Problem Set 2 - Machine Learning Theories

April 16, 2025

Problem 1

Consider two multi-dimensional random vectors $\mathbf{x} \in \mathbb{R}^M$ and $\mathbf{y} \in \mathbb{R}^D$ following marginal and conditional distributions which are Gaussian:

$$p(\mathbf{x}) = N(\mathbf{x}; \mu, \Sigma)$$
$$p(\mathbf{y}|\mathbf{x}) = N(\mathbf{y}; A\mathbf{x} + \mathbf{b}, S),$$

where $\mu \in \mathbb{R}^M$, $\Sigma \in \mathbb{R}^{M \times M}$, $A \in \mathbb{R}^{D \times M}$, $\mathbf{b} \in \mathbb{R}^D$, and $S \in \mathbb{R}^{D \times D}$.

(1) Determine the joint distribution $p(\mathbf{x}, \mathbf{y})$.

(2) Determine the marginal distribution $p(\mathbf{y})$.

(3) Consider two random vectors $\mathbf{x}, \mathbf{z} \in \mathbb{R}^D$ following Gaussian distributions $p(\mathbf{x}) = N(\mathbf{x}; \mu_x, \Sigma_x)$ and $p(\mathbf{z}) = N(\mathbf{z}; \mu_z, \Sigma_z)$, respectively. Use the above results to find the marginal distribution of $\mathbf{y} = \mathbf{x} + \mathbf{z}$, by considering the marginal distribution $p(\mathbf{x})$ and the conditional distribution $p(\mathbf{y}|\mathbf{x})$.

Use the following equations if necessary:

(a)

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} = \begin{pmatrix} M & -MBD^{-1} \\ -D^{-1}CM & D^{-1} + D^{-1}CMBD^{-1} \end{pmatrix},$$

where $M = (A - BD^{-1}C)^{-1}$.

(b) When the random vector $\mathbf{x} = \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{pmatrix}$ follows a Gaussian with mean $\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$ and covariance $\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$,

$$\mathbb{E}[\mathbf{x}_1|\mathbf{x}_2] = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(\mathbf{x}_2 - \mu_2), \quad \mathrm{Cov}[\mathbf{x}_1|\mathbf{x}_2] = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$$
$$\mathbb{E}[\mathbf{x}_1] = \mu_1, \quad \mathrm{Cov}[\mathbf{x}_1] = \Sigma_{11}.$$

(1)

Problem 2

Consider a Gaussian likelihood function for one-dimensional data $X = \{x_i\}_{i=1}^N$, where each $x_i \in \mathbb{R}$, and the parameters are the mean μ and precision τ. The likelihood is given by:

$$p(X|\mu, \tau) = \prod_{i=1}^{N} \left(\frac{\tau}{2\pi}\right)^{1/2} \exp\left(-\frac{\tau}{2}(x_i - \mu)^2\right).$$

(1) When μ is fixed, the conjugate prior for τ is the Gamma distribution, which has the form

$$Gam(\tau|a, b) = \frac{1}{\Gamma(a)} b^a \tau^{a-1} \exp(-b\tau),$$

where $\Gamma(a) = \int_0^\infty x^{a-1} e^{-x} dx$ is the Gamma function.
Using this conjugate prior, derive the posterior distribution

$$p(\tau|X) = Gam(\tau|a_N, b_N),$$

by computing the updated parameters a_N and b_N.

(2) The conjugate prior for joint parameters (μ, τ) is the normal-Gamma distribution, which has the form

$$\begin{aligned} p(\mu, \tau) &= p(\mu|\tau) \cdot p(\tau) \\ &= N(\mu|\mu_0, (\beta\tau)^{-1}) \cdot Gam(\tau|a, b) \\ &= \sqrt{\frac{\beta\tau}{2\pi}} \exp\left(-\frac{\beta\tau}{2}(\mu - \mu_0)^2\right) \cdot \frac{1}{\Gamma(a)} b^a \tau^{a-1} \exp(-b\tau). \end{aligned}$$

Using this conjugate prior, derive the posterior distribution

$$p(\mu, \tau|X) = N(\mu|\mu_N, (\beta_N\tau)^{-1}) \cdot Gam(\tau|a_N, b_N),$$

by computing the updated parameters μ_N, β_N, a_N, b_N.

