Introduction to Machine Learning. Part 2: Data Bivariate and Multivariate Analysis

Univariate Analysis Recap

Population vs. Sample.

- **Population:** all objects/instances under study.
- **Sample:** a subset of the population selected randomly.

Analytical Goals. Given a sample, reconstruct the probability distribution from which it was drawn.

Parameters. A statistical parameter is a quantity (characteristic) identifying a specific probability distribution in a family.

Statistics. A sample statistic is an *estimator* of a statistical parameter for the population from a given sample.

Univariate Parameters and Statistics. The analysis of a single random variable (given a sample) uses the following statistcs to estimate the parameters of the underlying distribution:

Distribution Parameter	Sample Statistic	Type
Population mean	Sample mean	Measure of central tendency
Population median	Sample median	Measure of central tendency
Population mode	Sample mode	Measure of central tendency
Value range	Sample range	Measure of dispersion
Population Variance	Sample variance	Measure of dispersion
Population standard deviation	Sample standard deviation	Measure of dispersion

Data Transformations. A univariate sample $X = (x_1 \dots, x_m)^T$ can be tansformed in a number of ways.

- Centering:

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

- Standardization: a.k.a., converstion to z-scores:

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

- Normalization:

$$X_{norm} = \begin{pmatrix} \frac{x_1 - \min(X)}{\max(X) - \min(X)} \\ \frac{x_2 - \min(X)}{\max(X) - \min(X)} \\ \vdots \\ \frac{x_m - \min(X)}{\max(X) - \min(X)} \end{pmatrix}$$

Bivariate Analysis

Bivariate Analysis considers two attributes/features/random variables at the same time.

The dataset X has the shape:

$$D = \begin{pmatrix} \begin{array}{cc} X_1 & X_2 \\ \hline x_{11} & x_{12} \\ x_{21} & x_{22} \\ \vdots & \vdots \\ x_{m1} & x_{m2} \end{array} \end{pmatrix}$$

Empirical Joint Probability Mass Function. Empirical joint probability mass function for X is

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

where $I(\mathbf{x})$ is the indicator function:

$$I(\mathbf{x} = \mathbf{x_i}) = \begin{cases} 1 & \text{if } x_{i1} = x_1 \text{ and } x_{i2} = x_2 \\ 0 & \text{otherwise} \end{cases}$$

Goals. What information about bivariate distributions can be measured? It boils down to:

- Measures of centrality and dispersion (same as in the univariate case)
- Measures of association (new measures)

Measures of location (centrality) and dispersion

Mean. The *bivariate mean* is the expected value of the vector of the univariate means:

$$\mu = E[X] = E\left[\begin{pmatrix} X_1 \\ X_2 \end{pmatrix}\right] = \begin{pmatrix} E[X_1] \\ E[X_2] \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$$

Sample mean. Bivariate sample mean is computed similarly to the univariate sample mean.

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

Variance. The **sample variance** of X given the sample D is

$$var(D) = \hat{\sigma_1}^2 + \hat{\sigma_2}^2$$

Measures of Association

Given two random variables, a natural question is *How are these variables related to each other?*

There are a number of measures of association that can be computed.

Covariance. The *covariance* between two random variables X_1 and X_2 is defined as

$$\sigma_{12} = E[(X_1 - \mu_1)(X_2 - \mu_2)]$$

We can represent covariance as follows:

$$\sigma_{12} = E[(X_1 - \mu_1)(X_2 - \mu_2)] = E[X_1X_2 - X_1\mu_2 - X_2\mu_1 + \mu_1\mu_2] =$$

$$= E[X_1X_2] - \mu_2 E[X_1] - \mu_1 E[X_2] + \mu_1\mu_2 = E[X_1X_2] - \mu_1\mu_2 - \mu_1\mu_2 + \mu_1\mu_2 =$$

$$= E[X_1X_2] - \mu_1\mu_2 = E[X_1X_2] - E[X_1]E[X_2]$$

If X_1 and X_2 are independent, then $E[X_1X_2] = E[X_1]E[X_2]$, which implies, that the covariance of two independent variables is equal to 0.

The converse is not true. If $\sigma_{12} = 0$, then *there is no* ***linear dependency*** *between* X_1 *and* X_2, but there may be more complex relationship between them.

