Basic statistics for applied machine learning

Sabber Ahamed
sabbers@gmail.com
linkedin.com/in/sabber-ahamed
github.com/msahamed

July 24, 2023

Contents

1 Descriptive Statistics

Statistics are crucial in machine learning, providing a foundation for understanding data and building models. This section covers some essential statistical concepts and their applications in machine learning.

Descriptive statistics provide a concise overview of the main characteristics of a dataset, helping us better understand its structure and tendencies. In data science, understanding the underlying statistics of a dataset is essential before applying any machine learning model, as this foundational knowledge aids in feature selection, preprocessing, and model evaluation.

Given a dataset: $[2, 4, 6, 8, 10]$

1.1 Averaging Methods

Averaging techniques allow us to represent datasets with a single, central value. They offer insights into the "typical" value of a dataset. Here, we will discuss common averaging methods using our dataset for consistent examples.

1.1.1 Arithmetic Mean

The arithmetic mean represents the sum of all values divided by the number of values.

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

For our dataset, the arithmetic mean is:

$$\bar{x} = \frac{2+4+6+8+10}{5} = 6$$

The arithmetic mean offers a straightforward way of understanding the "center" of our data.

1.1.2 Geometric Mean

The geometric mean is particularly relevant when considering data that exhibits multiplicative behavior.

$$\bar{x}_{\text{geo}} = \sqrt[n]{\prod_{i=1}^{n} x_i}$$

For our dataset:

$$\bar{x}_{\text{geo}} = \sqrt[5]{2 \times 4 \times 6 \times 8 \times 10} \approx 5.48$$

The geometric mean offers a clearer picture of average growth in contexts like growth rates.

1.1.3 Harmonic Mean

The harmonic mean is crucial when considering rates or ratios.

$$\bar{x}_{\text{harm}} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$$

Using our dataset:

$$\bar{x}_{\text{harm}} = \frac{5}{\frac{1}{2}+\frac{1}{4}+\frac{1}{6}+\frac{1}{8}+\frac{1}{10}} \approx 3.41$$

In contexts like speed, the harmonic mean provides an average skewed towards the smaller values.

1.1.4 Weighted Mean

Some values might be inherently more significant than others, necessitating a weighted mean.

$$\bar{x}_{\text{weighted}} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

For our dataset with weights $[1, 2, 3, 4, 5]$:

$$\bar{x}_{\text{weighted}} = \frac{1 \times 2 + 2 \times 4 + 3 \times 6 + 4 \times 8 + 5 \times 10}{15} = 7.33$$

The weighted mean accounts for the importance of each data point, offering a more nuanced average.

1.2 Variance

Variance measures the spread of the data, indicating how far data points deviate from the mean on average.

$$\sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}$$

For our dataset, where $\bar{x} = 6$:

$$\sigma^2 = \frac{(2-6)^2 + (4-6)^2 + (6-6)^2 + (8-6)^2 + (10-6)^2}{5} = 8$$

Variance aids in understanding the dispersion of our data, which can impact the model's accuracy and precision.

1.3 Standard Deviation

The standard deviation represents the average distance between the data points and the mean.

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}}$$

For our dataset:

$$\sigma = \sqrt{8} \approx 2.83$$

A lower standard deviation indicates data points are closer to the mean, whereas a higher value suggests more spread-out data.

1.4 Covariance

Covariance indicates how two variables change together. If they tend to increase and decrease simultaneously, covariance is positive; if one increases while the other decreases, it's negative.

Given two datasets X and Y:

$$\text{cov}(X, Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n}$$

For machine learning, understanding covariance is essential. Two highly correlated features convey similar information, and one might be dropped to reduce dimensionality. In algorithms like PCA, covariance matrices help transform data to capture the most variance with the fewest features.

2 Correlation Methods

Correlation measures the strength and direction of the relationship between two variables. There are several correlation methods, both linear and non-linear, each with its unique characteristics and applications. This section'll discuss Pearson, Spearman, and Alternating Conditional Expectation (ACE) correlation methods.

2.1 Pearson Correlation

Pearson correlation measures the linear relationship between two variables. It is the most commonly used correlation method. The Pearson correlation coefficient (r) ranges from -1 to 1, where -1 indicates a strong negative correlation, 1 indicates a strong positive correlation, and 0 indicates no correlation.

