Andhra Pradesh State Skill Development Corporation

Skill AP
APSSDC

Skill AP
Learn Anytime Anywhere
Andhra Pradesh State Skill Development Corporation

Machine Learning

Bayesian Concept Learning

CHAPTER 11

Bayesian Concept Learning

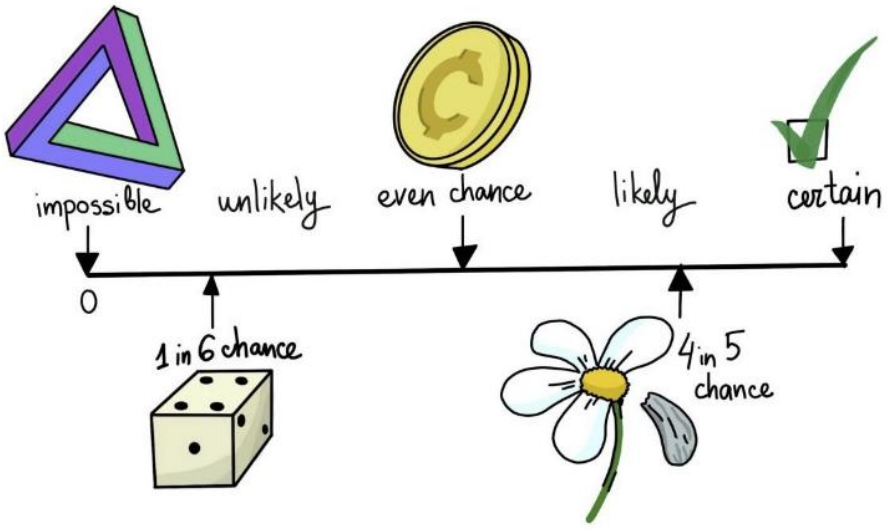
Understanding Probability and Bayes Theorem

Machine learning is about developing predictive models from uncertain data. Uncertainty means working with imperfect or incomplete information. As it is fair to say that probability is required to effectively work through a machine learning predictive modeling project. Uncertainty involves making decisions with incomplete information, and this is the way we generally operate in the world Handling uncertainty is typically described using everyday words like chance, luck, and risk.

Probability is a field of mathematics that gives us the language and tools to quantify the uncertainty of events and reason in a principled manner. We can assign and quantify the likelihood of things we care about, such as outcomes, events, or numerical values.
There are three main sources of uncertainty in machine learning, they are: noisy data, incomplete coverage of the problem domain and imperfect models.

Nevertheless, we can manage uncertainty using the tools of probability.

As Machine Learning practitioners, we must have an understanding of probability in order to manage the uncertainty we see in each project.
As Probability is a field of mathematics that quantifies uncertainty. It is undeniably a pillar of the field of machine learning, as probability makes more sense to a practitioner once they have the context of the applied Machine Learning process in which to interpret it.



Classification predictive modeling problems are those where an example is assigned a given label. Framing the problem as a prediction of class membership simplifies the modeling problem and makes it easier for a model to learn. It allows the model to capture ambiguity in the data, which allows a process downstream, such as the user to interpret the probabilities in the context of the domain. ambiguity in the data, which allows a process downstream, such as the user to interpret the probabilities in the context of the domain.
Instead of predicting class values directly for a classification problem, it can be convenient to predict the probability of an observation belonging to each possible class.

Predicting probabilities allows some flexibility including deciding how to interpret the probabilities, presenting predictions with uncertainty, and providing more nuanced ways to evaluate the skill of the model.

Predicted probabilities that match the expected distribution of probabilities for each class are referred to as calibrated. The problem is, not all machine learning models are capable of predicting calibrated probabilities. This choice of a class membership framing of the problem interpretation of the predictions made by the model requires a basic understanding of probability.

There are algorithms that are specifically designed to harness the tools and methods from probability. These range from individual algorithms, like **Naive Bayes algorithm**, which is constructed using **Bayes Theorem** with some simplifying assumptions. As Bayes Theorem provides a principled way for calculating a conditional probability Let us recall the Probability concepts.

The conditional probability is the probability of one event given the occurrence of another event, often described in terms of events A and B from two dependent random variables e.g. X and Y.

- **Conditional Probability**: Probability of one (or more) event given the occurrence of another event, e.g. P(A given B) or P(A | B).

