Contents

1 Introduction

1.1 Unsupervised Learning: Overview

The goal of unsupervised learning is to learn some "structure" in current data that generalizes to future data. To understand this better, let us review the concept of supervised learning that we are already familiar with.

Supervised Learning

The goal of supervised learning is to learn from current data to make predictions about future data. The current data consists of (input, output) pairs, denoted as $(x_1, y_1), \ldots, (x_n, y_n)$, which we often call the labeled data. Here, $x_i = (x_{i1}, \ldots, x_{ip}) \in \mathbb{R}^p$ is called the features (or feature vector) of the i-th data point, and $y_i \in \mathbb{R}$ is called the outcome or response; also called the label or class if it is categorical (e.g., $y_i \in \{0, 1\}$).[1] When y_i is categorical, we call the supervised learning problem a classification problem; otherwise, we call it a regression problem.

The goal is to find $f \colon \mathbb{R}^p \to \mathbb{R}$, which we call the prediction function, that performs well on future data, namely, $f(x_{n+1}) \approx y_{n+1}$ for future data (x_{n+1}, y_{n+1}) before seeing y_{n+1} so that $f(x_{n+1})$ serves as a good predictor of the unobserved output y_{n+1}. Here, $f(x_{n+1}) \approx y_{n+1}$, which means that the prediction function f learned from the data *generalizes* to new/unseen data, can be rigorously formulated in different ways depending on the application. Statistical learning theory concerns principles and guarantees for this generalization under probabilistic/statistical frameworks.

In supervised learning, we can validate the prediction function by evaluating its generalization performance. This is typically done by (1) splitting the data into training and testing sets, say $(x_1, y_1), \ldots, (x_m, y_m)$ for training and $(x_{m+1}, y_{m+1}), \ldots, (x_n, y_n)$ for testing, (2) learning f on the training set, and (3) evaluating its performance on the testing set based on the test error, e.g., mean squared error $\frac{1}{n-m} \sum_{i=m+1}^{n} (f(x_i) - y_i)^2$.

Unsupervised Learning

In unsupervised learning, we have unlabeled data, namely, only the features $x_1, \ldots, x_n \in \mathbb{R}^p$, which may or may not come with the corresponding outcomes or labels. Regardless, the focus is on discovering certain structures or patterns in the data, which can provide valuable insights and useful interpretations. We focus on the following unsupervised learning tasks.

- Dimension reduction: find a low-dimensional representation of the data.
- Clustering: group data points into clusters based on appropriate similarity.
- Mixture modeling: approximate the data by multi-modal distributions.
- Generative modeling: generate new data points similar to the observed data.

[1] Often, we consider $y_i \in \mathbb{R}^q$ when the output is multi-dimensional (e.g., multiple measurements or categories).

Dimension reduction and clustering are the most fundamental tasks, which have been integrated as fundamental building blocks in many data analysis pipelines. Mixture modeling aims to approximate the data by a mixture of distributions, which can be used for clustering and density estimation. Generative modeling aims to mimic the underlying data generating process to generate new data points, which has been increasingly important in recent years with the development of generative AI models.

As in supervised learning, we want to learn insights or knowledge from the current data that are *generalizable* to new data. However, unlike supervised learning, validation is infeasible in general because unsupervised learning concerns unlabeled data. Hence, there is no simple answer to how we measure whether the learned structure is "good" or "useful". This is often more of an art than science, which requires domain knowledge and human judgment to interpret the results.

Mathematical Preliminaries: Linear Algebra and Probability

Machine learning algorithms are often formulated as optimization problems, which can be solved by efficient numerical algorithms available in standard software packages (e.g., Python). Many such algorithms involve operations on the data matrix $X = (x_{ij}) \in \mathbb{R}^{n \times p}$, whose i-th row is the feature vector $x_i \in \mathbb{R}^p$ of the i-th data point. Particularly, most fundamental algorithms involve linear algebra operations, such as eigenvalue or singular value decomposition. Hence, it is crucial to have a solid understanding of linear algebra. In addition, basic knowledge of probability and elementary statistical inference is required for statistical learning theory to establish theoretical guarantees for machine learning algorithms.

