Introduction to Machine Learning. Part 2: Data Probabilistic View

Probabilistic View of Data

Recall our definition of a dataset:

Dataset. Let $\mathcal{A} = \{A_1, \ldots, A_n\}$ be a set of features, with domains $dom(A_1), \ldots, dom(A_n)$ respectively. Let $X = \{\mathbf{x_1}, \ldots, \mathbf{x_m}\}$, where

$$(\forall i = 1 \ldots m)\mathbf{x_i} = (x_{i1}, \ldots, x_{in}),$$

and

$$(\forall j = 1 \ldots n)x_{ij} \in dom(A_j).$$

X is called a *collection of datapoints*, or a *dataset*.

The algebraic view of dataset X approached X as a matrix, and individual rows as vectors in n-dimensional space representing individual data points and individual features as dimensions.

The probabilistic view of X assumes that:

- Each feature A_j is a *random variable* with some distribution over its domain $dom(A_j)$.
- Each data point $\mathbf{x}$ is a collection of instantiations of the random variables $A_1, \ldots, A_n$.

Probability distributions.

Discrete Random Variable. Let S be a finite or countably infinite set. A function $\mathcal{X} : S \longrightarrow \mathbb{R}$ is called a discrete random variable.

Probability Mass Function. Let $\mathcal{X}$ a discrete random variable. A function $f : \mathbb{R} \longrightarrow [0,1]$ is called a *probability mass function* iff

$$\sum_{s \in S} f(\mathcal{X}(s)) = 1$$

Continuous Random Variable. A function $\mathcal{X} : \mathbb{R} \longrightarrow \mathbb{R}$ is called a continuous random variable.

Probability density function. Given a continuous random variable $\mathcal{X}$, a function $f : \mathbb{R} \longrightarrow [0,1]$ is called a probability density function, iff

$$\int_{-\infty}^{\infty} f(x)dx = 1$$

Given a continuous random variable $\mathcal{X}$ with probability density function $f()$ on it, the probability $P(\mathcal{X} \in [a,b]$ is defined as

$$P(\mathcal{X} \in [a,b]) = \int_a^b f(x)dx.$$

Cumulative density function. Let $\mathcal{X}$ be a random variable. A cumulative distribution function $F : \mathbb{R} \longrightarrow [0,1]$ of $\mathcal{X}$ is defined as

$$F(x) = P(\mathcal{X} \leq x)$$

If $\mathcal{X}$ is discrete, then

$$F(x) = \sum_{y \leq x} P(\mathcal{X} = y) = \sum_{y \leq x} f(x).$$

If $\mathcal{X}$ is continuous, then

$$F(x) = \int_{-\infty}^{x} f(y)dy$$

Bivariate Random Variables.

Consider a dataset $X = \begin{pmatrix} | & | \\ X_1 & X_2 \\ | & | \end{pmatrix}$.

Instead of considering X_1 and X_2 are individual random variables, we can view X as a *single bivariate probability distribution*: $X : dom(X_1) \times dom(X_2) \longrightarrow \mathbf{R}^2$.

Here, X assigns each outcome from $dom(X_1) \times dom(X_2)$ a 2-dimensional vector $\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} in \mathbb{R}^2$.

Joint probability mass function. If X_1 and X_2 are discrete random variables, then X has a joint probability mass function defined as

$$f(x_1, x_2) = P(X_1 = x_1, X_2 = x_2)$$

$f(x_1, x_2)$ must satisfy the following conditions:

$$f(x_1, x_2) \geq 0$$

for all pairs $(x_1, x_2) \in \mathcal{R}^2$, and

$$\sum_{s_1 \in dom(X_1)} \sum_{s_2 \in dom(X_2)} f(X_1(s_1), X_2(s_2)) = 1$$

Joint probability distribution function. If X_1 and X_2 are continuous, then X has a joint probability distribution function f:

$$P((x_1, x_2)^T \in W) = \iint_{(x_1,x_2)^T \in W} f(x_1, x_2) dx_1 dx_2.$$

We must observe:

$$f(x_1, x_2) \leq 0$$

for all $(x_1, x_2) \in \mathbb{R}^2$, and

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x_1, x_2) dx_1 dx_2 = 1.$$

Joint Cumulative Distribution Function. The joint cumulative distribution function F for X_1 and X_2 is defined as

$$F(x_1, x_2) = P(X_1 \leq x_1, X_2 \leq x_2)$$

Statistical Independence. Two random variables X_1 and X_2 are *statistically independent* iff, for all subsets $[a_1, b_1], [a_2, b_2] \in \mathbb{R}$

$$P(X_1 \in [a_1, b_1], X_2 \in [a_2, b_2] = P(X_1 \in [a_1, b_1]) \cdot P(X_2 \in [a_2, b_2]).$$

Specifically, if X_1 and X_2 are independent and have probability density/mass functions f_1 and f_2, and cumulative density functions F_1 and F_2, then

$$F(x_1, x_2) = F_1(x_1) \cdot F_2(x_2)$$

and

$$f(x_1, x_2) = f_1(x_1) \cdot f_2(x_2)$$

Multivariate Random Variables

We can extend the notion of bivariate random variable to a multivariate random variable.