(Hint: Begin by writing down the joint posterior up to a normalization constant:

$$p(\mu, \tau|X) \propto p(X|\mu, \tau)p(\mu, \tau).$$

First, make a perfect square form (i.e., complete the square) in μ within the exponential to identify μ_N and β_N. Then, using the remaining terms, extract the exponential and power terms of τ to identify a_N and b_N.)

Problem 3

Consider modeling a sequence of coin flips. Let $\theta \in [0, 1]$ be the probability of the coin landing heads. The Bernoulli distribution can serve as the likelihood for a single flip: $p(x|\theta) = \theta^x(1-\theta)^{1-x}$, where $x = 1$ for heads and $x = 0$ for tails.
Assume we use a Beta distribution as the prior distribution for θ:

$$p(\theta) = \mathrm{Beta}(\theta|\alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

where $\alpha, \beta > 0$ are hyperparameters.
Suppose we observe a dataset $\mathcal{D} = \{x_1, x_2, \ldots, x_N\}$ from N independent coin flips, where $m = \sum_{i=1}^{N} x_i$ is the number of heads, and $N - m$ is the number of tails. The likelihood function for the dataset is $p(\mathcal{D}|\theta) = \theta^m(1-\theta)^{N-m}$.

(1) Using Bayes' theorem, derive the posterior distribution $p(\theta|\mathcal{D})$. Show that the posterior distribution has the same functional form (a Beta distribution) as the prior distribution. Explain why this demonstrates that the Beta distribution is a conjugate prior for the Bernoulli likelihood.

(2) Express the parameters (the new α' and β') of the derived posterior distribution in terms of the prior parameters (α, β) and the data (N, m).

(3) If the prior distribution is $\mathrm{Beta}(\theta|2, 2)$ and we observe 7 heads in 10 coin flips ($N = 10, m = 7$), what is the posterior distribution $p(\theta|\mathcal{D})$?

Problem 4

Consider a Gaussian likelihood $p(\mathbf{x}|\mu) = N(\mathbf{x}; \mu, \Sigma)$ for $\mathbf{x} \in \mathbb{R}^D$, where $\mu \in \mathbb{R}^D$ is the mean parameter and $\Sigma \in \mathbb{R}^{D\times D}$ is a fixed covariance matrix. Assume further that the posterior distribution over the mean parameter μ is Gaussian: $p(\mu|X) = N(\mu; \mu_N, \Sigma_N)$, where $\mu_N \in \mathbb{R}^D$ and $\Sigma_N \in \mathbb{R}^{D\times D}$. Derive the predictive distribution $p(\mathbf{x}|X)$ by marginalizing over μ:

$$p(\mathbf{x}|X) = \int p(\mathbf{x}|\mu)p(\mu|X)d\mu.$$

Use the Woodbury matrix identity if needed:

$$(A + UCV)^{-1} = A^{-1} - A^{-1}U(C^{-1} + VA^{-1}U)^{-1}VA^{-1}.$$

Problem 5

Consider a simple 1D linear regression model $y = wx + \epsilon$, where w is the unknown weight (slope), and ϵ is Gaussian noise with zero mean and known variance σ^2, i.e., $\epsilon \sim \mathcal{N}(0, \sigma^2)$. The likelihood for a data point (x, y) is therefore:

$$p(y|x, w, \sigma^2) = \mathcal{N}(y|wx, \sigma^2)$$

We place a Gaussian prior distribution on the weight w, with zero mean and variance α^{-1} (where $\alpha > 0$ is the precision):

$$p(w|\alpha) = \mathcal{N}(w|0, \alpha^{-1}).$$

Assume we are given a dataset $\mathcal{D} = \{(\mathbf{x}, \mathbf{y})\} = \{(x_1, y_1), \ldots, (x_N, y_N)\}$ consisting of N data points.

