VSB TECHNICAL UNIVERSITY OF OSTRAVA | FACULTY OF ELECTRICAL ENGINEERING AND COMPUTER SCIENCE | DEPARTMENT OF COMPUTER SCIENCE

Fundamentals of Machine Learning

Statistical Data Features

Jan Platos

November 15, 2023

Statistical Data Features

Statistical Data Features

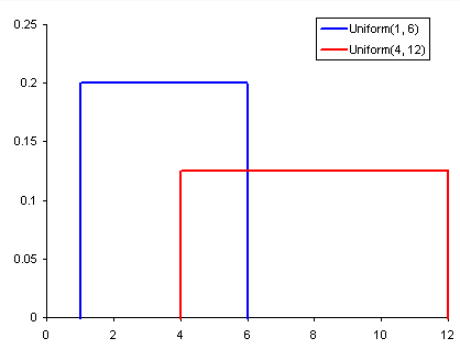
- What kind of statistical features exits?
- Which are important for a data analysis?
- May statistics compute the similarity of two features?

Statistical Data Features - Population

- Populations vs. Sample
- Population is a set of all objects in a group.
- Sample is a subset of the populations.
- Randomness of a sample is a biggest question.

Statistical Data Features - Probability distribution

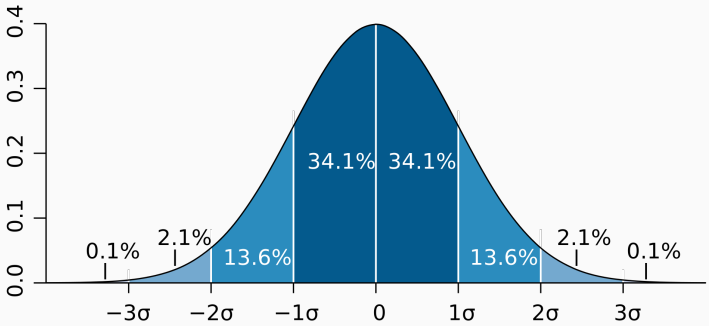
- Probability distribution is a function that shows the probabilities of the outcomes of an event or experiment.



https://towardsdatascience.com/the-5-basic-statistics-concepts-data-scientists-need-to-know-2c96740377ae

Statistical Data Features - Probability distribution

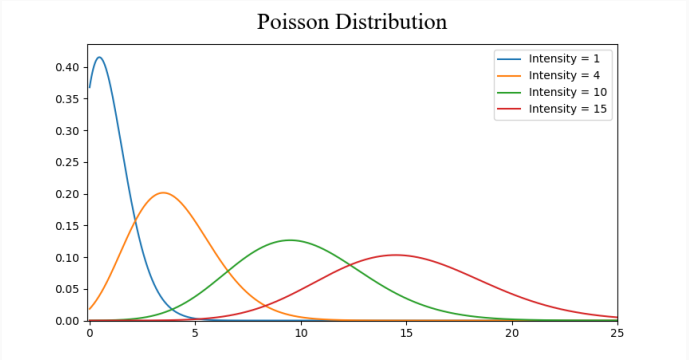
- Probability distribution is a function that shows the probabilities of the outcomes of an event or experiment.



https://towardsdatascience.com/10-must-know-statistical-concepts-for-data-scientists-645619783c08

Statistical Data Features - Probability distribution

- Probability distribution is a function that shows the probabilities of the outcomes of an event or experiment.



https://towardsdatascience.com/the-5-basic-statistics-concepts-data-scientists-need-to-know-2c96740377ae

Statistical Data Features - Central tendency

- Central tendency is the central (or typical) value of a probability distribution. The most common measures of central tendency are mean, median, and mode.
- **Mean** is the average of the values in series.
- **Median** is the value in the middle when values are sorted in ascending or descending order.
- **Mode** is the value that appears most often.

https://towardsdatascience.com/10-must-know-statistical-concepts-for-data-scientists-645619783c08

Statistical Data Features - Variance and Standard deviation

- Variance is a measure of the variation among values.

$$Variance = \frac{\sum (x_i - \overline{x})^2}{N}$$

- Standard deviation is a measure of how spread out the values are.

