In the Name of God

Introduction to Machine Learning (25737-2)

Problem Set 1

Spring Semester 1403-04

Department of Electrical Engineering

Sharif University of Technology

Instructor: Dr. S. Amini

Due on Esfand 25, 1403 at 23:59

(∗) starred problems are optional!

1 Multivariate Gaussian Distribution

In this problem, We do some execises related to Multivariate Gaussian Distribution to get a sense about it and also see why it is important.

Let X be a random variable taking values in $\mathbb{R}^n$. It is normally distributed with mean μ and covariance matrix Σ. Recall that the probability density function (pdf) $p_X(x)$, sometimes denoted $p(X = x)$, for X is given by

$$p_X(x) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right)$$

1.1 Normalization constant

Show how to obtain the normalization constant $\frac{1}{\sqrt{(2\pi)^n|\Sigma|}}$ for the multivariate Gaussian, starting from the fact that

$$p_X(x) \propto \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right).$$

Hints: It's fine to assume $\mu = 0$ (Why?). A useful (and cool!) fact is that

$$\int_{\mathbb{R}} \exp\left(-\frac{1}{2}x^2\right) dx = \sqrt{2\pi}.$$

1.2 A few special cases

1. Let the random variable $Y = 2X$. What is the pdf of Y?
2. What can we say about the distribution of X if Σ is the identity matrix, I? Does this imply anything about factorization of the pdf?

1.3 Disturbution Shape

In all parts of this problem assume $\mu = [0 \;\; 0]^T$ if not mentioned.

1. What can we say about the distribution of X if Σ is

$$\begin{bmatrix} 10 & 0 \\ 0 & 1 \end{bmatrix}$$

Approximately what shape do equiprobability contours (i.e., sets $\{x \in \mathbb{R}^n : p_X(x) = c\}$ for some c) have? What happens if $\mu = [1 \;\; 4]^T$

2. What can we say about the distribution of X if Σ is

$$\begin{bmatrix} 10 & -4 \\ -4 & 10 \end{bmatrix}$$

Approximately what shape do equiprobability contours of this distribution have?

3. Is

$$\begin{bmatrix} 2 & 10 \\ 10 & 2 \end{bmatrix}$$

a valid Σ? How can you tell?

2 The different types of machine learning problems.

Determine whether the tasks described below involve supervised learning or unsupervised learning. For supervised learning problems, identify them as regression, classification, or probabilistic classification.

1. Predict the risk of an accident at an intersection, given features such as the time of day and weather.
2. Identify cars, bicyclists, and pedestrians in video taken by an autonomous vehicle's cameras.
3. Determine the probability that there is a stop sign in an image.
4. Generate new road scenarios (generate streets, place stop signs and intersections) for testing autonomous vehicles in a simulation.

3 Posterior

3.1 Dirichlet

Assume that A is a categorical random variable such that $A \sim \text{Cat}(\theta)$. Now, assume we have N i.i.d. samples of A, denoted as $D = \{A_i\}_{i=1}^N$.

1. Calculate the likelihood of D in terms of θ, i.e., $P(D \mid \theta)$.
2. Assume that θ has a Dirichlet prior distribution, $\theta \sim \text{Dir}(\alpha)$, where $\alpha_i > 1$ for all i. Show that the posterior distribution of θ after observing D is also a Dirichlet distribution, $\text{Dir}(\alpha')$, and derive the updated parameters α'.

3.2 Gamma

Assume we have a random variable X drawn from a Poisson distribution, $X \sim \text{Po}(\lambda)$, where λ has a prior Gamma distribution with parameters α and β, i.e., $\lambda \sim \text{Gamma}(\alpha, \beta)$. Given N i.i.d. samples from X, denoted as $D = \{x_i\}_{i=1}^N$:

1. Show that the Gamma distribution is a conjugate prior for the Poisson distribution by deriving the posterior distribution of λ given the data D.