1.2 Review of Linear Algebra

For a vector $u = (u_1, \ldots, u_p) \in \mathbb{R}^p$, we define the following norms.

- Euclidean norm (L2 norm): $\|u\|_2 := \sqrt{\langle u, u \rangle} = \sqrt{\sum_{j=1}^p u_j^2}$.
- Manhattan norm (L1 norm): $\|u\|_1 := \sum_{j=1}^p |u_j|$.
- Max norm (infinity norm): $\|u\|_\infty := \max_{j=1,\ldots,p} |u_j|$.

For two vectors $u = (u_1, \ldots, u_p) \in \mathbb{R}^p$ and $v = (v_1, \ldots, v_p) \in \mathbb{R}^p$, their inner product is defined as

$$\langle u, v \rangle = \sum_{j=1}^p u_j v_j.$$

We say that u and v are orthogonal if $\langle u, v \rangle = 0$. We say that a set of vectors $w_1, \ldots, w_k \in \mathbb{R}^p$ is orthonormal if $\langle w_i, w_j \rangle = 0$ for all $i \neq j$ and $\|w_i\|_2 = 1$ for all $i = 1, \ldots, k$.

We call a set of vectors $w_1, \ldots, w_k \in \mathbb{R}^p$ linearly independent if $\sum_{i=1}^k \alpha_i w_i = 0$ for some $\alpha_1, \ldots, \alpha_k \in \mathbb{R}$ implies that $\alpha_1 = \cdots = \alpha_k = 0$. For a set of vectors $w_1, \ldots, w_k \in \mathbb{R}^p$, we define the span of the vectors as $\text{span}(w_1, \ldots, w_k) := \{\sum_{i=1}^k \alpha_i w_i : \alpha_1, \ldots, \alpha_k \in \mathbb{R}\}$, which is a linear subspace

of $\mathbb{R}^p$. We call a set of vectors $w_1, \ldots, w_p \in \mathbb{R}^p$ a basis of $\mathbb{R}^p$ if they are linearly independent and $\text{span}(w_1, \ldots, w_p) = \mathbb{R}^p$; if they are orthonormal as well, we call them an orthonormal basis of $\mathbb{R}^p$.

For a matrix $A \in \mathbb{R}^{n \times m}$, the transpose of A is denoted by $A^\top \in \mathbb{R}^{m \times n}$; the column (resp. row) space of A is the span of the columns (resp. rows) of A, and the rank of A, denoted by $\text{rank}(A)$, is the dimension of the column space of A, which is equal to the dimension of the row space of A.

For a square matrix $A = (A_{ij}) \in \mathbb{R}^{n \times n}$, let $\text{tr}(A) := \sum_{i=1}^n A_{ii}$ denote the trace of A, and let $\det(A)$ denote the determinant of A; when $\det(A) \neq 0$, we say that A is invertible and denote its inverse by A^{-1}, namely, $A^{-1}A = AA^{-1} = I_n$, where I_n is the identity matrix of size $n \times n$. For two square matrices $A, B \in \mathbb{R}^{n \times n}$, we have $\text{tr}(AB) = \text{tr}(BA)$ and $\det(AB) = \det(A)\det(B)$.

For $a_1, \ldots, a_n \in \mathbb{R}$, let $\text{diag}(a_1, \ldots, a_n) \in \mathbb{R}^{n \times n}$ denote the diagonal matrix whose diagonal entries are $a_1, \ldots, a_n$. We call a matrix $Q \in \mathbb{R}^{n \times n}$ orthogonal (or orthonormal) if $Q^\top Q = QQ^\top = I_n$, i.e., $Q^{-1} = Q^\top$; this means that the columns (or rows) of Q are orthonormal.