The joint probability can be calculated using the conditional probability; for example:

- **P(A, B) = P(A | B) * P(B)**

This is called the product rule. Importantly, the joint probability is symmetrical, meaning that:

- **P(A, B) = P(B, A)**

The conditional probability can be calculated using the joint probability; for example:

- **P(A | B) = P(A, B) / P(B)**

The conditional probability is not symmetrical; for example:

- **P(A | B) != P(B | A)**

Specifically, one conditional probability can be calculated using the other conditional probability; for example:

- **P(A|B) = P(B|A) * P(A) / P(B)**

The reverse is also true; for example:

- **P(B|A) = P(A|B) * P(B) / P(A)**

This alternate calculation of the conditional probability is referred to as Bayes Rule or Bayes Theorem, named for **Reverend Thomas Bayes**, who is credited with first describing it. It is grammatically correct to refer to it as **Bayes' Theorem** (with the apostrophe), but it is common to omit the apostrophe for simplicity.

Bayes' theorem, named after 18th-century British mathematician Thomas Bayes, is a mathematical formula for determining Conditional Probability. Conditional Probability is the likelihood of an outcome occurring, based on a previous outcome occurring. Bayes' theorem provides a way to revise existing predictions or theories (update probabilities) given new or additional evidence. In finance, Bayes' theorem can be used to rate the risk of lending money to potential borrowers.

Applications of the theorem are widespread and not limited to the financial realm. As an example, Bayes' theorem can be used to determine the accuracy of medical test results by taking into consideration how likely any given person is to have a disease and the general accuracy of the test. Bayes' theorem relies on incorporating **Prior Probability** distributions in order to generate **Posterior Probabilities**. Prior probability, in Bayesian statistical inference, is the probability of an event before new data is collected. This is the best rational assessment of the probability of an outcome based on the current knowledge before an experiment is performed. **Posterior probability** is the revised probability of an event occurring after taking into consideration new information. Posterior probability is calculated by updating the prior probability by using Bayes' theorem. In statistical terms, the posterior probability is the probability of event A occurring given that event B has occurred.

Formula For Bayes' Theorem

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

where:

$P(A) =$ The probability of A occurring

$P(B) =$ The probability of B occurring

$P(A|B) =$ The probability of A given B

$P(B|A) =$ The probability of B given A

$P\left(A \bigcap B\right)) =$ The probability of both A and B occurring

The terms in the Bayes Theorem equation are given names depending on the context where the equation is used.

It can be helpful to think about the calculation from these different perspectives and help to map your problem onto the equation.

Firstly, in general, the result P(A|B) is referred to as the **Posterior Probability** and P(A) is referred to as the **Prior Probability**.

- **P(A|B): Posterior probability.**
- **P(A): Prior probability.**

Sometimes P(B|A) is referred to as the **Likelihood** and P(B) is referred to as the **Evidence**.

- **P(B|A): Likelihood.**
- **P(B): Evidence.**

This allows Bayes Theorem to be restated as:

- **Posterior = Likelihood * Prior / Evidence**

This set of rules of probability allows one to update their predictions of events occurring based on new information that has been received, making for better and more dynamic estimates.

Introduction to Naive Bayes Classifier:

Consider a case where you have created features, you know about the importance of features and you are supposed to make a classification model that is to be presented in a very short period of time?

What will you do? You have a very large volume of data points and very less few features in your data set. In that situation if I had to make such a model I would have used '*Naive Bayes*', that is considered to be a really fast algorithm when it comes for classification tasks.

Naive Bayes is a machine learning model that is used for large volumes of data, even if you are working with data that has millions of data records the recommended approach is Naive Bayes. It gives very good results when it comes to NLP tasks such as sentimental analysis. It is a fast and uncomplicated classification algorithm.

To understand the naive Bayes classifier we need to understand the Bayes theorem,as we have already discussed about it

Bayes Theorem

It is a theorem that works on conditional probability. Conditional probability is the probability that something will happen, given that something else has already occurred. The conditional probability can give us the probability of an event using its prior knowledge.

