CS5487 Problem Set 1

Probability Theory and Linear Algebra Review

Antoni Chan
Department of Computer Science
City University of Hong Kong

Probability Theory

Problem 1.1 Linear Transformation of a Random Variable

Let x be a random variable on $\mathbb{R}$, and $a, b \in \mathbb{R}$. Let $y = ax + b$ be the linear transformation of x. Show the following properties:

$$\mathbb{E}[y] = a\mathbb{E}[x] + b, \tag{1.1}$$

$$\mathrm{var}(y) = a^2 \mathrm{var}(x). \tag{1.2}$$

Now, let x be a vector r.v. on $\mathbb{R}^d$, and $A \in \mathbb{R}^{m \times d}$, $b \in \mathbb{R}^m$. Let $y = Ax + b$ be the linear transformation of x. Show the following properties:

$$\mathbb{E}[y] = A\mathbb{E}[x] + b, \tag{1.3}$$

$$\mathrm{cov}(y) = A\,\mathrm{cov}(x) A^T. \tag{1.4}$$

. . . .

Problem 1.2 Properties of Independence

Let x and y be statistically independent random variables ($x \perp y$). Show the following properties:

$$\mathbb{E}[xy] = \mathbb{E}[x]\mathbb{E}[y], \tag{1.5}$$

$$\mathrm{cov}(x, y) = 0. \tag{1.6}$$

. . . .

Problem 1.3 Uncorrelated vs Independence

Two random variables x and y are said to be *uncorrelated* if their covariance is 0, i.e., $\mathrm{cov}(x, y) = 0$. For statistically independent random variables, their covariance is always 0 (see Problem 1.2), and hence independent random variables are always uncorrelated. However, the converse is generally not true; uncorrelated random variables are not necessarily independent.

Consider the following example. Let the pair of random variables (x, y) take values $(1, 0)$, $(0, 1)$, $(-1, 0)$, and $(0, -1)$, each with equal probability (1/4).

(a) Show that $\mathrm{cov}(x, y) = 0$, and hence x and y are uncorrelated.

(b) Calculate the marginal distributions, $p(x)$ and $p(y)$. Show that the $p(x, y) \neq p(x)p(y)$ and thus x and y are not independent.

(c) Now consider a more general example. Assume that x and y satisfy

$$\mathbb{E}[x|y] = \mathbb{E}[x], \tag{1.7}$$

i.e., the mean of x is the same regardless of whether y is known or not (the above example satisfies this property). Show that x and y are uncorrelated.

....

Problem 1.4 Sum of Random Variables

Let x and y be random variables (possibly dependent), show the following property:

$$\mathbb{E}[x + y] = \mathbb{E}[x] + \mathbb{E}[y]. \tag{1.8}$$

Furthermore, if x and y are statistically independent ($x \perp y$), show that

$$\mathrm{var}(x + y) = \mathrm{var}(x) + \mathrm{var}(y). \tag{1.9}$$

However, in general this is not the case when x and y are dependent.

....

Problem 1.5 Expectation of an Indicator Variable

Let x be an indicator variable on $\{0, 1\}$. Show that

$$\mathbb{E}[x] = p(x = 1), \tag{1.10}$$

$$\mathrm{var}(x) = p(x = 0)p(x = 1). \tag{1.11}$$

....

Problem 1.6 Multivariate Gaussian

The multivariate Gaussian is a probability density over real vectors, $x = \begin{bmatrix} x_1 \\ \vdots \\ x_d \end{bmatrix} \in \mathbb{R}^d$, which is parameterized by a mean vector $\mu \in \mathbb{R}^d$ and a covariance matrix $\Sigma \in \mathbb{S}^d_+$ (i.e., a d-dimensional positive-definite symmetric matrix). The density function is

$$p(x) = \mathcal{N}(x|\mu, \Sigma) = \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} e^{-\frac{1}{2}\|x-\mu\|^2_\Sigma}, \tag{1.12}$$

where $|\Sigma|$ is the determinant of Σ, and

$$\|x - \mu\|^2_\Sigma = (x - \mu)^T \Sigma^{-1} (x - \mu) \tag{1.13}$$

is the *Mahalanobis distance*. In this problem, we will look at how different covariance matrices affect the shape of the density.