Spectral Properties

For a square matrix $A \in \mathbb{R}^{n \times n}$, we say that a complex number $\lambda \in \mathbb{C}$ is an eigenvalue of A if there exists a nonzero vector $v \in \mathbb{C}^n$ such that $Av = \lambda v$; the vector v is called an eigenvector corresponding to the eigenvalue λ. If A is symmetric, i.e., $A = A^\top$, then all eigenvalues of A are real, and there exists an orthonormal basis of $\mathbb{R}^n$ consisting of eigenvectors of A, leading to the following eigenvalue decomposition of A:

$$A = V \Lambda V^\top,$$

where $V \in \mathbb{R}^{n \times n}$ is an orthogonal matrix whose columns are the eigenvectors of A, and $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_n) \in \mathbb{R}^{n \times n}$ is a diagonal matrix whose diagonal entries are the eigenvalues of A, where we typically order the eigenvalues in descending order $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$; in this case, we have $\text{tr}(A) = \sum_{i=1}^n \lambda_i$ and $\det(A) = \prod_{i=1}^n \lambda_i$, and the rank of A is equal to the number of nonzero eigenvalues. The multiplicity of an eigenvalue λ is the number of times it appears on the diagonal of Λ, which is the dimension of the eigenspace corresponding to λ; as noted, $n - r$ is the multiplicity of the eigenvalue 0.

For a square matrix $A \in \mathbb{R}^{n \times n}$, we say that A is positive semidefinite (PSD) if $v^\top Av \geq 0$ for all $v \in \mathbb{R}^n$. We say that A is positive definite (PD) if $v^\top Av > 0$ for all nonzero $v \in \mathbb{R}^n$. If A is symmetric, then A is PSD (resp. PD) if and only if all eigenvalues of A are nonnegative (resp. positive). To see this, let $A = V\Lambda V^\top$ be the eigenvalue decomposition of A; then, it suffices to observe that for any $v \in \mathbb{R}^n$, we have

$$v^\top Av = v^\top V \Lambda V^\top v = \sum_{i=1}^n \lambda_i w_i^2,$$

where $V^\top v = (w_1, \ldots, w_n)$. If A is symmetric and positive semidefinite, we can define the square root of A as $A^{1/2} := V\Lambda^{1/2}V^\top$, where $\Lambda^{1/2} := \text{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_n})$; then, we have $A^{1/2}$ is also symmetric and positive semidefinite, and $A^{1/2}A^{1/2} = A$.

Matrix Norms

Like vectors, we can define norms for matrices to measure their sizes in different ways. For a matrix $A \in \mathbb{R}^{n\times m}$, a natural way to define a norm is to treat A as a vector in $\mathbb{R}^{nm}$ by vectorizing A, i.e., stacking the columns of A into a single column vector in $\mathbb{R}^{nm}$; then, we can define the vector norms of A as the matrix norms of A, leading to the following matrix norms.

- Frobenius norm: $\|A\|_{\mathrm{F}} := \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{m} A_{ij}^2}$.
- L1 norm: $\|A\|_1 := \sum_{i=1}^{n}\sum_{j=1}^{m} |A_{ij}|$.
- Max norm: $\|A\|_\infty := \max_{\substack{i=1,\dots,n\\ j=1,\dots,m}} |A_{ij}|$.

The Frobenius norm comes from the Euclidean geometry, together with the matrix inner product: for $A = (A_{ij}) \in \mathbb{R}^{n\times m}$ and $B = (B_{ij}) \in \mathbb{R}^{n\times m}$, we define the inner product of A and B as

$$\langle A, B\rangle := \sum_{i=1}^{n}\sum_{j=1}^{m} A_{ij}B_{ij} = \mathrm{tr}(A^\top B) = \mathrm{tr}(AB^\top).$$

Then, $\|A\|_{\mathrm{F}} = \sqrt{\langle A, A\rangle}$.

Another way to define a matrix norm is to measure how much the linear transformation $v \mapsto Av$ stretches a vector $v \in \mathbb{R}^m$, leading to the following operator norm:

$$\|A\|_2 := \max_{v\in\mathbb{R}^m, v\neq 0} \frac{\|Av\|_2}{\|v\|_2},$$

which is also called the spectral norm of A. One can show that

$$\|A\|_2 = \sqrt{\lambda_{\max}(A^\top A)} = \sigma_{\max}(A),$$

where $\lambda_{\max}(A^\top A)$ is the largest eigenvalue of $A^\top A$, and $\sigma_{\max}(A)$ is the largest singular value of A. This can be shown by the singular value decomposition (SVD) of A, which we will discuss later when studying the principal component analysis (PCA).