Naive Bayes Classifier

- It is a kind of classifier that works on Bayes theorem.
- Prediction of membership probabilities is made for every class such as the probability of data points associated to a particular class.
- The class having maximum probability is appraised as the most suitable class.
- This is also referred as Maximum A Posteriori (MAP).
- The MAP for a hypothesis is:
  - $MAP\ (H) = \max P((H|E))$
  - $MAP\ (H) = \max P((H|E)\ *(P(H))\ /P(E))$
  - $MAP\ (H) = \max(P(E|H) * P(H))$
  - $P\ (E)$ is evidence probability, and it is used to normalize the result. The result will not be affected by removing $P(E)$.
- NB classifiers conclude that all the variables or features are not related to each other.The Existence or absence of a variable does not impact the existence or absence of any other variable.

Types Of Naive Bayes Algorithms

1.**Gaussian Naïve Bayes:** When characteristic values are continuous in nature then an assumption is made that the values linked with each class are dispersed according to Gaussian that is Normal Distribution.

2. **Multinomial Naïve Bayes:** Multinomial Naive Bayes is favored to use on data that is multinomial distributed. It is widely used in text classification in NLP. Each event in text classification constitutes the presence of a word in a document.

3.**Bernoulli Naïve Bayes:** When data is dispensed according to the multivariate Bernoulli distributions then Bernoulli Naive Bayes is used. That means there exist multiple features but each one is assumed to contain a binary value. So, it requires features to be binary-valued.

Advantages And Disadvantages Of Naive Bayes

Advantages:

- It is a highly extensible algorithm which is very fast.
- It can be used for both binaries as well as multiclass classification.
- It has mainly three different types of algorithms that are GaussianNB, MultinomialNB, BernoulliNB.
- It is a famous algorithm for spam email classification.
- It can be easily trained on small datasets and can be used for large volumes of data as well.

Disadvantages:

- The main disadvantage of the NB is considering all the variables independent that contributes to the probability.

Applications of Naive Bayes Algorithms

- **Real-time Prediction:** Being a fast learning algorithm can be used to make predictions in real-time as well.
- **MultiClass Classification:** It can be used for multi-class classification problems also.
- **Text Classification:** As it has shown good results in predicting multi-class classification so it has more success rates compared to all other algorithms. As a result, it is majorly used in sentiment analysis & spam detection.

Implementation of Naive Bayes - Data Collection

Before implementing the Gaussian Naive Bayes classifier we should note two simple assumptions:

- Our data is normally distributed
- We expect our data columns to be conditionally independent of each other

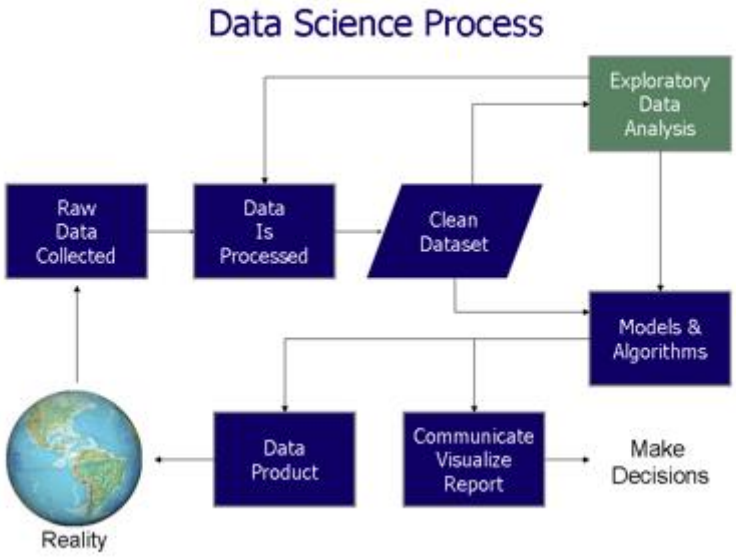
We will be working on the heart disease dataset as it contains 76 attributes, but all published experiments refer to using a subset of 14 of them. In particular, the Cleveland database is the only one that has been used by ML researchers to this date. The "goal" field refers to the presence of heart disease in the patient. It is integer valued from 0 (no presence) to 4.The names and social security numbers of the patients were recently removed from the database, replaced with dummy values.

Data collection is the process of gathering and measuring information from countless different sources. In order to use the data we collect to develop practical artificial intelligence (AI) and machine learning solutions, it must be collected and stored in a way that makes sense for the business problem at hand.