First, consider the case where Σ is a *diagonal matrix*, i.e., the off-diagonal entries are 0,

$$\Sigma = \begin{bmatrix} \sigma_1^2 & & 0 \\ 0 & \ddots & \\ & & \sigma_d^2 \end{bmatrix}. \tag{1.14}$$

(a) Show that with a diagonal covariance matrix, the multivariate Gaussian is equivalent to assuming that the elements of the vector are independent, and each is distributed as a univariate Gaussian, i.e.,

$$\mathcal{N}(x|\mu, \Sigma) = \prod_{i=1}^{d} \mathcal{N}(x_i|\mu_i, \sigma_i^2). \tag{1.15}$$

Hint: the following properties of diagonal matrices will be useful:

$$|\Sigma| = \prod_{i=1}^{d} \sigma_i^2, \quad \Sigma^{-1} = \begin{bmatrix} \frac{1}{\sigma_1^2} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{\sigma_d^2} \end{bmatrix}. \tag{1.16}$$

(b) Plot the Mahalanobis distance term and probability density function for a 2-dimensional Gaussian with $\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$, and $\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 0.25 \end{bmatrix}$. How is the shape of the density affected by the diagonal terms?

(c) Plot the Mahalanobis distance term and pdf when the variances of each dimension are the same, e.g., $\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$, and $\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$. This is sometimes called an i.i.d. (independently and identically distributed) covariance matrix, isotropic covariance matrix, or circular covariance matrix.

Next, we will consider the general case for the covariance matrix.

(d) Let $\{\lambda_i, v_i\}$ be the eigenvalue/eigenvector pairs of Σ, i.e.,

$$\Sigma v_i = \lambda_i v_i, \quad i \in \{1, \cdots, d\}. \tag{1.17}$$

Show that Σ can be written as

$$\Sigma = V \Lambda V^T, \tag{1.18}$$

where $V = [v_1, \cdots, v_d]$ is the matrix of eigenvectors, and $\Lambda = \text{diag}(\lambda_1, \cdots, \lambda_d)$ is a diagonal matrix of the eigenvalues.

(e) Let $y = V^T(x - \mu)$. Show that the Mahalanobis distance $\|x - \mu\|_\Sigma^2$ can be rewritten as $\|y\|_\Lambda^2$, i.e., a Mahalanobis distance with a diagonal covariance matrix. Hence, in the space of y, the multivariate Gaussian has a diagonal covariance matrix. (Hint: use Problem 1.12)

(f) Consider the transformation from y to x: $x = Vy + \mu$. What is the effect of V and μ?

(g) Plot the Mahalanobis distance term and probability density function for a 2-dimensional Gaussian with $\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$, and $\Sigma = \begin{bmatrix} 0.625 & 0.375 \\ 0.375 & 0.625 \end{bmatrix}$. How is the shape of the density affected by the eigenvectors and eigenvalues of Σ?

....

Problem 1.7 Product of Gaussian Distributions

Show that the product of two Gaussian distributions, $\mathcal{N}(x|\mu_1, \sigma_1^2)$ and $\mathcal{N}(x|\mu_2, \sigma_2^2,)$, is a scaled Gaussian,

$$\mathcal{N}(x|\mu_1, \sigma_1^2)\mathcal{N}(x|\mu_2, \sigma_2^2) = Z\mathcal{N}(x|\mu_3, \sigma_3^2), \tag{1.19}$$

where

$$\mu_3 = \sigma_3^2(\sigma_1^{-2}\mu_1 + \sigma_2^{-2}\mu_2), \tag{1.20}$$

$$\sigma_3^2 = \frac{1}{\sigma_1^{-2} + \sigma_2^{-2}}, \tag{1.21}$$

$$Z = \frac{1}{\sqrt{2\pi(\sigma_1^2 + \sigma_2^2)}} e^{\frac{-1}{2(\sigma_1^2+\sigma_2^2)}(\mu_1-\mu_2)^2} = \mathcal{N}(\mu_1|\mu_2, \sigma_1^2 + \sigma_2^2). \tag{1.22}$$

. . . .