By measuring the stretching in different vector norms, we can also define the following operator norms:

$$\|A\|_{1,1} := \max_{v\in\mathbb{R}^m, v\neq 0} \frac{\|Av\|_1}{\|v\|_1},$$
$$\|A\|_{\infty,\infty} := \max_{v\in\mathbb{R}^m, v\neq 0} \frac{\|Av\|_\infty}{\|v\|_\infty}.$$

One can show that $\|A\|_{1,1}$ is the maximum absolute column sum of A, and $\|A\|_{\infty,\infty}$ is the maximum absolute row sum of A, i.e.,

$$\|A\|_{1,1} = \max_{j=1,\dots,m} \sum_{i=1}^{n} |A_{ij}|,$$
$$\|A\|_{\infty,\infty} = \max_{i=1,\dots,n} \sum_{j=1}^{m} |A_{ij}|.$$

In the literature, $\|A\|_{1,1}$ is often denoted by $\|A\|_1$, and $\|A\|_{\infty,\infty}$ is often denoted by $\|A\|_\infty$, which should not be confused with the L1 and infinity norms defined earlier via vectorization.

1.3 Review of Probability

For a random variable X, its distribution P_X is a probability measure on the real line that maps each measurable set $A \subset \mathbb{R}$ to the probability $\mathbb{P}(X \in A)$, namely, $P_X(A) = \mathbb{P}(X \in A)$; its cumulative distribution function (CDF) is defined as $F(x) := \mathbb{P}(X \le x) = P_X((-\infty, x])$ for $x \in \mathbb{R}$. If X is a continuous random variable whose density function with respect to the Lebesgue measure is given by $f_X \colon \mathbb{R} \to [0, \infty)$, then we have $\mathbb{P}(X \in A) = \int_A f_X(x)\,\mathrm{d}x$ for any measurable set $A \subset \mathbb{R}$, and $F(x) = \int_{-\infty}^{x} f_X(z)\,\mathrm{d}z$ for $x \in \mathbb{R}$. If X is a discrete random variable taking values in a countable set $\mathcal{X} \subset \mathbb{R}$, then we have the probability mass function $f_X(x) := \mathbb{P}(X = x)$ for $x \in \mathcal{X}$, so that $\mathbb{P}(X \in A) = \sum_{x \in A \cap \mathcal{X}} f_X(x)$ for any measurable set $A \subset \mathbb{R}$.

The covariance and correlation between two random variables X and Y are defined as

$$\begin{aligned} \mathrm{Cov}(X, Y) &:= \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y], \\ \mathrm{Corr}(X, Y) &:= \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}. \end{aligned}$$

By the Cauchy-Schwarz inequality, we have $|\mathrm{Cov}(X, Y)| \le \sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}$ and $\mathrm{Corr}(X, Y) \in [-1, 1]$. We say that X and Y are uncorrelated if $\mathrm{Cov}(X, Y) = 0$. We say X and Y are independent if $\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B)$ for any measurable sets $A, B \subset \mathbb{R}$. If X and Y are independent, then $\mathrm{Cov}(X, Y) = 0$, but the converse is not true in general.

Multivariate Analysis

A random vector $X = (X_1, \ldots, X_p) \in \mathbb{R}^p$ is a collection of p random variables, whose distribution P_X is a probability measure on $\mathbb{R}^p$ that maps each measurable set $A \subset \mathbb{R}^p$ to the probability $\mathbb{P}(X \in A)$, namely, $P_X(A) = \mathbb{P}(X \in A)$; its mean vector is $\mathbb{E}[X] = (\mathbb{E}[X_1], \ldots, \mathbb{E}[X_p])^\top \in \mathbb{R}^p$; its covariance matrix $\mathrm{Cov}(X)$ has entries $\mathrm{Cov}(X_i, X_j)$ for $i, j = 1, \ldots, p$, which is given by

$$\mathrm{Cov}(X) = \mathbb{E}[(X - \mathbb{E}[X])(X - \mathbb{E}[X])^\top] = \mathbb{E}[XX^\top] - \mathbb{E}[X]\mathbb{E}[X]^\top.$$