PREPROCESSING Preprocessing can be done by several methods. Real world data are generally inconsistent, noisy and incomplete data. In this system filtering method is used for data reduction (removing the repeated data), mean method is used for incomplete data (i.e., missing values) and also removes the inconsistent data (impossible data combination).

EDA and Modelling

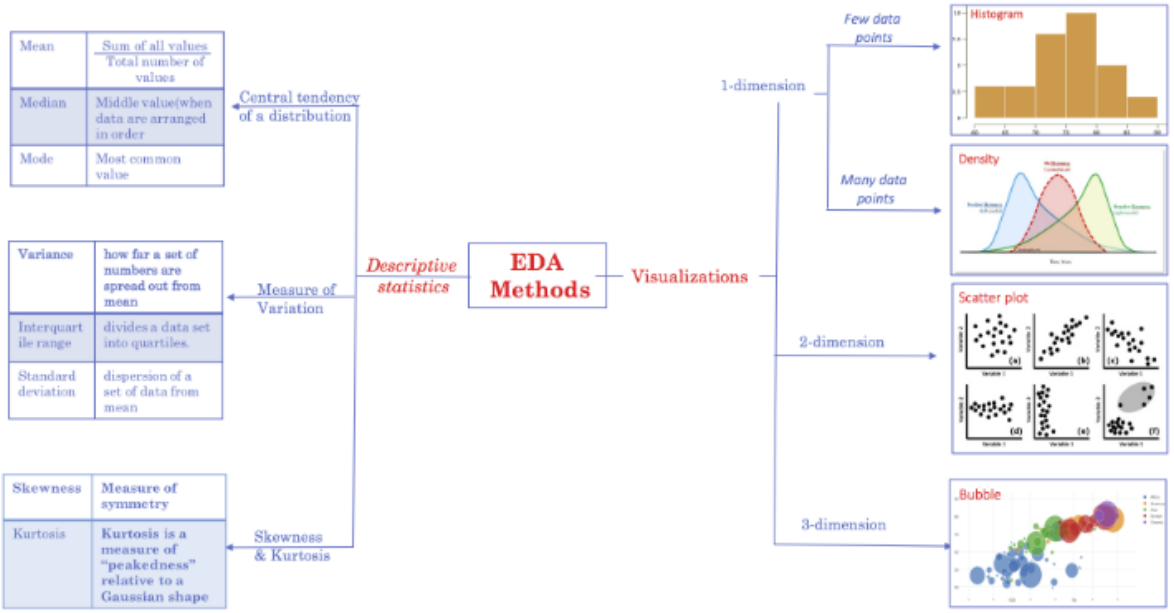
The phrase exploratory data analysis was coined by John W. Tukey in 1977. What, is interesting is that even though he coined the phrase, it is difficult, if not impossible, to find a precise definition of EDA in Tukey's writings, partly because he preferred working with vague concepts, things that could be made precise in several ways.



A good data exploration is a preamble to the formal analysis. It allows the data scientist to:

- Verify expected relationships actually exist in the data, thus formulating and validating planned techniques of analysis.
- To find some unexpected structure in the data that must be taken into account, thereby suggesting some changes in the planned analysis.
- Deliver data-driven insights to business stakeholders by confirming they are asking the right questions and not biasing the investigation with their assumptions.
- Provide the context around the problem to make sure the potential value of the data scientist's output can be maximized.

EDA is generally cross-classified in two ways. First, each method is either non-graphical or graphical. And second, each method is either univariate or multivariate (usually just bivariate). Non-graphical methods generally involve calculation of summary statistics, while graphical methods obviously summarize the data in a diagrammatic or pictorial way.So we will visualize the data and understand the relationship between the features.



Then we can evaluate our results by performing performance metrics which includes Classification Report,Confusion Matrix and so on.

Multinomial Naive Bayes Implementation

With an ever-growing amount of textual information stored in electronic form such as legal documents, policies, company strategies, etc., Automatic Text Classification is becoming increasingly important. This requires a supervised learning technique that classifies every new document by assigning one or more class labels from a fixed or predefined class. The **Multinomial Naive Bayes** classifier is suitable for classification with discrete features (e.g., word counts for text classification). The **Multinomial** distribution normally requires integer feature counts.\
Text can be divided into different primitives:

- Document
- Sentences
- Words
- Characters

Document is a large collection of text,it contains sentences.Each sentence is formed with a collection of words and each word is a group of characters.
Feature extraction and Selection are the most important sub-tasks in pattern classification.The three main criteria of good features are:

- Salient: The features should be meaningful and important to the problem
- Invariant: The features are resistant to scaling, distortion and orientation etc.
- Discriminatory: For training of classifiers, the features should have enough information to distinguish between patterns.

