Mathematics Applicable in Data Science

AlvineW.W

September 24, 2023

In python Scikit learn - mathematics is covered you dont need to know every bit of it to make progress in Data Science but it is crucial this presents a snapshort. To let you know that this is what underpins most of algorithms used in Data science.

Vectors

Vectors are fundamental in the fields of physics, engineering, and data science. They can be used to represent physical quantities such as forces or velocities that have both magnitude and direction.
1. **Vector Definition**: A vector in a n-dimensional space is a set of ordered numbers. For example, a vector in 2D space can be represented as

$$\vec{v} = (v_1, v_2)$$

2. **Vector Addition**: Two vectors can be added together to form a new vector. If we have two vectors

$$\vec{a} = (a_1, a_2)$$

and

$$\vec{b} = (b_1, b_2)$$

, their sum is given by

$$\vec{a} + \vec{b} = (a_1 + b_1, a_2 + b_2)$$

3. **Scalar Multiplication**: A vector can be multiplied by a scalar (a real number). If we have a vector

$$\vec{v} = (v_1, v_2)$$

and a scalar c, the scalar multiplication is given by

$$c\vec{v} = (cv_1, cv_2)$$

4. **Dot Product**: The dot product of two vectors is a scalar quantity. If we have two vectors

$$\vec{a} = (a_1, a_2)$$

and

$$\vec{b} = (b_1, b_2)$$

, their dot product is given by

$$\vec{a} \cdot \vec{b} = a_1 b_1 + a_2 b_2$$

5. **Magnitude of a Vector**: The magnitude (or length) of a vector

$$\vec{v} = (v_1, v_2)$$

is given by

$$||\vec{v}|| = \sqrt{v_1^2 + v_2^2}$$

. This comes from the Pythagorean theorem
6. **Unit Vector**: A unit vector is a vector with magnitude 1. It's often used to represent direction. A unit vector in the same direction as

$$\vec{v}$$

is given by

$$\hat{v} = \frac{\vec{v}}{||\vec{v}||}$$

7. **Vector Projection**: The projection of vector

$$\vec{a}$$

onto another vector

$$\vec{b}$$

is given by

$$proj_{\vec{b}}\vec{a} = \frac{\vec{a} \cdot \vec{b}}{||\vec{b}||^2} \cdot \vec{b}$$

. This represents how much of

$$\vec{a}$$

lies in the direction of

$$\vec{b}$$

These are some of the basic operations involving vectors that are crucial in data science for operations such as calculating distances and understanding directions of data points in multi-dimensional space.

Matrix

A matrix is a two-dimensional data structure where numbers are arranged into rows and columns. For example, a matrix A with 2 rows and 3 columns is represented as:

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix}$$

Matrix Operations Matrices can be added and subtracted element by element, and multiplied according to the rules of matrix multiplication. For example, if A and B are two matrices of the same size, their sum is given by $A + B$, and their difference is given by $A - B$.

Matrix in Python In Python, matrices can be represented using lists or using the NumPy library's array object. Here's an example of how to create a matrix in Python:

```
import numpy as np

A = np.array([[1, 2, 3], [4, 5, 6]])
print(A)
```

Matrix Operations in Python Python's NumPy library also supports matrix operations. For example, you can add two matrices using the ‘+‘ operator or the ‘np.add()‘ function, subtract them using the ‘-‘ operator or the ‘np.subtract()‘ function, and multiply them using the ‘@‘ operator or the ‘np.dot()‘ function.

Application in Data Science Matrices are used in various fields of data science. For example, in machine learning algorithms like linear regression and logistic regression, the dataset is usually represented as a matrix where each row is a data point and each column is a feature.

Statistics

Central Tendencies

Central tendency refers to the measure that determines the center of a distribution. The most common measures of central tendency are:

- **Mean**: The average of all data points. If we have n data points $x_1, x_2, \ldots, x_n$, the mean μ is given by $\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$.

- **Median**: The middle value in a sorted dataset. If the dataset has an odd number of observations, the median is the middle value. If it has an even number of observations, the median is the average of the two middle values.

- **Mode**: The most frequently occurring value in a dataset.

Dispersion

Dispersion refers to measures that determine how spread out the values in a dataset are. Common measures of dispersion include:

- **Range**: The difference between the maximum and minimum values.

- **Variance**: The average of the squared differences from the mean. It is given by $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$.

- **Standard Deviation**: The square root of the variance, denoted as σ.

Covariance

Covariance is a measure that indicates the extent to which two variables change in tandem. If we have two variables X and Y with means μ_X and μ_Y respectively, their covariance is given by $cov(X, Y) = E[(X - \mu_X)(Y - \mu_Y)]$.