The covariance matrix $\mathrm{Cov}(X)$ is always symmetric and positive semidefinite. To see this, for any $v \in \mathbb{R}^p$, we have

$$v^\top \mathrm{Cov}(X) v = \mathbb{E}[v^\top (X - \mathbb{E}[X])(X - \mathbb{E}[X])^\top v] = \mathbb{E}[(v^\top (X - \mathbb{E}[X]))^2] \ge 0.$$

If $X = (X_1, \ldots, X_p)$ is a continuous random vector whose density function with respect to the Lebesgue measure is given by $f_X \colon \mathbb{R}^p \to [0, \infty)$, then we have

$$\mathbb{P}(X \in A) = \int_A f_X(x_1, \ldots, x_p)\,\mathrm{d}x_1 \cdots \mathrm{d}x_p$$

for any measurable set $A \subset \mathbb{R}^p$. The j-th coordinate of X is a continuous random variable with density function $f_{X_j} \colon \mathbb{R} \to [0, \infty)$, which is obtained by marginalizing f_X over the other coordinates; for instance, the first coordinate has the density function

$$f_{X_1}(x_1) = \int f_X(x_1, x_2, \ldots, x_p)\,\mathrm{d}x_2 \cdots \mathrm{d}x_p.$$

Similarly, for two coordinates X_i and X_j, we can obtain the joint density function $f_{X_i,X_j}\colon \mathbb{R}^2 \to [0,\infty)$ by marginalizing f_X over the other coordinates.

For a random vector $X = (X_1, \ldots, X_p) \in \mathbb{R}^p$ such that $X_1, \ldots, X_p$ are discrete random variables taking values in a countable set $\mathcal{X} \subset \mathbb{R}$, we have the probability mass function

$$f_X(x_1, \ldots, x_p) := \mathbb{P}(X = (x_1, \ldots, x_p)) = \mathbb{P}(X_1 = x_1, \ldots, X_p = x_p)$$

for $(x_1, \ldots, x_p) \in \mathcal{X}^p$, so that $\mathbb{P}(X \in A) = \sum_{x \in A \cap \mathcal{X}^p} f_X(x)$ for any measurable set $A \subset \mathbb{R}^p$. The j-th coordinate of X is a discrete random variable with probability mass function $f_{X_j}\colon \mathcal{X} \to [0,1]$, which is obtained by marginalizing f_X over the other coordinates. Similarly, one can obtain the joint probability mass function $f_{X_i,X_j}\colon \mathcal{X}^2 \to [0,1]$ for two coordinates X_i and X_j by marginalizing f_X over the other coordinates.

For two random vectors $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$, their covariance matrix $\mathrm{Cov}(X,Y) \in \mathbb{R}^{p \times q}$ has entries $\mathrm{Cov}(X_i, Y_j)$, which is given by

$$\mathrm{Cov}(X,Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])^\top].$$

We say that X and Y are uncorrelated if $\mathrm{Cov}(X,Y) = 0$. We say that X and Y are independent if $\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B)$ for any measurable sets $A \subset \mathbb{R}^p$ and $B \subset \mathbb{R}^q$. If X and Y are independent, then $\mathrm{Cov}(X,Y) = 0$, but the converse is not true in general.

Multivariate Gaussian Distribution

Given $\mu \in \mathbb{R}^p$ and a positive definite matrix $\Sigma \in \mathbb{R}^{p \times p}$, the multivariate Gaussian distribution $N(\mu, \Sigma)$ has the following density function: for $x \in \mathbb{R}^p$,

$$\mathcal{N}(x \,|\, \mu, \Sigma) = ((2\pi)^p \det(\Sigma))^{-1/2} \exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right),$$

which has mean vector μ and covariance matrix Σ. The standard multivariate Gaussian distribution is $N(0, I_p)$, which can be represented as a collection of p independent standard Gaussian random variables, i.e., $X = (X_1, \ldots, X_p) \sim N(0, I_p)$, where $X_1, \ldots, X_p \sim N(0,1)$ are independent. Given $X \sim N(0, I_p)$, let $Y := \mu + \Sigma^{1/2} X$ for any $\mu \in \mathbb{R}^p$ and a positive definite matrix $\Sigma \in \mathbb{R}^{p \times p}$, where $\Sigma^{1/2}$ is the square root of Σ; then, we have $Y \sim N(\mu, \Sigma)$.

