Problem Set 1 - Machine Learning Theories

March 20, 2025

Problem 1

(1) We are given a function $f : \mathbb{R}^D \to \mathbb{R}$, $\mathbf{w} \mapsto f(\mathbf{w}) = (\mathbf{w}^\top \mathbf{X}\mathbf{w})^2$, where $\mathbf{w} \in \mathbb{R}^D$ and $\mathbf{X} \in \mathbb{R}^{D\times D}$. Find the derivative of $f(\mathbf{w})$ with respect to the vector $\mathbf{w}$. Use the definition of the vector derivative $\left[\frac{df}{d\mathbf{w}}\right]_k = \frac{\partial f}{\partial w_k}$, where w_k is the k-th element of $\mathbf{w}$.

(2) We are given a function $f : \mathbb{R}^{m\times n} \to \mathbb{R}$, $\mathbf{A} \mapsto f(\mathbf{A}) = \mathbf{x}^\top \mathbf{A}\mathbf{y}$, where $\mathbf{A} \in \mathbb{R}^{m\times n}$, $\mathbf{x} \in \mathbb{R}^m$, and $\mathbf{y} \in \mathbb{R}^n$. Find the derivative of $f(\mathbf{A})$ with respect to the matrix $\mathbf{A}$. Use the definition of the matrix derivative $\left[\frac{df}{d\mathbf{A}}\right]_{ij} = \frac{\partial f}{\partial A_{ij}}$, where A_{ij} is the (i, j) element of $\mathbf{A}$.

(3) We are given a function $f : \mathbb{R}^{D\times D} \to \mathbb{R}$, $\mathbf{B} \mapsto f(\mathbf{B}) = Tr(\mathbf{A}\mathbf{B})$, where $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{D\times D}$, $Tr(\cdot)$ is the trace operator for matrices defined as $Tr(\mathbf{X}) = \sum_{i=1}^{D} X_{ii}$, and X_{ii} is the (i, i) element of $\mathbf{X}$. Find the derivative of $f(\mathbf{B})$ with respect to the matrix $\mathbf{B}$.

Problem 2
Consider the following function,

$$f(\mathbf{x}; \mathbf{w}) = \frac{1}{1 + \exp\left(-\mathbf{w}^\top \mathbf{x}\right)}.$$

For a given dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$ with $\mathbf{x}_i \in \mathbb{R}^D$ and $y_i \in \{0, 1\}$, $i = 1, \ldots, N$, find the gradient descent update rule for $\mathbf{w}$ with learning rate η to minimize the following loss:

$$L = \frac{1}{2} \sum_{i=1}^{N} (f(\mathbf{x}_i; \mathbf{w}) - y_i)^2.$$

Problem 3

Evaluate the following integrals:

(1)

$$\int_0^\infty xe^{-x^2}\,dx$$

(2)

$$\int_{-\infty}^\infty e^{-x^2}\,dx$$

(Hint: Consider squaring the integral by multiplying $\int_{-\infty}^\infty e^{-y^2}dy$ and transforming the variables (x,y) to polar coordinates.)

(3)

$$\int_{-\infty}^\infty x\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2\sigma^2}(x-\mu)^2}dx$$

(4)

$$\int_{-\infty}^\infty x^2\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2\sigma^2}(x-\mu)^2}dx$$

(Hint: Differentiate both sides of the identity $\int_{-\infty}^\infty \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2\sigma^2}(x-\mu)^2}dx = 1$ with respect to σ^2.)

Problem 4

The probability density function of a D-dimensional Gaussian distribution is given by:

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{D}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^\top \Sigma^{-1}(\mathbf{x}-\mu)\right),$$

where $\mathbf{x} \in \mathbb{R}^D$ and $\mu \in \mathbb{R}^D$ and $\Sigma \in \mathbb{R}^{D\times D}$.

Now, suppose we partition the vectors and the covariance matrix as follows:

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{pmatrix}, \quad \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22}. \end{pmatrix}$$

Show that the probability density function $p(\mathbf{x})$ can be written as:

$$p(\mathbf{x}) \propto \exp\left(-\frac{1}{2}(\mathbf{x}_1-\mu_1')^\top {\Sigma_1'}^{-1}(\mathbf{x}_1-\mu_1') - \frac{1}{2}(\mathbf{x}_2-\mu_2)^\top \Sigma_{22}^{-1}(\mathbf{x}_2-\mu_2)\right).$$

Note that the inverse of Σ can be written as follows:

$$\Sigma^{-1} = \Lambda = \begin{pmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{pmatrix},$$

where

$$\Lambda_{11} = (\Sigma/\Sigma_{22})^{-1}, \quad \Lambda_{12} = -(\Sigma/\Sigma_{22})^{-1}\Sigma_{12}\Sigma_{22}^{-1}, \quad \Lambda_{21} = \Lambda_{12}^\top, \quad \Lambda_{22} = \Sigma_{22}^{-1}+\Sigma_{22}^{-1}\Sigma_{21}(\Sigma/\Sigma_{22})^{-1}\Sigma_{12}\Sigma_{22}^{-1},$$

and the Schur complement of Σ_{22} in Σ is $\Sigma/\Sigma_{22} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$.