Problem 1.8 Product of Multivariate Gaussian Distributions

Show that the product of two d-dimensional multivariate Gaussians distributions, $\mathcal{N}(x|a, A)$ and $\mathcal{N}(x|b, B)$, is a scaled multivariate Gaussian,

$$\mathcal{N}(x|a, A)\mathcal{N}(x|b, B) = Z\mathcal{N}(x|c, C), \tag{1.23}$$

where

$$c = C(A^{-1}a + B^{-1}b), \tag{1.24}$$

$$C = (A^{-1} + B^{-1})^{-1}, \tag{1.25}$$

$$Z = \frac{1}{(2\pi)^{\frac{d}{2}} |A + B|^{\frac{1}{2}}} e^{-\frac{1}{2}(a-b)^T(A+B)^{-1}(a-b)} = \mathcal{N}(a|b, A + B). \tag{1.26}$$

Hint: after expanding the exponent term, apply the result from Problem 1.10 and (1.35).

. . . .

Problem 1.9 Correlation between Gaussian Distributions

Using the result from Problem 1.8, show that the correlation between two multivariate Gaussian distributions is

$$\int \mathcal{N}(x|a, A)\mathcal{N}(x|b, B)dx = \mathcal{N}(a|b, A + B). \tag{1.27}$$

. . . .

Linear Algebra

Problem 1.10 Completing the square

Let $x, b \in \mathbb{R}^n$, $A \in \mathbb{S}^n$, $c \in \mathbb{R}$, and let $f(x)$ be a quadratic function of x,

$$f(x) = x^T A x - 2x^T b + c. \tag{1.28}$$

Show that $f(x)$ can be rewritten in the form

$$f(x) = (x-d)^T A (x-d) + e, \tag{1.29}$$

where

$$d = A^{-1} b, \tag{1.30}$$
$$e = c - d^T A d = c - b^T A^{-1} b. \tag{1.31}$$

Rewriting the quadratic function in (1.28) as (1.29) is a procedure known as "completing the square", which is very useful when dealing with products of Gaussian distributions.

....

Problem 1.11 Eigenvalues

Let $\{\lambda_i\}$ be the eigenvalues of $A \in \mathbb{R}^{n \times n}$. Derive the following properties:

$$\mathrm{tr}(A) = \sum_{i=1}^{n} \lambda_i, \qquad |A| = \prod_{i=1}^{n} \lambda_i. \tag{1.32}$$

....

Problem 1.12 Eigenvalues of an inverse matrix

Let $\{\lambda_i, x_i\}$ be the eigenvalues/eigenvectors of $A \in \mathbb{R}^{n \times n}$. Show that $\{\frac{1}{\lambda_i}, x_i\}$ are the eigenvalues/eigenvectors of A^{-1}.

....

Problem 1.13 Positive definiteness

Derive the following properties:

1. A symmetric matrix $A \in \mathbb{S}^n$ is positive definite if all its eigenvalues are greater than zero.
2. For any matrix $A \in \mathbb{R}^{m \times n}$, $G = A^T A$ is positive semidefinite.

....

Problem 1.14 Positive definiteness of inner product and outer product matrices

Let $X = [x_1, \cdots, x_n] \in \mathbb{R}^{d\times n}$ be a matrix of n column vectors $x_i \in \mathbb{R}^d$. We can think of each column vector x_i as a sample in our dataset X.

(a) outer-product: Prove that $\Sigma = XX^T$ is always *positive semi-definite* ($\Sigma \succeq 0$). When will Σ be strictly *positive definite*?

(b) inner-product: Prove that $G = X^T X$ is always *positive semi-definite*. When will G be strictly *positive definite*?

Note: If $\{x_1, \cdots, x_n\}$ are zero mean samples, then Σ is n times the sample covariance matrix. G is sometimes called a *Gram matrix* or *kernel matrix*.