Given $x_1, \ldots, x_n \in \mathbb{R}^p$, suppose we want to find $\mu \in \mathbb{R}^p$ and a positive definite matrix $\Sigma \in \mathbb{R}^{p \times p}$ such that $N(\mu, \Sigma)$ best fits the data in the sense of maximum likelihood estimation. The likelihood function of the parameters (μ, Σ) is defined as

$$L(\mu, \Sigma) = \prod_{i=1}^n \mathcal{N}(x_i \,|\, \mu, \Sigma) = ((2\pi)^p \det(\Sigma))^{-n/2} \exp\left(-\frac{1}{2}\sum_{i=1}^n (x_i-\mu)^\top \Sigma^{-1}(x_i-\mu)\right),$$

and the log-likelihood function is given by

$$\ell(\mu, \Sigma) = -\frac{np}{2}\log(2\pi) - \frac{n}{2}\log\det(\Sigma) - \frac{1}{2}\sum_{i=1}^n (x_i-\mu)^\top \Sigma^{-1}(x_i-\mu).$$

To find the maximum likelihood estimator (MLE) of (μ, Σ), we compute the derivatives of the log-likelihood function with respect to μ and Σ and set them to zero. A useful trick here is to consider $\Omega := \Sigma^{-1}$ as a variable instead of Σ, considering the log-likelihood function as a function of (μ, Ω):

$$\bar{\ell}(\mu, \Omega) := -\frac{np}{2}\log(2\pi) + \frac{n}{2}\log\det(\Omega) - \frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^\top \Omega (x_i - \mu),$$

where we use the fact $\det(\Sigma^{-1}) = \det(\Sigma)^{-1}$. First, from

$$\frac{\partial \bar{\ell}}{\partial \mu} = \Omega \sum_{i=1}^{n}(x_i - \mu) = 0,$$

we obtain the MLE of μ as

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Meanwhile, we have

$$\frac{\partial \bar{\ell}}{\partial \Omega} = \frac{n}{2}\Omega^{-1} - \frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)(x_i - \mu)^\top = 0,$$

where we use the following matrix derivative formulas: for a square matrix $A \in \mathbb{R}^{n\times n}$ and a vector $z \in \mathbb{R}^n$, we have

$$\frac{\partial}{\partial A}\log\det(A) = A^{-1} \quad \text{and} \quad \frac{\partial}{\partial A}(z^\top A z) = zz^\top.$$

Combined with the MLE $\hat{\mu}$ of μ, we deduce that the MLE of Ω should satisfy

$$\hat{\Omega}^{-1} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})(x_i - \hat{\mu})^\top,$$

meaning that the MLE of Σ is

$$\hat{\Sigma} = \hat{\Omega}^{-1} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})(x_i - \hat{\mu})^\top.$$

Hence, the MLE of μ is the sample mean, and the MLE of Σ is the sample covariance matrix with n in the denominator, instead of the usual $n - 1$.

Conditional Distributions

Let X and Y be random vectors in $\mathbb{R}^p$ and $\mathbb{R}^q$, whose distributions are given by P_X and P_Y, respectively. The random vector (X, Y) in $\mathbb{R}^{p+q}$ has a joint distribution $P_{X,Y}$. We are interested in understanding the conditional distributions of X given $Y = y$.