(1) Expand the quadratic term $-\frac{1}{2}(\mathbf{x}-\mu)^\top\Sigma^{-1}(\mathbf{x}-\mu)$ in the exponent of $p(\mathbf{x})$ and express it as a perfect square in terms of $\mathbf{x}_1$ while treating $\mathbf{x}_2$ as a constant as follows:

$$-\frac{1}{2}(\mathbf{x}_1-\mu_1')^\top \Sigma_1'(\mathbf{x}_1-\mu_1') + \mathit{const}.$$

Then, derive explicit expressions for μ_1' and Σ_1' in terms of $\mu_1, \mu_2, \Sigma_{11}, \Sigma_{12}, \Sigma_{21}$, and Σ_{22}.

(2) Express the remaining terms (that is, those not involving $\mathbf{x}_1$) from part (1) as a perfect square in terms of $\mathbf{x}_2$. Specifically, show that these remaining terms take the form:

$$-\frac{1}{2}(\mathbf{x}_2-\mu_2)^\top \Sigma_{22}^{-1}(\mathbf{x}_2-\mu_2).$$

(3) Combining the results from (1) and (2) to conclude that $p(\mathbf{x})$ can be rewritten as:

$$p(\mathbf{x}) \propto \exp\left(-\frac{1}{2}(\mathbf{x}_1-\mu_1')^\top {\Sigma_1'}^{-1}(\mathbf{x}_1-\mu_1') - \frac{1}{2}(\mathbf{x}_2-\mu_2)^\top \Sigma_{22}^{-1}(\mathbf{x}_2-\mu_2)\right).$$

Problem 5

Suppose that $\mathbf{x} \in \mathbb{R}^D$ follows a Gaussian distribution with mean $\mu \in \mathbb{R}^D$ and covariance matrix $\Sigma \in \mathbb{R}^{D \times D}$, i.e., $\mathbf{x} \sim N(\mu, \Sigma)$. Show that the linearly transformed variable

$$\mathbf{z} = A\mathbf{x} + b$$

where $A \in \mathbb{R}^{D \times D}$ is an invertible matrix and $b \in \mathbb{R}^D$, also follows a Gaussian distribution. Derive its probability density function $p(\mathbf{z})$.

(Hint: Use the coordinate transformation rule of the probability density functions, i.e.,

$$p(\mathbf{z}) = p(\mathbf{x}) \left| \frac{\partial \mathbf{x}}{\partial \mathbf{z}} \right|,$$

where $\frac{\partial \mathbf{x}}{\partial \mathbf{z}} \in \mathbb{R}^{D \times D}$ is the Jacobian of $\mathbf{x}$ with respect to $\mathbf{z}$. Then plug in $\mathbf{x} = A^{-1}(\mathbf{z} - b)$.)

Problem 6
Derive the following results:

(1) Let $\mathbf{x} \sim N(\mu, \Sigma)$ be a Gaussian random variable with mean $\mu \in \mathbb{R}^D$ and covariance matrix $\Sigma \in \mathbb{R}^{D\times D}$. Show that $\mathbb{E}[\mathbf{x}\mathbf{x}^\top] = \mu\mu^\top + \Sigma$ using the properties $\mathbb{E}[\mathbf{x}] = \mu$ and $\mathbb{E}[(\mathbf{x}-\mu)(\mathbf{x}-\mu)^\top] = \Sigma$.

(2) Let $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N \in \mathbb{R}^D$ be independent and identically distributed (i.i.d.) samples from a Gaussian distribution $N(\mu, \Sigma)$. Show that

$$\mathbb{E}[\mathbf{x}_m\mathbf{x}_n^\top] = \mu\mu^\top + I_{mn}\Sigma,$$

where $\mathbb{E}[\cdot]$ denotes the expectation with respect to $\mathbf{x}_m, \mathbf{x}_n$ and I_{mn} is the (m, n) entry of the identity matrix $I \in \mathbb{R}^{N\times N}$.

(3) Using the above results, show that for the maximum likelihood estimators of the mean and covariance $\hat{\mu}_{ML} = \frac{1}{N}\sum_{i=1}^N \mathbf{x}_i$ and $\hat{\Sigma}_{ML} = \frac{1}{N}\sum_{i=1}^N (\mathbf{x}_i - \hat{\mu}_{ML})(\mathbf{x}_i - \hat{\mu}_{ML})^\top$, the expectations satisfy:

$$\mathbb{E}[\hat{\mu}_{ML}] = \mu,$$
$$\mathbb{E}[\hat{\Sigma}_{ML}] = \frac{N-1}{N}\Sigma,$$

where $\mathbb{E}[\cdot]$ denotes the expectation with respect to $\mathbf{x}_1, \ldots, \mathbf{x}_N$.

Problem 7

Given a dataset $\mathcal{D} = \{m_i\}_{i=1}^N$ where $m_i \in \{0, 1, \ldots, M\}$ with $M \in \mathbb{N}$, determine the Maximum Likelihood Estimator (MLE) for the parameter $\mu \in [0, 1]$ in the following probability model:

$$p(m_i; \mu) = \binom{M}{m_i} \mu^{m_i} (1 - \mu)^{M - m_i},$$

where $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ denotes the binomial coefficient.