Sample Covariance. The sample covariance between X_1 and X_2 given a sample D is defined as

$$\hat{\sigma}_{12} = \frac{1}{n}\sum_{i=1}^{m}(x_{i1} - \hat{\mu_1})(x_{i2} - \hat{\mu_2}).$$

Correlation. **Correlation** between two variables is the *standardized covariance*:

$$\rho_{12} = \frac{\sigma_{12}}{\sigma_1\sigma_2} = \frac{\sigma_{12}}{\sqrt{\sigma_1^2\sigma_2^2}}$$

Sample Correlation. The **sample correlation** of X_1 and X_2 given a sample D is

$$\hat{\rho}_{12} = \frac{\hat{\sigma}_{12}}{\hat{\sigma}_1\hat{\sigma}_2} = \frac{\sum_{i=1}^{m}(x_{i1} - \hat{\mu_1})(x_{i2} - \hat{\mu_2})}{\sqrt{\sum_{i=1}^{m}(x_{i1} - \hat{\mu})^2\sum_{i=1}^{m}(x_{i2} - \hat{\mu})^2}}$$

Geometry of Sample Correlation. Let

$$Z_1 = \begin{pmatrix} x_{11} - \hat{\mu} \\ x_{12} - \hat{\mu} \\ \vdots \\ x_{1m} - \hat{\mu} \end{pmatrix}$$

$$Z_2 = \begin{pmatrix} x_{21} - \hat{\mu} \\ x_{22} - \hat{\mu} \\ \vdots \\ x_{2m} - \hat{\mu} \end{pmatrix}$$

i.e, Z_1 and Z_2 are centered samples (residual vectors). The sample covariance is then

$$\sigma_{12} = \frac{Z_1^T Z_2}{n},$$

i.e., it is *the dot product of the vectors of residuals of the feature samples normalized by sample size*.

Sample correlation is

$$\hat{\rho_{12}} = \frac{Z_1^T Z_2}{\sqrt{Z_1^T Z_1}\sqrt{Z_2^T Z_2}} = \frac{Z_1^T Z_2}{\|Z_1\|\|Z_2\|} = \left(\frac{Z_1}{\|Z_1\|}\right)^T\left(\frac{Z_2}{\|Z_2\|}\right) = \cos(X_1, X_2)$$

Covariance matrix. The variance-covariance matrix (or simply covariance matrix) of X_1 and X_2 is a matrix

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{pmatrix}$$

Given a sample

$$D^T = \begin{pmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{pmatrix},$$

the sample covariance matrix is

$$\hat{\Sigma} = \begin{pmatrix} \hat{\sigma}_1^2 & \hat{\sigma}_{12} \\ \hat{\sigma}_{21} & \hat{\sigma}_2^2 \end{pmatrix}$$

Let

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix},$$

where

$$\mu_1 = \frac{x_{11} + x_{21}}{2} \text{ and } \mu_2 = \frac{x_{12} + x_{22}}{2}.$$

Let

$$Z_D^T = \begin{pmatrix} x_{11} - \mu_1 & x_{12} - \mu_2 \\ x_{21} - \mu_1 & x_{22} - \mu_2 \end{pmatrix} = D - \mathbf{1}\mu^T,$$

where $\mathbf{1} = (1,1)^T$.

Then

$$\frac{1}{2} Z_D^T Z_D = \frac{1}{2} \begin{pmatrix} x_{11} - \mu_1 & x_{12} - \mu_2 \\ x_{21} - \mu_1 & x_{22} - \mu_2 \end{pmatrix} \cdot \begin{pmatrix} x_{11} - \mu_1 & x_{21} - \mu_1 \\ x_{12} - \mu_2 & x_{22} - \mu_2 \end{pmatrix} =$$

$$= \frac{1}{2} \begin{pmatrix} (x_{11} - \mu_1)(x_{11} - \mu_1) + (x_{12} - \mu_2)(x_{12} - \mu_2) & (x_{11} - \mu_1)(x_{21} - \mu_1) + (x_{12} - \mu_2)(x_{22} - \mu_2) \\ (x_{21} - \mu_1)(x_{11} - \mu_1) + (x_{22} - \mu_2)(x_{12} - \mu_2) & (x_{21} - \mu_1)(x_{21} - \mu_1) + (x_{22} - \mu_2)(x_{22} - \mu_2) \end{pmatrix} =$$

$$= \begin{pmatrix} \frac{1}{2}(\mathbf{x_1} - \mu)^T(\mathbf{x_1} - \mu) & \frac{1}{2}(\mathbf{x_1} - \mu)^T(\mathbf{x_2} - \mu) \\ \frac{1}{2}(\mathbf{x_2} - \mu)^T(\mathbf{x_1} - \mu) & \frac{1}{2}(\mathbf{x_2} - \mu)^T(\mathbf{x_2} - \mu) \end{pmatrix} = \begin{pmatrix} \hat{\sigma}_1^2 & \hat{\sigma}_{12} \\ \hat{\sigma}_{21} & \hat{\sigma}_2^2 \end{pmatrix} = \hat{\Sigma}$$