When X and Y are discrete, where X and Y take values in countable sets $\mathcal{X}$ and $\mathcal{Y}$, respectively, we compute the conditional probability mass function of X given $Y = y$ as

$$\mathbb{P}(X = x \,|\, Y = y) = \frac{\mathbb{P}(X = x, Y = y)}{\mathbb{P}(Y = y)}.$$

This allows us to compute the conditional probability that X falls in a set $A \subset \mathcal{X}$ given $Y = y$ as

$$\mathbb{P}(X \in A \,|\, Y = y) = \sum_{x \in A} \mathbb{P}(X = x \,|\, Y = y).$$

When X and Y are continuous, where $f_{X,Y} \colon \mathbb{R}^{p+q} \to [0, \infty)$ is the density of (X, Y), recall that the marginal densities of X and Y, namely, $f_X \colon \mathbb{R}^p \to [0, \infty)$ and $f_Y \colon \mathbb{R}^q \to [0, \infty)$ are given by

$$f_X(x) = \int_{\mathbb{R}^q} f_{X,Y}(x, y) \, \mathrm{d}y \quad \text{and} \quad f_Y(y) = \int_{\mathbb{R}^p} f_{X,Y}(x, y) \, \mathrm{d}x.$$

Then, the conditional density of X given $Y = y$ is given as follows: for y such that $f_Y(y) > 0$,

$$f_{X|Y}(x \,|\, y) = \frac{f_{X,Y}(x, y)}{f_Y(y)}.$$

This allows us to compute the conditional probability that X falls in a set $A \subset \mathbb{R}^p$ given $Y = y$ as

$$\mathbb{P}(X \in A \,|\, Y = y) = \int_A f_{X|Y}(x \,|\, y) \, \mathrm{d}x.$$

Comparing Distributions

Let X, Y be random vectors in $\mathbb{R}^p$ with distributions P_X and P_Y, respectively. We are interested in measuring the similarity or difference between these two distributions. The total variation distance between P_X and P_Y is defined as

$$\mathsf{TV}(P_X, P_Y) = \sup_{A \subset \mathbb{R}^p} |P_X(A) - P_Y(A)|,$$

which seeks a measurable set $A \subset \mathbb{R}^p$ on which P_X and P_Y differ the most.

When P_X and P_Y admit densities f_X and f_Y with respect to the Lebesgue measure, we can quantify the difference between P_X and P_Y through the average difference of the densities, given by

$$L_1(P_X, P_Y) = \int_{\mathbb{R}^p} |f_X(x) - f_Y(x)| \, \mathrm{d}x,$$

which is called the L_1 distance between the distributions P_X and P_Y.

In this case, the Kullback-Leibler (KL) divergence between the distributions P_X and P_Y is defined as

$$\mathsf{KL}(P_X \,\|\, P_Y) = \int_{\mathbb{R}^p} f_X(x) \log \frac{f_X(x)}{f_Y(x)} \, \mathrm{d}x,$$

which is well-defined when $f_Y(x) = 0$ implies $f_X(x) = 0$ for almost every $x \in \mathbb{R}^p$. Unlike the above distances, the KL divergence is not symmetric, i.e., $\mathsf{KL}(P_X \,\|\, P_Y) \neq \mathsf{KL}(P_Y \,\|\, P_X)$ in general. That said, the KL divergence is nonnegative because

$$\mathsf{KL}(P_X \,\|\, P_Y) = -\int_{\mathbb{R}^p} f_X(x) \log \frac{f_Y(x)}{f_X(x)} \, \mathrm{d}x \geq -\log \int_{\mathbb{R}^p} f_X(x) \frac{f_Y(x)}{f_X(x)} \, \mathrm{d}x = -\log \int_{\mathbb{R}^p} f_Y(x) \, \mathrm{d}x = 0,$$

where the inequality follows from Jensen's inequality, since the negative logarithm function is convex. From this, one can deduce that the KL divergence is zero if and only if $P_X = P_Y$.

When P_X and P_Y correspond to discrete random variables, the KL divergence can be written in terms of probability vectors consisting of the probability mass functions. Suppose X and Y take values in a countable set $\mathcal{S}$. Then,

$$\mathsf{KL}(P_X \,\|\, P_Y) = \sum_{s \in \mathcal{S}} \mathbb{P}(X = s) \log \frac{\mathbb{P}(X = s)}{\mathbb{P}(Y = s)}.$$