$$StdDev = \sigma = \sqrt{Variance}$$

Statistical Data Features - Expected value

- The expected value of a random variable is the weighted average of all possible values of the variable.
- Discreet variables may be evaluated in a direct way

$$E(X) = \sum p(x_i)\, x_i$$

- Continuous variables needs approximation using PDF - Probability density function.
- PDF specifies the probability of a random variable taking value within a particular range.

$$E(X) = \int_{x_{min}}^{x_{max}} x\, PDF(x)$$

Statistical Data Features - Covariance and Correlation

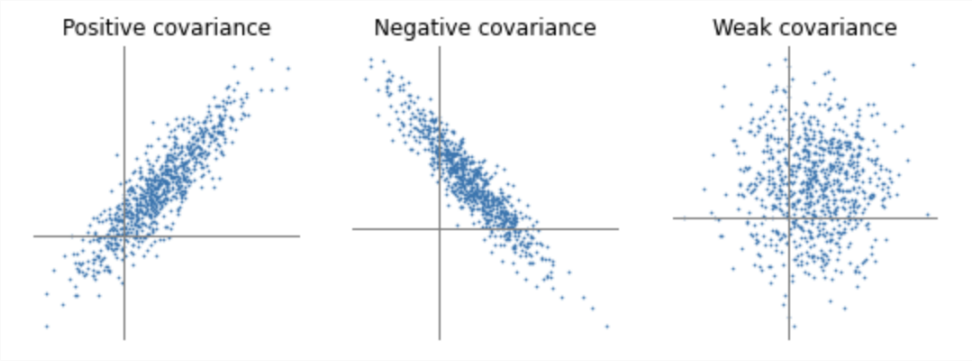
- Covariance is a quantitative measure that represents how much the variations of two variables match each other.

$$Cov(X, Y) = E\left[(X - E(X))\,(Y - E(Y))\right]$$

- For discreet variables:

$$Cov(X, Y) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})\,(y_i - \overline{y})$$

Statistical Data Features - Covariance and Correlation



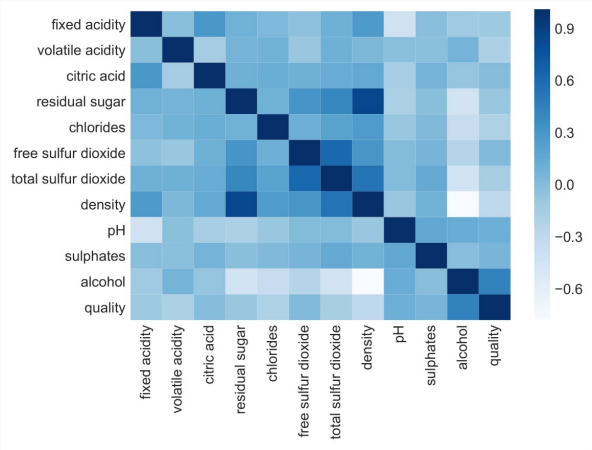
https://programmathically.com/covariance-and-correlation/

Statistical Data Features - Covariance and Correlation

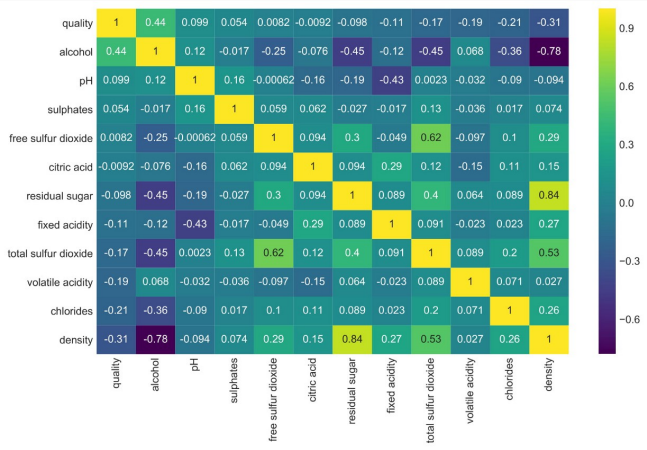
- Correlation is a normalization of covariance by the standard deviation of each variable.

$$Corr(X, Y) = \frac{Cov(X, Y)}{\sigma_x \sigma_y}$$

Statistical Data Features - Covariance and Correlation



Statistical Data Features - Covariance and Correlation



Questions