Bayes Theorem

Definition:

Bayes Theorem is a method in probability theory used to calculate the probability of a hypothesis based on prior knowledge and new evidence. It provides a way to update the probability of a hypothesis as more data becomes available. Bayes Theorem is widely used in fields like machine learning, statistics, medicine, and risk analysis.

Formula:

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

Where:

- **P(A|B)**: Posterior Probability — Probability of A given B.
- **P(B|A)**: Likelihood — Probability of observing B when A is true.
- **P(A)**: Prior Probability — Initial probability of A before observing B.
- **P(B)**: Marginal Probability — Total probability of B across all possible scenarios.

Importance:

Bayes Theorem allows us to incorporate prior knowledge with observed data to refine predictions. It provides a foundation for probabilistic reasoning and decision-making.

Key Insights:

1. **Update Beliefs**: Prior beliefs (P(A)) are updated using new evidence (P(B|A)) to produce a revised belief (P(A|B)).
2. **Relationship Between Events**: Bayes Theorem connects conditional probabilities, making it useful in scenarios involving dependencies between events.
3. **Normalization**: The denominator (P(B)) ensures the probabilities sum to 1 by considering all possible causes of B.

Applications:

- **Medical Diagnosis**: Estimating the probability of a disease given symptoms and test results.
- **Spam Filtering**: Determining whether an email is spam based on its content.
- **Risk Assessment**: Evaluating the likelihood of financial or operational risks.
- **Machine Learning**: Used in Bayesian models and Naive Bayes classifiers.
- **Forensic Analysis**: Assessing evidence probabilities in criminal investigations.

Advantages:

- Facilitates probabilistic reasoning in uncertain situations.
- Easy to compute and apply in various real-world scenarios.
- Helps incorporate both new evidence and prior knowledge.

Limitations:

- Requires accurate prior probabilities and likelihood estimates.
- May not work well with large numbers of interdependent variables.

Example:

In a medical context:

- **1%** of people have a rare disease (P(D)=0.1).
- A diagnostic test is **95% accurate for positive cases** (P(T|D)=0.95) and gives **5% false positives** (P(T|¬D)=0.05).
- If a person tests positive (TT), what is the probability they actually have the disease (P(D|T))?

Using Bayes Theorem:

$$P(D|T) = \frac{P(T|D) \cdot P(D)}{P(T)}$$

Where $P(T) = P(T|D) \cdot P(D) + P(T|\neg D) \cdot P(\neg D)$.

Substituting values:

$$P(D|T) = \frac{(0.95)(0.01)}{(0.95)(0.01) + (0.05)(0.99)} \approx 0.16$$

Naive Bayes

Definition:

Naive Bayes is a supervised learning algorithm based on Bayes Theorem. It is called "naive" because it assumes all features are independent of each other, which simplifies computations. Despite this unrealistic assumption, Naive Bayes performs remarkably well in many real-world tasks, especially for text classification and spam detection.

Formula:

For a class C and features $x_1, x_2, \ldots, x_n$:

$$P(C|x_1, x_2, \ldots, x_n) = \frac{P(C) \cdot P(x_1|C) \cdot P(x_2|C) \cdot \ldots \cdot P(x_n|C)}{P(x_1, x_2, \ldots, x_n)}$$

- The denominator is constant for all classes, so it is ignored when finding the most probable class.

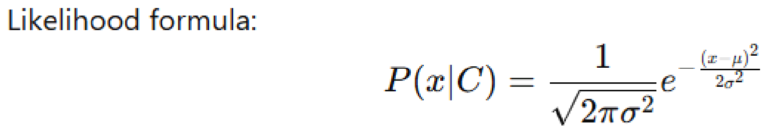
Steps in Classification:

1. **Calculate Prior Probabilities (P(C))**:
   - Determine the proportion of each class in the training data.
2. **Calculate Likelihood (P(xi|C)**:
   - For each feature given a class, estimate its probability using the frequency or distribution (e.g., Gaussian for continuous data).
3. **Compute Posterior Probabilities**:
   - Multiply the prior and likelihoods for each class.
4. **Choose the Class**:
   - Select the class with the highest posterior probability as the predicted class.