The Pearson correlation coefficient is calculated as follows:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Example: Consider two variables, $x = [1, 2, 3, 4, 5]$ and $y = [2, 4, 5, 6, 8]$. The Pearson correlation coefficient can be calculated as:

$$r \approx 0.97$$

This high positive value indicates a strong positive linear relationship between the two variables.

2.2 Spearman Correlation

Spearman correlation, also known as Spearman's rank correlation, measures the monotonic relationship between two variables. It is a non-parametric method that evaluates the correlation based on the ranks of the data rather than the actual values. The Spearman correlation coefficient (ρ) ranges from -1 to 1.

The Spearman correlation coefficient is calculated as follows:

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2 - 1)}$$

where d_i is the difference between the ranks of corresponding values in the two variables.

Example: Using the same variables as before, $x = [1, 2, 3, 4, 5]$ and $y = [2, 4, 5, 6, 8]$, the ranks are $R_x = [1, 2, 3, 4, 5]$ and $R_y = [1, 2, 3, 4, 5]$. The Spearman correlation coefficient can be calculated as:

$$\rho = 1$$

This value indicates a perfect positive monotonic relationship between the two variables.

2.3 Alternating Conditional Expectation (ACE)

Alternating Conditional Expectation (ACE) is a non-parametric and non-linear correlation method that estimates the relationship between two variables by iteratively minimizing the conditional variance. Unlike Pearson and Spearman correlations, ACE does not assume any specific form for the relationship between the variables. ACE is an iterative algorithm that generally involves the following steps:

1. Initialize the transformed variables as the original variables.
2. Estimate the conditional expectations.
3. Update the transformed variables.
4. Repeat steps 2 and 3 until convergence.

ACE does not have a simple equation like Pearson or Spearman correlation coefficients. It requires more advanced statistical techniques and computational methods to calculate the correlation between the variables. The output of ACE is typically visualized as a scatter plot or curve representing the relationship between the two variables.

2.4 Usages of Correlation Methods

In machine learning, correlation is vital in understanding the relationships between variables and selecting the most relevant features for model building. Here are some key points on how correlation can be used in machine learning:

- **Feature Selection**: By calculating the correlation between input features and the target variable, we can identify which features have a strong relationship with the target and are more likely to provide valuable information for prediction.
- **Multicollinearity Detection**: Correlation helps identify multicollinearity, where two or more highly correlated input features. Multicollinearity can lead to unstable models and reduce the interpretability of the feature importance.
- **Addressing Multicollinearity**: By analyzing the correlation between input features, we can detect multicollinearity and address it, for instance, by removing one of the highly correlated features or using dimensionality reduction techniques such as Principal Component

Analysis (PCA).

- **Understanding Relationships**: Correlation analysis can reveal relationships between variables, providing insights into the underlying structure of the dataset and helping guide feature engineering and model selection.

In summary, correlation is a powerful tool in machine learning for feature selection, detecting and addressing multicollinearity, and understanding relationships between variables. We can build more reliable, accurate, and interpretable models by leveraging correlation analysis.

3 Basic probability theory

3.1 Concepts

Probability is a mathematical framework for quantifying uncertainty. It provides a way to model random events and reason about them.

Experiment: A procedure that yields one of several possible outcomes, e.g., tossing a coin or rolling a die. Sample space (Ω): The set of all possible outcomes of an experiment, e.g., $\Omega = H, T$ for a coin toss. Event (A): A subset of the sample space, e.g., $A = H$, when we are interested in the coin landing heads up. Probability function (P): A function that assigns a probability value to each event, satisfying the following axioms: $0 \leq P(A) \leq 1$ for any event A. $P(\Omega) = 1$. If $A_1, A_2, \ldots$ are mutually exclusive events (i.e., $A_i \cap A_j = \emptyset$ for $i \neq j$), then $P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$. **Example:** Tossing a fair coin twice.

Sample space: $\Omega = (H, H), (H, T), (T, H), (T, T)$ Event of interest: $A = (H, T), (T, H)$ (exactly one head) Probability function: $P(A) = \frac{2}{4} = \frac{1}{2}$

3.2 Conditional Probability

Conditional probability is the probability of an event occurring, given that another event has occurred. It is denoted as $P(A|B)$, read as "the probability of event A given event B."