. . . .

Problem 1.15 Useful Matrix Inverse Identities

Show that the following identities are true:

$$(P^{-1} + B^T R^{-1} B)^{-1} B^T R^{-1} = PB^T (BPB^T + R)^{-1}, \tag{1.33}$$

$$(A^{-1} + B^{-1})^{-1} = A(A+B)^{-1}B = B(A+B)^{-1}A, \tag{1.34}$$

$$(A^{-1} + B^{-1})^{-1} = A - A(A+B)^{-1}A = B - B(A+B)^{-1}B, \tag{1.35}$$

$$(A^{-1} + UC^{-1}V^T)^{-1} = A - AU(C + V^T AU)^{-1} V^T A \tag{1.36}$$

The last one is called the Matrix Inversion Lemma (or Sherman-Morrison-Woodbury formula) Hint: these can be verified by multiplying each side by an appropriate matrix term.

. . . .

Problem 1.16 Useful Matrix Determinant Identities

Verify that the following identities are true:

$$\left|I + AB^T\right| = \left|I + B^T A\right| \tag{1.37}$$

$$\left|I + ab^T\right| = 1 + b^T a \tag{1.38}$$

$$\left|A^{-1} + UV^T\right| = \left|I + V^T AU\right| \left|A^{-1}\right| \tag{1.39}$$

The last one is called the Matrix Determinant Lemma.

. . . .

Problem 1.17 Singular Value Decomposition (SVD)

The singular value decomposition (SVD) of a real $n\times m$ matrix A is a set of three matrices $\{U, S, V\}$, such that

$$A = USV^T, \tag{1.40}$$

where

- $U \in \mathbb{R}^{n\times m}$ is an orthonormal matrix of left-singular vectors (columns of U), i.e., $U^T U = I$.
- $S \in \mathbb{R}^{m\times m}$ is a diagonal matrix of *singular values*, i.e. $S = \text{diag}(s_1, \cdots, s_m)$. The singular values are usually ordered $s_1 \geq s_2 \geq \cdots \geq s_m \geq 0$
- $V \in \mathbb{R}^{m\times m}$ is an orthonormal matrix of right-singular vectors (columns of V), i.e., $V^T V = I$.

The SVD has an intuitive interpretation, which shows how the matrix A acts on a vector $x \in \mathbb{R}^m$. Consider the matrix-vector product

$$z = Ax = USV^T x. \tag{1.41}$$

This shows that the matrix A performs 3 operations on x, namely rotation by V^T, scaling along the axis via S, and another rotation by U. The SVD is also closely related to the eigen-decomposition, matrix inverse, and pseudoinverses.

(a) Show that the singular values of A are the square roots of the eigenvalues of the matrix $B = AA^T$, and that the left-singular vectors (columns of U) are the associated eigenvectors.

(b) Show that the singular values of A are the square roots of the eigenvalues of the matrix $C = A^T A$, and that the right-singular vectors (columns of V) are the associated eigenvectors.

(c) Suppose $A \in \mathbb{R}^{n\times n}$ is a square matrix of rank n. Show that the inverse of A can be calculated from the SVD,

$$A^{-1} = VS^{-1}U^T. \tag{1.42}$$

(d) Suppose $A \in \mathbb{R}^{n\times m}$ is a "fat" matrix ($n < m$) of rank n. The Moore-Penrose pseudoinverse of A is given by

$$A^\dagger = A^T(AA^T)^{-1}. \tag{1.43}$$

Likewise, for a "tall" matrix ($n > m$) of rank m, the pseudoinverse is

$$A^\dagger = (A^T A)^{-1} A^T. \tag{1.44}$$

Show that in both cases the pseudoinverse can also be calculated using the SVD,

$$A^\dagger = VS^{-1}U^T. \tag{1.45}$$

Note: Properties in (c) and (d) also apply for lower rank matrices. In these cases, some of the singular values will be 0, and S^{-1} is hence replaced with $S^\dagger$, where $S^\dagger$ replaces all non-zero diagonal elements by its reciprocal.

....