Assumptions:

- All features are conditionally independent given the class label.
- Each feature contributes equally to the outcome.

Types of Naive Bayes:

1. **Gaussian Naive Bayes**:
   - Used for continuous data. Assumes features follow a normal distribution.
   - 

     Likelihood formula:

     $$P(x|C) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$
   - Parameters: Mean (μ\mu) and variance (σ2\sigma^2) of the data.
2. **Multinomial Naive Bayes**:
   - Suitable for discrete data, such as word counts in documents.
   - Used in text classification, spam filtering, and sentiment analysis.
3. **Bernoulli Naive Bayes**:
   - Designed for binary features (presence/absence).
   - Commonly applied in text analysis to check the presence of specific words.

Advantages:

- Simple and computationally efficient.
- Works well with large datasets and high-dimensional data.
- Requires less training data compared to complex models.
- Effective for real-time predictions.

Disadvantages:

- Relies on the independence assumption, which is rarely true.
- May perform poorly with highly correlated features.
- Sensitive to imbalanced data.

Applications:

- **Text Classification**: Classifying documents, emails, and news articles.
- **Spam Filtering**: Detecting spam emails based on word usage.
- **Sentiment Analysis**: Analyzing user opinions or reviews.
- **Medical Diagnosis**: Predicting diseases from symptoms and test results.

Example:

Suppose you are classifying emails into "spam" or "not spam". Words like "Buy", "Free", and "Win" are considered features.

- Calculate the prior probabilities (P(spam) and P(not spam)).
- Estimate the likelihoods (P(word|spam)) for each word in the email.
- Compute the posterior probabilities and classify the email based on the class with the higher posterior probability.

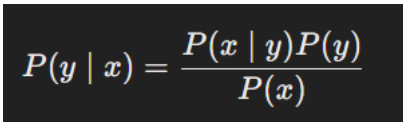
Bayes Optimal Classifier

The *Bayes Optimal Classifier* is a theoretical model that provides the most accurate classification of a new instance based on the training data. It operates under the principles of Bayes' theorem, calculating the conditional probabilities of different outcomes and selecting the one with the highest probability. This classifier is often referred to as the Bayes optimal learner, and it serves as a benchmark for evaluating the performance of other classifiers in machine learning.

(The **Bayes Optimal Classifier** is the best way to classify data because it calculates the most probable class based on both prior knowledge (probability of each class) and the observed data)

Step1-Bayes' Theorem for Classification

Bayes' Theorem allows us to compute the **posterior probability** P(y|x) of class y given the observed data x:



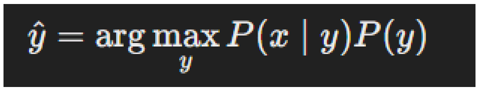
$$P(y \mid x) = \frac{P(x \mid y)P(y)}{P(x)}$$

Where:

- P(y|x) is the **posterior probability**: the probability that xxx belongs to class yyy.
- P(x|y) is the **likelihood**: the probability of observing xxx given class yyy.
- P(y) is the **prior probability** of class yyy.
- P(x) is the **marginal likelihood** (also called evidence): the overall probability of observing xxx across all classes.

Step2-Decision Rule for Bayes Optimal Classifier

The Bayes Optimal Classifier assigns the class y that maximizes the posterior probability:



$$\hat{y} = \arg\max_{y} P(x \mid y)P(y)$$

(argmax: The "argument of the maximum" means the value that maximizes the function)

Advantages

- Theoretical Foundation: The Bayes Optimal Classifier is grounded in solid statistical principles, making it a reliable benchmark for classification tasks.
- Optimal Performance: It provides the best possible classification accuracy under the given conditions, outperforming other classifiers on average

Applications

The Bayes Optimal Classifier is widely used in various fields, including:

- **Medical Diagnosis**: It helps in predicting diseases based on symptoms and patient data.
- **Spam Detection**: Used to classify emails as spam or non-spam based on content features.
- **Image Recognition**: Assists in identifying objects within images by analysing pixel data.