The formula for conditional probability is:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

where $P(A \cap B)$ is the probability of both events A and B occurring. The conditional probability of A given B is equal to the probability of both events occurring divided by the probability of B occurring. This formula can be applied to any two events A and B,

provided that $P(B) > 0$.

Example: Consider drawing a card from a standard deck of 52 playing cards. Let A be the event of drawing an Ace, and B be the drawing of a heart.

$P(A) = \frac{4}{52} = \frac{1}{13}$ $P(B) = \frac{13}{52} = \frac{1}{4}$ There is only one Ace of Hearts in the deck, so $P(A \cap B) = \frac{1}{52}$ $P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{\frac{1}{52}}{\frac{1}{4}} = \frac{1}{13}$ The conditional probability of drawing an Ace given that the drawn card is a heart is $\frac{1}{13}$.

3.3 Bayes' Theorem

Bayes' theorem is a way to calculate the probability of an event given prior knowledge of conditions that might be related to the event. It is named after Reverend Thomas Bayes, who published a paper on the theorem in 1763.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

where $P(A)$ is the probability of event A occurring, $P(B)$ is the probability of event B occurring, and $P(A|B)$ is the probability of event A occurring given that B is true. $P(B|A)$ is the probability of event B occurring given that A is true.

In machine learning, Bayes' theorem is used in Bayesian statistics, a framework for updating the probability of a hypothesis as more evidence or information becomes available. It is also used in Bayesian inference, a method of statistical inference that uses Bayes' theorem to update the probability for a hypothesis as more evidence or information becomes available.

4 Distributions

Probability distributions describe the probabilities of all possible outcomes for a random variable. They can be discrete or continuous, depending on the nature of the random variable.

4.1 Discrete Distributions

Discrete distributions are used for random variables that have a finite or countably infinite set of possible outcomes.

4.1.1 Bernoulli Distribution:

The Bernoulli distribution models a binary outcome (success or failure) with probability p for success and $1 - p$ for failure. It is useful for modeling binary classification problems in machine learning, where the output is 0 or 1. Figure 1 shows the Bernoulli distribution.

The following equation represents the Bernoulli distribution:

$$P(X = k) = p^k(1 - p)^{1-k}$$

where, k is the outcome (either 0 or 1) and p is the probability of success (1). And p^k is the probability of k successes and $(1-p)^{1-k}$ is the probability of $1 - k$ failures.

4.1.2 Binomial Distribution:

The binomial distribution models the number of successes in n independent Bernoulli trials with the same probability of success p. It can be used in machine learning to analyze the performance of a binary classifier on multiple instances. Figure 1 shows the Binomial distribution. Each bar represents the probability of k successes in n trials.

$$P(X = k) = \binom{n}{k} p^k(1 - p)^{n-k}$$

Where $\binom{n}{k}$ is the binomial coefficient, which is the number of ways of choosing k items from a set of n items, where the order of the items does not matter.

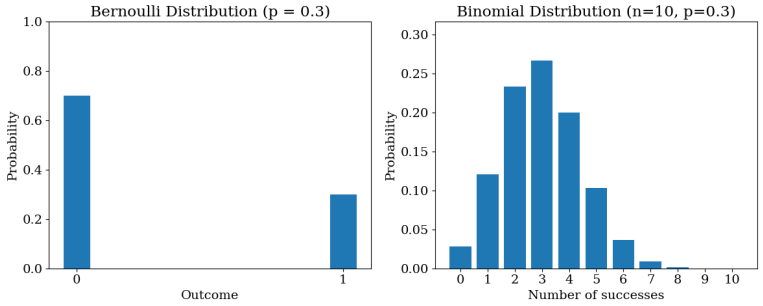

Figure 1: The plot shows the Bernoulli distribution (which essentially shows the probability of success and failure) and then the Binomial distribution, which represents the probability distribution of the number of successes in a fixed number of Bernoulli trials. In this example, the probability of success (e.g., flipping heads on a biased coin) is set to $p = 0.3$, and the number of trials for the Binomial distribution is set to $n = 10$.

And p^k is the probability of k successes and $(1-p)^{n-k}$ is the probability of $n-k$ failures.