Correlation

Correlation is a measure that determines how strong a relationship is between two variables. Correlation values range from -1 to 1. A value of 1 means a perfect positive correlation, -1 means a perfect negative correlation, and 0 means no correlation. The correlation between two variables X and Y is given by $corr(X, Y) = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$.

These concepts are fundamental in statistics and data science as they provide insights into the data and form the basis for many statistical tests and models.

In Python, these concepts can be easily computed using libraries such as NumPy and pandas. For example, to compute the mean of a list of numbers, you can use the ‘numpy.mean()‘ function. Similarly, you can use ‘numpy.var()‘ for variance, ‘numpy.std()‘ for standard deviation, ‘numpy.cov()‘ for covariance, and ‘numpy.corrcoef()‘ for correlation.

Probability

Probability is a branch of mathematics that deals with the occurrence of a random event. The value is expressed from zero to one. Probability has been introduced in Maths to predict how likely events are to happen. The meaning of probability is basically the extent to which something is likely to happen.
In more formal terms, probability is a measure of the likelihood of an event to occur. Many events cannot be predicted with total certainty. We can predict only the chance of an event to occur i.e., how likely they are going to happen. Probability can range from 0 to 1, where 0 means the event to be an impossible one and 1 indicates a certain event

Dependence and Independence

Two events A and B are independent if the occurrence of A does not affect the occurrence of B, and vice versa. Mathematically, this is expressed as $P(A \cap B) = P(A)P(B)$. If this equation does not hold, then the events are dependent.

Conditional Probability

Conditional probability is the probability of an event given that another event has occurred. If we are interested in the probability of event A given that event B has occurred, we write this as $P(A|B)$, which is calculated as $\frac{P(A \cap B)}{P(B)}$.

Bayes's Theorem

Bayes's theorem describes the probability of an event based on prior knowledge of conditions that might be related to the event. It is stated mathematically as $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$.

Random Variable

A random variable is a variable whose possible values are numerical outcomes of a random phenomenon. There are two types of random variables, discrete and continuous.

Continuous Distributions

A continuous distribution describes the probabilities of the possible values of a continuous random variable. A continuous random variable is a random variable with a set of possible values (known as the range) that is infinite and uncountable.

Normal Distribution

The normal distribution is a probability function that describes how the values of a variable are distributed. It is a symmetric distribution where most of the observations cluster around the central peak and the probabilities for values further away from the mean taper off equally in both directions.

Central Limit Theorem

The central limit theorem states that if you have a population with mean μ and standard deviation σ and take sufficiently large random samples from the population with replacement, then the distribution of the sample means will be approximately normally distributed.

Hypothesis and Inference

Hypothesis testing is a statistical method that is used in making statistical decisions using experimental data. A hypothesis is an assumption that we make about the population parameter. This assumption may or may not be true. Hypothesis testing is a critical tool in inferential statistics, for determining what outcomes of a study would lead to a rejection of the null hypothesis for a pre-specified level of significance.

Statistical Hypothesis Testing

Statistical hypothesis testing is a method of making decisions using data from a scientific study. In statistics, a result is called statistically significant if it has been predicted as unlikely to have occurred by chance alone, according to a pre-determined threshold probability, the significance level. The process of hypothesis testing involves setting up two competing hypotheses, the null hypothesis and the alternative hypothesis. One selects a random sample (or multiple samples when there are more comparison groups), computes summary statistics and then assesses the likelihood that the observed data fall under the null or alternative hypotheses.

Type I and Type II Errors

Type I error occurs when the null hypothesis is true, but is rejected. It is denoted by alpha. In other words, this is the error of overreacting and by rejecting the null hypothesis when it is true. On the other hand, we have a Type II error when we do not reject the null hypothesis and it is false. It is denoted by beta. This is an error of under reaction or failing to act, i.e., failing to reject the null hypothesis when you should have done so.

P-values

The p-value is used in hypothesis testing to help you support or reject the null hypothesis. It represents the probability that the results of your test occurred at random. If p-value $p - value \leq 0.05$, you reject the null hypothesis. The smaller the p-value, the stronger the evidence that you should reject the null hypothesis.

Confidence Intervals

A confidence interval gives an estimated range of values which is likely to include an unknown population parameter, the estimated range being calculated from a given set of sample data. If independent samples are taken repeatedly from the same population, and a confidence interval calculated for each sample, then a certain percentage (confidence level) of the intervals will include the unknown population parameter.

Bayesian Inference

Bayesian inference is a method of statistical inference in which Bayes' theorem is used to update the probability for a hypothesis as more evidence or information becomes available. Bayesian inference has found application in a wide range of activities, including science, engineering, philosophy, medicine, sport, and law.