An n-dimensional multivariate random variable $X = (X_1, \ldots, X_n)^T$ is a function $X : dom(X_1) \times \ldots \times dom(X_n) \longrightarrow \mathbb{R}^n$.

If $X_1, \ldots X_n$ are all discrete random variables, then X has a probability mass function

$$f(x_1, \ldots, x_n) = P(X_1 = x_1, \ldots, X_n = x_n)$$

such that

$f(x_1, \ldots, x_n) \geq 0$ for all $(x_1, \ldots, x_n)$, and

$$\sum_{(x_1,\ldots,x_n) \in dom(X_1) \times \ldots \times dom(X_n)} f(x_1, \ldots, x_n) = 1$$

If $X_1, \ldots, X_n$ are all continuous random variables, then X has a probability density function

$$P(X \in W) = \int \ldots \int_{(x_1,\ldots,x_n)^T \in W} f(x_1, \ldots, x_n) dx_1 \ldots dx_n$$

Random Sampling and Statistics

The key to understanding the *probabilistic view* of the data is the notion of *random sampling*.

Population. A **population** is a set of all objects/instances under study.

Sample. A subset of the **population** selected/drawn randomly.

A dataset $X = \{\mathbf{x_1}, \ldots, \mathbf{x_m}\}$ is viewed as a *random sample* with individual values x_{ij} drawn from the distributions of the random variables $X_1, \ldots, X_n$.

Statistics. We study samples by computing *sample statistics.*

Univariate Sample. Let X be a random variable. A *random sample of size* m from X is a set of m *independent and identically distributed* random variables $S_1, \ldots, S_m$ which have the same probability mass/density functions as X.

Given a sample $x_1, \ldots, x_m$, the independence assumption implies that

$$f(x_1, \ldots, x_m) = \prod_{i=1}^{m} f_X(x_i)$$

Multivariate Sample. Consider a multivariate random variable $X = (X_1, \ldots, X_n)$. The multivariate sample of size m is a set of *independent and identically distributed* random variables $\mathbf{x_1}, \ldots, \mathbf{x_m}$, $\mathbf{x_i} = (x_{i1}, \ldots, x_{in})$, with joint probability distribution

$$f(\mathbf{x_1}, \ldots, \mathbf{x_m}) = \prod_{i=1}^{m} f_X(\mathbf{x_1})$$

Attribute Independence assumption. Often, we assume that *individual random variables* $X_1, \ldots, X_n$ forming the multivariate random variable X are *independent*. We call this the attribute independence assumption.

Under the attribute independence assumption:

$$f(\mathbf{x_1}, \ldots, \mathbf{x_m}) = \prod_{i=1}^{m} f_X(\mathbf{x_1}) = \prod_{i=1}^{m} \prod_{j=1}^{n} f_{X_j}(x_{ij})$$

In this case, we can replace the need to know the joint probability distribution f_X with knowing the individual, marginal probability distributions $f_{X_1}, \ldots, f_{X_n}$, which is usually "cheaper".

Statistics. Let $\mathbf{S} = \{\mathbf{S_1}, \ldots, \mathbf{S_m}\}$ be a random sample drawn from (multivariate) random variable X. A statistic $\hat{\theta}$ is a function

$$\hat{\theta} : \mathbf{S} \longrightarrow \mathbb{R}$$

that is an *estimator* (a.k.a. *point estimator*) of a corresponding population parameter.

Example. **Sample Mean.** Given a dataset $X = (\mathbf{x_1}, \ldots, \mathbf{x_m}$, the sample mean $\hat{\mu_X}$ is

$$\hat{\mu_X} = \frac{1}{m} \sum_{i=1}^{m} \mathbf{x_i}$$

We treat the *mean vector* of the dataset X as the *estimator for the true population mean* for the population from which the dataset has been drawn.

Univariate Analysis

Univariate Analysis is the analysis of a single attribute/feature, i.e., the analysis of a single random variable in isolation.

Univariate data matrix. Given an attribute X with domain $dom(X)$, the data matrix for a set of data points (dataset) D is

$$D = \begin{pmatrix} X \\ \hline x_1 \\ x_2 \\ \vdots \\ x_m \end{pmatrix}$$

We treat X as a random variable, and treat $x_1, \ldots x_m$ as being drawn from X based on its (possibly unknown) distribution.

Usually, neither *probability mass/distribution function*, not *cumulative probability function* for X are known ahead of time.

A typical goal of univariate analysis is to **estimate the probability distribution function** given the dataset D.