That is:

> *the sample covariance matrix is the matrix product of the transposed centered sample matrix with itself normalized by 2 (number of attributes).*

Total Variance. The total variance of X_1 and X_2 is

$$var(\mathbf{D}) = tr(\Sigma) = \sigma_1^2 + \sigma_2^2$$

Generalized Variance. The generalized variance of X_1 and X_2 is the determinant of the covariance matrix Σ:

$$|\Sigma| = det(\Sigma) = \sigma_1^2\sigma_2^2 - \sigma_{12}\sigma_{21} = \sigma_1^2\sigma_2^2 - \sigma_{12}^2 = (1 - \rho_{12})\sigma_1^2\sigma_2^2$$

Properties of Covariance Matrix. Covariance (and sample covariance) matrices are:

- Symmetric
- $tr(\Sigma) \geq 0$.
- $|\Sigma| \geq 0$.

Multivariate Analysis

Multivariate analysis extends the notions of *mean*, *variance* and *covariance* onto the case of multiple $n > 2$ attributes.

Let $X = (X_1, \ldots, X_n)$ and

$$\mathbf{D} = \left(\begin{array}{cccc} X_1 & X_2 & \ldots & X_n \\ \hline x_{11} & x_{12} & \ldots & x_{1n} \\ x_{21} & x_{22} & \ldots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & \ldots & x_{mn} \end{array} \right)$$

The individual data points $\mathbf{x_j} = (x_{j1}, x_{j2}, \ldots, x_{jn})^T$ are in the rows of $\mathbf{D}$.

Probabilistic view. Data points $\mathbf{x_j}$ are a random sample from X. Each point $\mathbf{x_j}$ is drawn independently from other points from $\mathbf{D}$ and it is identically distributed as X.

Mean. The *multivariate mean vector* is defined as

$$\mu = E[X] = \begin{pmatrix} E[X_1] \\ E[X_2] \\ \vdots \\ E[X_n] \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{pmatrix}$$

Covariance Matrix. The *covariance matrix* Σ is defined as

$$\Sigma = E[(X - \mu)(X - \mu)^T] = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \ldots & \sigma_{1n} \\ \sigma_{12} & \sigma_2^2 & \ldots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \ldots & \sigma_n^2 \end{pmatrix}$$

Recall that $\sigma_{ij} = \sigma_{ji}$. Thus, the covariance matrix is *symmetric*.

Positive semi-definite matrixes. A symmetric square matrix M is said to be *positive-definite* iff for *every vector* z

$$z^T M z > 0$$

A symmetric square matrix M is said to be *positive semi-definite* iff for *every vector* z

$$z^T M z \geq 0$$

Theorem. The covariance matrix Σ of the set of attributes X is *a positive semi-definite matrix*.

Proof. By definition of Σ:

$$z^T \Sigma z = z^T E[(X-\mu)(X-\mu)^T] z = E[z^T (X-\mu)(X-\mu)^T z] = E[Y^2]$$

for $Y = z^T (X - \mu)$.

(more about this later).

Total and generalized variance. Just as in binary case:

$$var(\Sigma) = tr(\Sigma) = \sigma_1^2 + \ldots + \sigma_n^2,$$

and

$$|\Sigma| = det(\Sigma)$$

is the generalized variance.

Sample Covariance Matrix. Given the collection of attributes X and a sample $\mathbf{D}$, the sample covariance matrix is

$$\hat{\Sigma} = E[(X - \hat{\mu})(X - \hat{\mu})^T]$$

Similarly to the binary case, let

$$\mathbf{Z} = \mathbf{D} - \mathbf{1} \cdot \hat{\mu}^T = \begin{pmatrix} \mathbf{x_1}^T - \hat{\mu}^T \\ \mathbf{x_2}^T - \hat{\mu}^T \\ \vdots \\ \mathbf{x_m}^T - \hat{\mu}^T \end{pmatrix}$$

be the vector of centered samples.

Then, $\hat{\Sigma}$ can be expressed as

$$\hat{\Sigma} = \frac{1}{n} \mathbf{Z}^T \mathbf{Z}.$$