Limitations

- **Computational Complexity**: The computation of posterior probabilities can be expensive, especially with large datasets and complex hypothesis spaces.
- **Intractability**: In many practical scenarios, calculating the Bayes Optimal Classifier can be intractable due to the high dimensionality of the data and the number of hypotheses

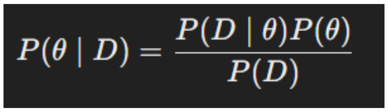
Bayesian Inference

Definition:

- Bayesian Inference is a method of statistical inference in which we update our belief about the parameters (expressed as a posterior distribution) using prior knowledge and the observed data. It relies on Bayes' Theorem.

Mathematical Formulation:

- Bayes' Theorem gives the posterior distribution:



$$P(\theta \mid D) = \frac{P(D \mid \theta)P(\theta)}{P(D)}$$

Where:

  - P(θ|D) is the posterior distribution of the parameters given the data.
  - P(D|θ) is the likelihood of the data given the parameters.
  - P(θ)is the prior distribution of the parameters.
  - P(D) is the marginal likelihood or evidence, which ensures that the posterior is normalized.

- **Posterior Distribution**: It represents our updated belief about the parameter θ after observing the data.
- **Prior Distribution**: Encodes our belief about θ before observing the data.

- **Likelihood**: Represents the probability of the data given the parameters.
- **Marginal Likelihood**: Ensures that the posterior is a valid probability distribution.

Example: Suppose we have prior knowledge about a coin's bias (the probability of heads), and we observe a series of coin flips. Bayesian inference would allow us to update our belief about the bias of the coin after seeing the flips.

Types of Inference- Maximum A Posteriori (MAP) Estimation:

- Maximum Likelihood Estimation (MLE)

Maximum Likelihood Estimation (MLE) vs Maximum A Posteriori Estimation (MAP)

Maximum Likelihood Estimation (MLE) and **Maximum A Posteriori Estimation (MAP)** are both methods used in statistical estimation to infer the parameters of a model based on observed data. However, they differ in how they treat the uncertainty of model parameters and the type of prior knowledge they incorporate. Below are detailed notes on both techniques:

---

1. Maximum Likelihood Estimation (MLE)

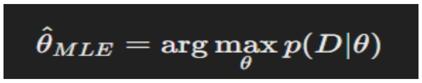
Overview:

- **Goal:** MLE aims to find the parameter values that maximize the likelihood of observing the given data.
- **Assumption:** MLE only considers the likelihood of the data and does **not** incorporate any prior knowledge about the parameters.
- **Objective:** Find the parameter θ that maximizes the likelihood function L(θ).

Mathematical Formulation:

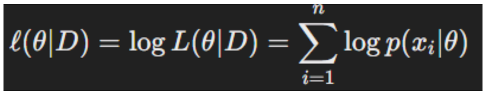
Given a dataset D={x1,x2,……,xn} and a likelihood function L(θ|D)=p(D|θ) (probability of data given the parameters θ

- The MLE is defined as:



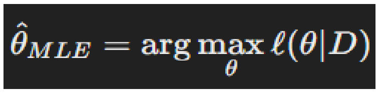
$$\hat{\theta}_{MLE} = \arg\max_{\theta} p(D|\theta)$$

- **Log-Likelihood:** In practice, it's more convenient to work with the log of the likelihood function (log-likelihood):



$$\ell(\theta|D) = \log L(\theta|D) = \sum_{i=1}^{n} \log p(x_i|\theta)$$

MLE seeks to maximize this log-likelihood function:



$$\hat{\theta}_{MLE} = \arg\max_{\theta} \ell(\theta|D)$$

Characteristics of MLE:

- **No Prior Information:** MLE does not incorporate any prior distribution over the parameters. It relies purely on the observed data.
- **Frequentist Approach:** MLE is a frequentist method, meaning it treats the parameters as fixed but unknown quantities.
- **Asymptotic Consistency:** As the sample size increases, MLE tends to converge to the true parameter values.
- **Example:** Estimating the mean of a normal distribution based on observed data.

Advantages of MLE:

- **Simple to Implement:** MLE only requires the likelihood function and the data.
- **Asymptotically Efficient:** MLE is efficient in large samples (i.e., it tends to give the true value as the sample size grows).

Disadvantages of MLE:

- **Overfitting Risk:** MLE can overfit when the sample size is small, particularly if the model is too complex.
- **No Prior Incorporation:** It does not account for prior knowledge, which can be limiting in some cases.