Empirical Cumulative Distribution Function. The *empirical cumulative distribution function* of X given D is

$$\hat{F}(x) = \frac{1}{m} I(x_i \leq x)$$

where

$$I(x_i \leq x) = \begin{cases} 1 & \text{if } x_i \leq x \\ 0 & \text{if } i_i > x \end{cases}$$

$I(.)$ is called a *binary indicator variable*.

Empirical Probability Mass Distribution Function. The empirical probability mass distribution function (PMF) is X given D is

$$\hat{f}(x) = P(X = x) = \frac{1}{n} \sum_{i=1}^{m} I(x_i = x)$$

where

$$I(x_i = x) = \begin{cases} 1 & \text{if } x_i = x \\ 0 & \text{if } i_i \neq x \end{cases}$$

Measures of Central Tendency

Mean. The *mean* or *expected value* of a random variable X is defined as follows. For discrete random variables:

$$\mu_X = E[X] = \sum_{x \in dom(X)} x f(x)$$

For continuous random variables:

$$\mu_X = E[X] = \int_{-\infty}^{\infty} x f(x) dx,$$

where $f(x)$ is the probability mass/probability density function of X.

Sample mean. Given a sample D of random variable X, the *sample mean* is a statistic computed as

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

We can show that $\hat{\mu}$ is an estimator of the (unknown) mean μ of X:

$$\sum_{x \in dom(X)} x\hat{f}(x) = \sum_{x \in dom(X)} x\left(\frac{1}{m}\sum_{i=1}^{m} I(x_i = x)\right) = \frac{1}{n}\sum_{i=1}^{m} x_i = \hat{\mu}$$

Unbiased Estimate. An estimate $\hat{\theta}$ is unbiased for parameter θ if $E[\hat{\theta}] = \theta$ for every possible value of θ.

We can show that the sample mean is unbiased.

Theorem. Let X and Y be two random variables. Then

$$E[aX + bY] = aE[x] + bE[Y],$$

i.e., *mean is additive*.

With this in mind:

$$E[\hat{\mu}] = E\left[\frac{1}{n}\sum_{i=1}^{m} x_i\right] = \frac{1}{m}\sum_{i=1}^{m} E[x_i] = \frac{1}{m}\sum_{i=1}^{m} \mu = \mu.$$

Robustness. A statistic is robust if it is not affected by extreme values. Sample mean is **not robust.**

Median. A median m of a random variable X is a value such that

$$P(X \leq m) \geq \frac{1}{2} \text{ and } P(X \geq m) \geq \frac{1}{2}.$$

Sample median is a **robust statistic.**

Mode. A *mode* of a random variable X is the value where $f(x)$ reaches its maximum value. A *sample mode* is defined as:

$$mode(x) = \arg \max_{x \in dom(X)} \hat{f}(x)$$

Measures of Dispersion

Range of *value range*. Given a random variable X, its *value range* is defined as

$$r = \max_{x \in dom(X)}(x) - \min_{x \in dom(X)}(x)$$

The sample range of X given a sample D is

$$\hat{r} = \max_{i=1}^{m}(x_i) - \min_{i=1}^{m}(x_i).$$

Variance. Recall that we already defined *variance* as the average squared distance of a point in a dataset from the center of the dataset when we discussed the algebraic view of data. In probabilistic view of data, the variance of a random variable X is defined as

$$\sigma^2 = var(X) = E[(X - \mu)^2]$$

For discrete X we have

$$\sigma^2 = var(X) = \sum_{x \in dom(X)} (x - \mu)^2 f(x)$$

For continuous X, the variance is

$$\sigma^2 = var(X) = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx$$

The *standard deviation*, σ is the positive square root of the variance.

Recall:

$$\sigma^2 = E[(X - \mu)^2] = E[X^2 - 2\mu X + \mu^2] = E[X^2] - 2\mu E[X] + \mu^2 =$$

$$= E[X^2] - 2\mu^2 + \mu^2 = E[X^2] + \mu^2 = E[X^2] - (E[X])^2.$$

Sample Variance. Given a random variable X and its sample D, the sample variance is defined as

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{m} (x_i - \hat{\mu})^2.$$

The *sample standard deviation*, $\hat{\sigma}$ is the positive root of $\hat{\sigma}^2$.

Standard Score. The standard score, a.k.a. the z-score of a sample value x_i is

$$z_i = \frac{x_i - \hat{\mu}}{\hat{\sigma}}.$$

The z-score of a sample value x_i is the deviation of x_i from the mean.

Consider a vector

$$Z = \begin{pmatrix} x_1 - \hat{\mu} \\ x_2 - \hat{\mu} \\ \vdots \\ x_m - \hat{\mu} \end{pmatrix}$$

This is often called a *residuals* vector. We can express $\hat{\sigma^2}$ as:

$$\hat{\sigma}^2 = \frac{1}{m}\sum_{i=1}^{m}(x_i - \hat{\mu})^2 = \frac{1}{m}Z^T Z = \frac{1}{m}\|Z\|^2.$$

This means that the *sample variance is the squared magnitude of the residuals vector*.