2)$ be a bivariate Normal random variable with mean $\mu = (1, 1)$ and covariance matrix $\Sigma = \begin{pmatrix} 3 & 1 \\ 1 & 2 \end{pmatrix}$. What is the mean and the variance of the conditional distribution of $Y = X_1 + X_2$ given $Z = X_1 - X_2 = 0$?

Exercise 5: Online Bayesian Linear Regression

As you have seen in the lecture, BLR with Gaussian prior and Gaussian likelihood has a closed-form posterior, which is also Gaussian and its mean and covariance matrix can be written in terms of the observed data and the noise level σ_n . Concretely, if X is the data matrix and y is the vector of responses, the posterior has the form

$$p(w \mid X, y) = \mathcal{N}(w; (X^\top X + \sigma_n^2 I)^{-1} X^\top y, (\sigma_n^{-2} X^\top X + I)^{-1})$$

Instead of assuming that the whole data is *available offline* (as a data matrix X), we are now in a situation where the data is *coming one by one in an i.i.d. stream*. That is, at round $t \in \mathbb{N}$, a new datapoint (x_t, y_t) is observed, where $x_t \in \mathbb{R}^d$ is the feature vector and $y_t \in \mathbb{R}$ is the response.

Our goal is to keep track of the posterior of w at the end of each round. That is, letting $p_t(w) := p(w \mid \{(x_i, y_i)\}_{i=1}^t)$ to be the posterior of w after round t , we want to be able to compute p_t and the predictive distribution for a desired data point x^* .

To achieve this task, one idea is to construct the data matrix $X^{(t)} \in \mathbb{R}^{t \times d}$ and response vector $y^{(t)} \in \mathbb{R}^t$ and do the computations described in the lecture *every round*. This idea gives the correct posterior; however, it is computationally prohibitive: as larger and larger amounts of data is collected, the data matrix grows larger, and the computation complexity, as well as memory requirement grows with t .

- (a) Can you design an algorithm that *updates* the posterior (as opposed to recalculating it from scratch) in a smarter way? The requirement is that the memory should not grow as $O(t)$. As a hint, write down p_{t+1} recursively as a function of p_t and try to do the computations in a smart way.
- (b) If d is large, computing the inverse every round is very expensive. Can you use the recursive structure you found in the previous question to bring down the computational complexity of every round to $O(d^2)$?