4 Gaussian System

We have a weight and height measurement system that measures height in centimeters (cm) and weight in kilograms (kg). The system first scales down the quantities to measure them with smaller tools. This conversion introduces a bias term. Additionally, if a person is heavy, it compresses the scale under his feet, making him appear shorter. The system also introduces additive white Gaussian noise to each quantity independently. The measured quantities are as follows:

$$h_m = h \cdot 0.5 + 1 - 0.01 \cdot w + n_h,$$

$$w_m = w \cdot 0.3 + 0.1 + 0.001 \cdot h + n_w,$$

Where: - w_m is the measured weight, - h_m is the measured height, - $n_h \sim \mathcal{N}(0, 0.02)$ is the noise for height, - $n_w \sim \mathcal{N}(0, 0.01)$ is the noise for weight(n_w and n_h are independent), and - w and h are the real values of weight and height, respectively.

Given $w_m = 18$ and $h_m = 80$:

1. Estimate w and h from their distributions.
2. Assume that human weight and height have a multivariate Gaussian distribution as follows:

$$\begin{bmatrix} h \\ w \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 180 \\ 80 \end{bmatrix}, \begin{bmatrix} 8 & 3 \\ 3 & 4 \end{bmatrix} \right).$$

   Based on this prior, estimate w and h from their distributions, and compare the results with the previous part.
3. Now assume $n_h = 0$ and $n_w = 0$, and repeat the previous part. Compare the results.
4. Now assume the measurements are fixed ($h = h_0, w = w_0$), and the noise variances are equal, i.e., $Var(n_h) = Var(n_w) = \sigma^2$. Show that the estimated height $\hat{h}$ can be expressed as a function of σ^2, i.e.,

$$\hat{h} = f(\sigma^2).$$

5. Determine the sign of the derivative $\frac{df}{d(\sigma^2)}$ in the interval $[0, 200]$. (You can do it either by deriving the equation or by inspection. Both approaches are accepted.)

Hint: Model the problem as a Linear Gaussian System and use the relations discussed in the slides.

Note: In all the parts, you should first calculate the distributions and then estimate the quantities.

5 Hitler

You are responsible for encrypting Nazi information during World War II. The number of messages you receive each day for encryption can be modeled as the random variable $N \sim \text{Poi}(\lambda)$.

Your encryption machine has become somewhat outdated, and each message is independently encrypted incorrectly with probability p. Let X be the random variable representing the number of correctly encrypted messages in a day. Let Y be the random variable representing the number of incorrectly encrypted messages in a day.

1. Find the probability mass function (PMF) of X.
2. Determine the joint probability mass function of the two random variables X and Y.
3. Are X and Y independent?

6 Probability and Gaussian Variables

6.1 Chernoff's Inequality

Show that for any $s > 0$, we can state the following inequality:

$$P(X > t) \leq e^{-st}\mathbb{E}[e^{sX}]$$

6.2 Gaussian Random Variable

Show that if X is a Gaussian random variable, the following holds:

$$P(|X| \geq t) \leq 2e^{-\frac{t^2}{2\sigma^2}}$$

6.3 Expectation of Maximum of a Random Variable

Show that for any random variable Z, we have:

$$\mathbb{E}[\max(0, Z)] = \int_0^\infty P(Z \geq x)dx.$$

6.4 Another Inequality

The random variable X follows the distribution P_X. The probability of the event ε occurring is given by $\mathbb{P}[\varepsilon] = \delta$. Prove that:

$$\mathbb{E}\left[f(X)|\varepsilon\right] \leq \mathbb{E}[f(X)] + \sqrt{\frac{\text{Var}(f(X))}{\delta}}$$

where f is a function mapping the random variable X to real numbers.

7 Dice and Dice

First, we roll a fair dice until we obtain a number X. Then, we roll the dice multiple times until we obtain a number Y such that $X \leq Y$ (we continue rolling the dice until we obtain a number that is greater than X).

1. Find the probability that $Y = 6$ given that $X = 2$.
2. Find the probability that $Y = 6$.
3. Find the probability that $X = 2$ given that $Y = 6$.

8 Bayesian Decision Theory

You have recently started working as a doctor in a world affected by the COVID-19 pandemic. It is known that approximately ten percent of the population is infected with the disease, and the average lifespan of the population is about 80 years.

The diagnostic test has a sensitivity of 0.875 and a specificity of 0.975. Additionally, the discovered treatment for this disease reduces the average lifespan of individuals by 8 years; in other words, its QALY (Quality-Adjusted Life Year) is equal to 8. (If you are not familiar with the concept of reduced lifespan due to drug consumption, research QALY (Quality-Adjusted Life Year).)