4.1.3 Poisson Distribution:

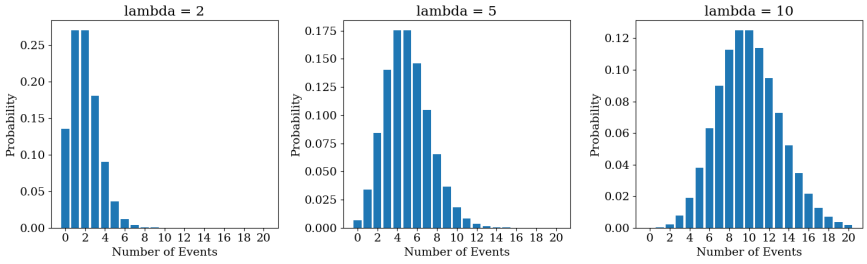

Figure 2: The plot shows the Poisson distributions with different λ.

The Poisson distribution models the number of events occurring in a fixed interval of time or space, given a constant average rate of occurrence λ. In machine learning, it can be used to model rare events, like the number of system failures or clicks on an ad within a time interval. Figure 2 shows the Poisson distribution.

$$P(X = k) = \frac{e^{-\lambda}\lambda^k}{k!}$$

Where, e is the base of the natural logarithm, λ is the average number of occurrences in a given interval, and $k!$ is the factorial of k.

4.2 Continuous Distributions

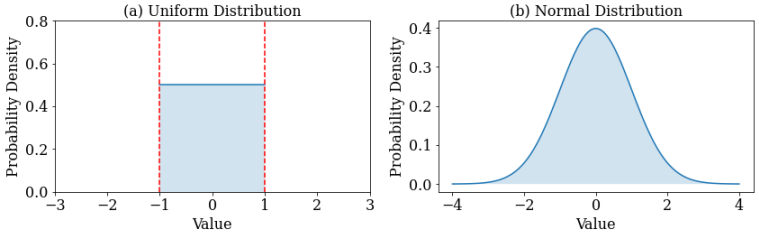

Figure 3: The plot shows two common probability distributions: the Uniform Distribution (a) and the Normal Distribution (b). These distributions represent how data points are spread across a range of values

Continuous distributions are used for random variables that have an infinite set of possible outcomes. They are often used to model real-valued quantities.

4.2.1 Uniform Distribution:

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{if } a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}$$

Where a is the lower bound of the interval, and b is the upper bound of the interval. The uniform distribution models a random variable with equal probability density for all values within a specific interval $[a, b]$.

Figure-3a shows a Uniform Distribution. This distribution has a constant probability density function (PDF) between two boundaries (red dashed lines). This means that any value between these boundaries is equally likely to occur. In this specific example, the boundaries are -1 and 1. The shaded area under the curve represents the probability of observing a value within the given range

In machine learning, it can be used as a prior distribution for parameters or for generating random numbers in specific intervals, e.g., initializing weights in a neural network.

4.2.2 Normal Distribution

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Where μ is the mean of the distribution, σ^2 is the variance of the distribution, and σ is the standard deviation of the distribution.

The normal distribution, also known as the Gaussian distribution, is a bell-shaped continuous probability distribution widely used in machine learning due to its attractive properties. Many natural phenomena and measurement errors follow normal distributions. In machine learning, it is commonly used as a prior distribution for parameters, noise models, and for generating random numbers.

Figure-3b shows a normal distribution known as the Gaussian or "bell curve" distribution. The PDF of this distribution has a symmetric bell shape, with a peak at the mean value (0 in this case) and tails on both sides. Most data points are concentrated around the mean, with fewer points appearing farther from the mean. The shaded area under the curve represents the probability of observing a value within the given range.

Normal distribution is important in Machine Learning

Many statistical methods and tests assume data follows a normal distribution, making it a fundamental concept in statistical hypothesis testing. Some of the key applications of the normal distribution in machine learning include:

1. **Central Limit Theorem:** Regardless of the original data's shape, the sample means distribution approaches a normal distribution as the sample size increases. This theorem underpins many machine learning algorithms, especially those based on statistical properties.
2. **Performance Metrics:** In regression models, assuming errors are normally distributed can be crucial for interpreting and validating the results.