(1) Using Bayes' rule, derive the expression for the log posterior of w, $\log p(w|\mathcal{D}, \sigma^2, \alpha)$, up to an additive constant (i.e., you may ignore terms that do not depend on w).

(2) By rearranging the terms in the log posterior into a quadratic form in w, show that the posterior distribution $p(w|\mathcal{D}, \sigma^2, \alpha)$ is a Gaussian distribution.

(3) Identify the mean m_N and precision α_N (or equivalently, the variance $\sigma_N^2 = \alpha_N^{-1}$) of the posterior distribution. Express your result in terms of the data $(\mathbf{x}, \mathbf{y})$, the noise variance (σ^2), and the prior precision (α). (Hint: Identify the coefficients of the quadratic form in w to identify the posterior parameters.)

(4) Given the dataset $(\mathbf{x}, \mathbf{y}) = \{(1, 2), (3, 4)\}$, noise variance $\sigma^2 = 1$, and prior precision $\alpha = 1$, calculate the mean m_N of the posterior distribution.

Problem 6

Consider some positive semi-definite kernel functions $k(\mathbf{x}, \mathbf{y})$ for $\mathbf{x}, \mathbf{y} \in \mathbb{R}^D$, and find the feature mapping $\phi : \mathbb{R}^D \to \mathbb{R}^F$ corresponding to the kernel function that satisfies

$$k(\mathbf{x}, \mathbf{y}) = \phi(\mathbf{x})^\top \phi(\mathbf{y}). \tag{1}$$

Represent $\phi(\mathbf{x}) \in \mathbb{R}^F$ in a vector form and find the dimensionality F of the feature space.

(1) $k(\mathbf{x}, \mathbf{y}) = (\mathbf{x}^\top \mathbf{y})^3$, where $\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \in \mathbb{R}^2$.

(2) $k(\mathbf{x}, \mathbf{y}) = 1 + \mathbf{x}^\top \mathbf{A} \mathbf{y}$, where $\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \in \mathbb{R}^2$, and $\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{pmatrix} \in \mathbb{R}^{2\times 2}$ is a symmetric positive-definite matrix. *(Hint: find $\mathbf{L} = \begin{pmatrix} a & 0 \\ b & c \end{pmatrix} \in \mathbb{R}^{2\times 2}$ that satisfies $\mathbf{A} = \mathbf{L}\mathbf{L}^\top$.)*

(3) $k(\mathbf{x}, \mathbf{y}) = k_1(\mathbf{x}, \mathbf{y}) + k_2(\mathbf{x}, \mathbf{y})$, where $k_1(\mathbf{x}, \mathbf{y}) = \psi(\mathbf{x})^\top \psi(\mathbf{y})$ with $\psi(\mathbf{x}) = [\psi_1(\mathbf{x}) \ldots \psi_{F_1}(\mathbf{x})]^\top \in \mathbb{R}^{F_1}$, and $k_2(\mathbf{x}, \mathbf{y}) = \xi(\mathbf{x})^\top \xi(\mathbf{y})$ with $\xi(\mathbf{x}) = [\xi_1(\mathbf{x}) \ldots \xi_{F_2}(\mathbf{x})]^\top \in \mathbb{R}^{F_2}$.

Problem 7
Consider the following linear dynamical system with Gaussian noise:

$$x_{t+1} = ax_t + \epsilon_t, \quad \epsilon_t \sim N(0, \sigma^2), \quad t = 0, 1, 2, 3, \ldots$$

where $0 < a < 1$, the initial state is distributed as $x_0 \sim N(1, \sigma_0^2)$, and the noise terms ϵ_t and $\epsilon_{t'}$ are independent for all $t \neq t'$. The collection of $(x_{t_1}, x_{t_2}, \ldots, x_{t_N})$ for distinct time indices $t_1, \ldots, t_N \in \mathbb{N} \cup \{0\}$ is jointly Gaussian hence this system defines a Gaussian process.
Find the mean function $m(t) = \mathbb{E}[x_t]$ and the covariance function $k(t_1, t_2) = \mathrm{Cov}(x_{t_1}, x_{t_2})$ of the Gaussian process $\{x_t\}_{t \geq 0}$.