8.1 Cure?

Four individuals with the following conditions visit you:

1. A 20-year-old individual who tested positive.
2. A 70-year-old individual who tested positive.
3. A 20-year-old individual who tested negative.
4. A 70-year-old individual who tested negative.

For each of these individuals, decide whether you should prescribe the treatment or not.

8.2 Threshold

Find an age threshold for prescribing the treatment.

8.3 New treatment

If a new treatment with a QALY of 0.5 is introduced to the market, how would your decisions change?

9 * A Surprising Statistical Reversal

Researchers compare two treatments for kidney stones, Treatment A (open surgery) and Treatment B (percutaneous nephrolithotomy). They record success rates overall and also split by stone size. The data are summarized in the table below:

	Overall	**Small Stones**	**Large Stones**
Treatment A	78% (273/350)	**93%** (81/87)	**73%** (192/263)
Treatment B	**83%** (289/350)	87% (234/270)	69% (55/80)

Table 1: Kidney Stone Treatment Success Rates

Observations:

- For both small and large stones, Treatment A is more successful than Treatment B.
- However, when the data are aggregated (ignoring stone size), Treatment B exhibits a higher overall success rate than Treatment A.

Questions:

1. Why does Treatment A appear superior in both subgroups (small and large stones) yet is inferior overall?
2. What confounding factor(s) in this data might cause such a paradoxical reversal of the apparent "best" treatment when moving from subgroup analysis to the combined data?
3. How would you advise a medical practitioner or policy maker to interpret these results correctly?

Supplementary Notes

Dirichlet Distribution

The Dirichlet distribution is a vector distribution for a random vector $x \in \mathbb{R}^n$ such that:

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \Rightarrow \begin{cases} 0 \leq x_i \leq 1, & i = 1, \ldots, n \\ \sum_{i=1}^n x_i = 1 \end{cases}$$

The probability density function of the Dirichlet distribution has a vector parameter $\alpha \in (\mathbb{R}^+)^n$. The mean and variance of each component of the random vector x are given by:

$$\mathrm{Dir}(x|\alpha) = \frac{1}{B(\alpha)} \prod_{i=1}^n x_i^{\alpha_i - 1}, \quad B(\alpha) = \frac{\prod_{i=1}^n \Gamma(\alpha_i)}{\Gamma\left(\sum_{i=1}^n \alpha_i\right)}$$

$$\begin{cases} E[X_k] = \frac{\alpha_k}{\alpha_0} \\ \mathrm{Var}[X_k] = \frac{\alpha_k(\alpha_0 - \alpha_k)}{\alpha_0^2(\alpha_0+1)}, & \alpha_0 = \sum_{i=1}^n \alpha_i \\ \mathrm{Mode}[X_k] = \frac{\alpha_k - 1}{\alpha_0 - n} \end{cases}$$

Bernoulli Distribution:

$$\mathrm{Ber}(y|\theta) = \theta^y (1-\theta)^{1-y}, \quad \begin{cases} y \in \{0, 1\} \\ \theta \in [0, 1] \end{cases}$$

Categorical Distribution:

$$\mathrm{Cat}(y|\theta) = \prod_{c=1}^C \theta_c^{y_c}, \quad \begin{cases} 0 \leq \theta_c \leq 1 \\ \sum_{c=1}^C \theta_c = 1 \end{cases}$$

Gamma Distribution

The Gamma distribution is a continuous probability distribution for a non-negative random variable $x \in \mathbb{R}^+$. It is characterized by two parameters: the shape parameter $k > 0$ and the scale parameter $\theta > 0$.

The probability density function (PDF) of the Gamma distribution is given by:

$$\mathrm{Gamma}(x|k, \theta) = \frac{1}{\Gamma(k)\theta^k} x^{k-1} e^{-\frac{x}{\theta}}, \quad x \geq 0$$

where $\Gamma(k)$ is the Gamma function. The mean, variance, and mode of the Gamma distribution are as follows:

$$\begin{cases} E[X] = k\theta \\ \mathrm{Var}[X] = k\theta^2 \\ \mathrm{Mode}[X] = (k-1)\theta, & \text{for } k \geq 1 \end{cases}$$