3. **Simplifies Complexity:** A lot of natural phenomena can be modeled using the normal distribution due to its simplicity and tractability.

Understanding your Data

1. **Mean (μ):** Represents the central tendency of the distribution. The mean, median, and mode are all equal in a normal distribution.
2. **Standard Deviation (σ):** Measures the variation or dispersion in the data. Approximately 68%
3. **Skewness:** Measures the asymmetry of the distribution. A normal distribution has a zero skewness, meaning it's perfectly symmetric.
4. **Kurtosis:** Describes the tails and sharpness of the distribution. A normal distribution has a kurtosis of three (excluding the subtraction of three, which is often used to ease calculations). Higher kurtosis indicates heavier tails, while lower kurtosis indicates lighter tails.

Transforming Data to a Normal Distribution

In real-world scenarios, data might only sometimes be normally distributed. There are many benefits to having normally distributed data, like the ability to use parametric statistical methods and the simplicity of the distribution.

However, various techniques can transform data to approximate a normal distribution:

- **Log Transformation:** Suitable for data that exhibits right-skewness.
- **Square Root or Cube Root Transformation:** Helps reduce the impact of outliers.
- **Box-Cox Transformation:** A family of power transformations that can stabilize variance and make the data more normal.

5 Central Limit Theorem

The Central Limit Theorem (CLT) is an important concept in statistics and machine learning. Before we move on to Central Limit Theorem, let's define some terms that will be used in the theorem:

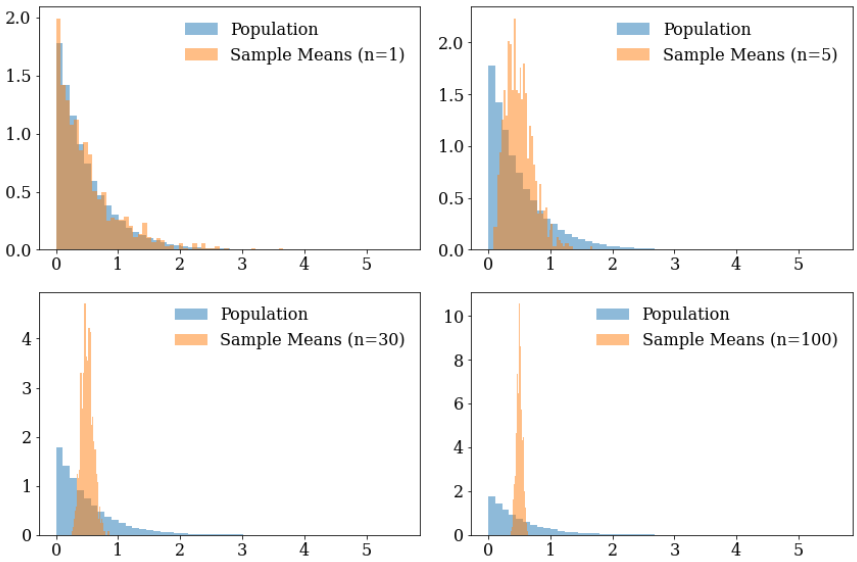

Figure 4: The plots show the population and the sample mean distributions for each sample size (1, 5, 30, and 100). As the sample size increases, the distribution of the sample means approaches a normal distribution following the Central Limit Theorem.

1. **Sample**: A sample is a subset of data points selected from a larger population. It is meant to represent the population and is used for analysis, making inferences, and drawing conclusions about its characteristics.
2. **Population**: A population is the complete set of data points that you want to make inferences about. It is usually too large to analyze directly.
3. **Parameter**: A parameter is a numerical value that describes a population. For Example, the mean and standard deviation of a population are parameters.

Central Limit Theorem: In simple terms, it states that when you take a large number of random samples from any population, the average (or sum) of these samples will have a normal distribution, no matter what the original population distribution looks like. Figure-4 shows the distribution of the sample means approaches a normal distribution with increasing the sample size.

Mathematically, let $X_1, X_2, \ldots, X_n$ be a random sample of size n from a distribution with mean μ and variance σ^2. The CLT states that as n approaches infinity, the distribution of the sample mean, $\bar{X}$, approaches a normal distribution with mean μ and variance $\frac{\sigma^2}{n}$:

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$$