---

2. Maximum A Posteriori Estimation (MAP)

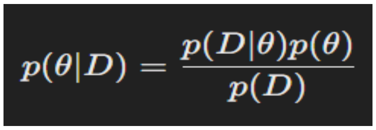
Overview:

- **Goal:** MAP estimation seeks to find the parameter value that maximizes the posterior distribution, combining both the likelihood and a prior belief about the parameters.
- **Assumption:** MAP incorporates **prior knowledge** or beliefs about the parameters, usually in the form of a prior distribution p(θ).

- **Objective:** Find the parameter θ that maximizes the posterior distribution p(θ|D) , which is the probability of the parameters given the data.

Mathematical Formulation:

Using Bayes' Theorem, the posterior distribution is given by:

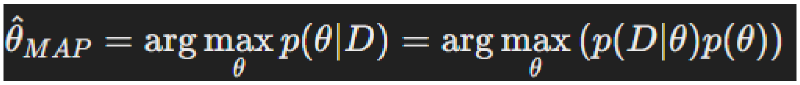


$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)}$$

where:

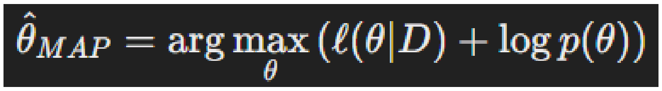
- p(D|θ) is the **likelihood** (same as in MLE).
- p(θ) is the **prior** distribution of θ.
- p(D) is the marginal likelihood or evidence (normalizing constant).

The MAP estimate θ^MAP is defined as:



$$\hat{\theta}_{MAP} = \arg\max_{\theta} p(\theta|D) = \arg\max_{\theta} \left(p(D|\theta)p(\theta)\right)$$

Since p(D) does not depend on θ, we can focus on the **posterior**'s unnormalized form:



$$\hat{\theta}_{MAP} = \arg\max_{\theta} \left(\ell(\theta|D) + \log p(\theta)\right)$$

Characteristics of MAP:

- **Prior Information:** MAP incorporates a **prior** distribution p(θ), representing prior knowledge or beliefs about the parameter values.
- **Bayesian Approach:** MAP is a Bayesian method, meaning it treats the parameters as random variables with distributions, rather than fixed quantities.
- **Trade-off Between Likelihood and Prior:** The relative influence of the data and the prior depends on the strength of the prior. If the prior is strong, it can dominate the posterior.

Advantages of MAP:

- **Incorporates Prior Knowledge:** MAP is more flexible than MLE because it allows the inclusion of prior knowledge about the parameters, which can guide the estimation, especially in small data regimes.
- **Regularization:** In some cases, the prior can act as a form of regularization, helping to prevent overfitting.

Disadvantages of MAP:

- **Choice of Prior:** The result of MAP estimation heavily depends on the choice of the prior distribution. If the prior is incorrect, it can bias the results.
- **Computational Complexity:** In some cases, computing the posterior may be challenging, especially when the prior is complex or non-conjugate.

Comparison Between MLE and MAP:

Aspect	**MLE**	**MAP**
Focus	Maximizes likelihood.	Maximizes posterior (likelihood + prior).
Prior Information	Does not use prior information.	Uses prior information in the form of p(θ).
Assumption	Frequentist approach (parameters are fixed but unknown).	Bayesian approach (parameters are random variables with a prior distribution).
Risk of Overfitting	Prone to overfitting in small data.	Less prone to overfitting due to regularization from prior.
Asymptotic Behaviour	As sample size increases, MLE is consistent and asymptotically efficient.	As sample size increases, MAP converges to MLE if the prior is non-informative.
When to Use	When there is no prior knowledge or the sample size is large.	When prior knowledge is available or when regularization is needed.

□ **MLE**: Since MLE treats parameters as **fixed but unknown** values, it doesn't incorporate any uncertainty about the parameters. The estimation process is solely data-driven, maximizing the likelihood.

□ **MAP**: MAP, on the other hand, treats parameters as **random variables** and incorporates uncertainty through the prior distribution. The result is a distribution over the parameters, not just a point estimate, which represents uncertainty about the true parameter value.