Bag of words is a commonly used model in Natural Language Processing. The idea behind this model is the creation of vocabulary that contains the collection of different words, and each word is associated with a count of how it occurs. Later, the vocabulary is used to create d-dimensional feature vectors.

For Example:

D1: Each country has its own constitution
D2: Every country has its own uniqueness
Vocabulary could be written as:
V= {each : 1, state : 1, has : 2, its : 2, own : 2, constitution : 1, every : 1, country : 2,}

Tokenization

It is the process of breaking down the text corpus into individual elements. These individual elements act as an input to Machine learning algorithms.

For Example:

Every country has its own uniqueness

every	country	has	its	own	uniqueness

Stop Words

Stop Words also known as un-informative words such as (so, and, or, the) should be removed from the document.

Stemming and Lemmatization

Stemming and Lemmatization are the process of transforming a word into its root form and aims to obtain the grammatically correct forms of words.
The above-mentioned process comes under the Bag of Words Model. Multinomial Naïve Bayes uses term frequency i.e. the number of times a given term appears in a document. Term frequency is often normalized by dividing the raw term frequency by the document length. After normalization, term frequency can be used to compute maximum likelihood estimates based on the training data to estimate the conditional probability.
TFIDF(TermFreq,Inverse Doc Freq) : It is the most common method in text handling.It is based on facts that important words are those which comes frequently in related document but not comes in every document.By applying this method each doc/text get a score vector of words.Higher score of any word means word is important for that document.

Tf-idf term weighting:

In a large text corpus, some words will be very present (e.g. “the”, “a”, “is” in English) hence carrying very little meaningful information about the actual contents of the document. If we were to feed the direct count data directly to a classifier those very frequent terms would shadow the frequencies of rarer yet more interesting terms.

In order to re-weight the count features into floating point values suitable for usage by a classifier it is very common to use the tf–idf transform.

Tf means term-frequency while tf–idf means term-frequency times inverse document-frequency

$$\textbf{tf-idf(t,d) = tf(t,d)*idf(t)}$$

Multinomial Naïve Bayes consider a feature vector where a given term represents the number of times it appears or very often i.e. frequency. On the other hand, Bernoulli is a binary algorithm used when the feature is present or not. At last Gaussian is based on continuous distribution.

Advantages:

- Low computation cost.
- It can effectively work with large datasets.
- For small sample sizes, Naive Bayes can outperform the most powerful alternatives.
- Easy to implement, fast and accurate method of prediction.
- Can work with multiclass prediction problems.
- It performs well in text classification problems.

Disadvantages:

It is very difficult to get the set of independent predictors for developing a model using Naive Bayes.

Applications:

- Naive Bayes classifier is used in Text Classification, Spam filtering and Sentiment Analysis. It has a higher success rate than other algorithms.
- Naïve Bayes along with Collaborative filtering are used in Recommended Systems.
- It is also used in disease prediction based on health parameters.
- This algorithm has also found its application in Face recognition.
- Naive Bayes is used in prediction of weather reports based on atmospheric conditions (temp, wind, clouds, humidity etc.)

We will use the sparse word count features from the 20 Newsgroups corpus to show how we might classify these short documents into categories.The 20 newsgroups dataset comprises around 18000 newsgroups posts on 20 topics split in two subsets: one for training (or development) and the other one for testing (or for performance evaluation). The split between the train and test set is based upon messages posted before and after a specific date.

This module contains two loaders. The first one, **sklearn.datasets.fetch_20newsgroups**, returns a list of the raw texts that can be fed to text feature extractors such as **sklearn.feature_extraction.text.CountVectorizer** with custom parameters so as to extract feature vectors.

The second one,**sklearn.datasets.fetch_20 newsgroups_vectorized**, returns ready-to-use features, i.e., it is not necessary to use a feature extractor.So we choose our category and train the data accordingly to predict() the result