Naïve Bayes Classifier

Table of Contents

[Probabilistic Classification 2](#_Toc106495432)

[Decision Theory 2](#_Toc106495433)

[Bayes Theorem 3](#_Toc106495434)

[Types of Models 4](#_Toc106495435)

[Generative Models 4](#_Toc106495436)

[Discriminative Models 4](#_Toc106495437)

[Discriminant Functions 4](#_Toc106495438)

[Graphical Models 5](#_Toc106495439)

[Naïve Bayes Classifiers 5](#_Toc106495440)

[Applications 7](#_Toc106495441)

[Advantages and Disadvantages 8](#_Toc106495442)

## Probabilistic Classification

Generally, with neural networks, we take some input and get some output . The output is an **array of values**, with each value corresponding to a **different class**. The values tell us the **probability** of that class being the correct one. Thus, this is called **probabilistic classification**. The models being used in these classification problems are called **probabilistic models**.

Formally, for each class , we find . We assign the class which has the highest probability, so .

A probabilistic classifier is a classifier that, given an observation as an input, can predict the probability distribution over a set of classes, instead of just outputting the most likely class that the observation should belong to.

## Decision Theory

Probability theory gives us a consistent framework for quantifying and manipulating uncertainty. **Decision Theory** is combined with probability theory to make optimal decisions in situations that involve uncertainty.

In decision theory, there are two parts, inference and decision. Suppose we have some input and a vector of classes, . The joint probability distribution gives a complete summary of the uncertainty associated with these variables, i.e. for each possible value of , what are the probabilities that each of the classes in are the correct ones.

The **inference step** involves determining from a set of training data. This is usually a very difficult problem. By comparison, the **decision step**, which for us is simply taking the class which has the highest probability for a given input , is an easy problem.

## Bayes Theorem

Consider that we obtain an x-ray image for a patient, and the output is whether or not the patient has a tumour, given by . According to **Bayes’ Theorem**:

Here, is the **prior probability**. For our example, the prior probability is the probability that the patient has a tumour before we have performed the x-ray. In contrast, is the **posterior probability**, the probability that the patient has cancer after we have performed the x-ray.

is the probability that the x ray we got happens to occur in the class . For example, the probability that a healthy lung x ray occurs in the cancer class is low. However, we do not know this value.

is the **normalization factor**. In our case, we will be comparing the values for different classes, i.e. comparing the probabilities of the patient having cancer and of the patient not having cancer. In both cases, the values will be the same (since it comes from the x ray), so we can ignore it.

## Types of Models

### Generative Models

A model which calculates using the Bayes theorem is called a **generative model**. Such a model must calculate the intermediate , which is difficult to calculate.

### Discriminative Models

By contrast, a model which calculates directly, without using the Bayes theorem, is called a **discriminative model**. Such a model does not need to calculate the difficult .

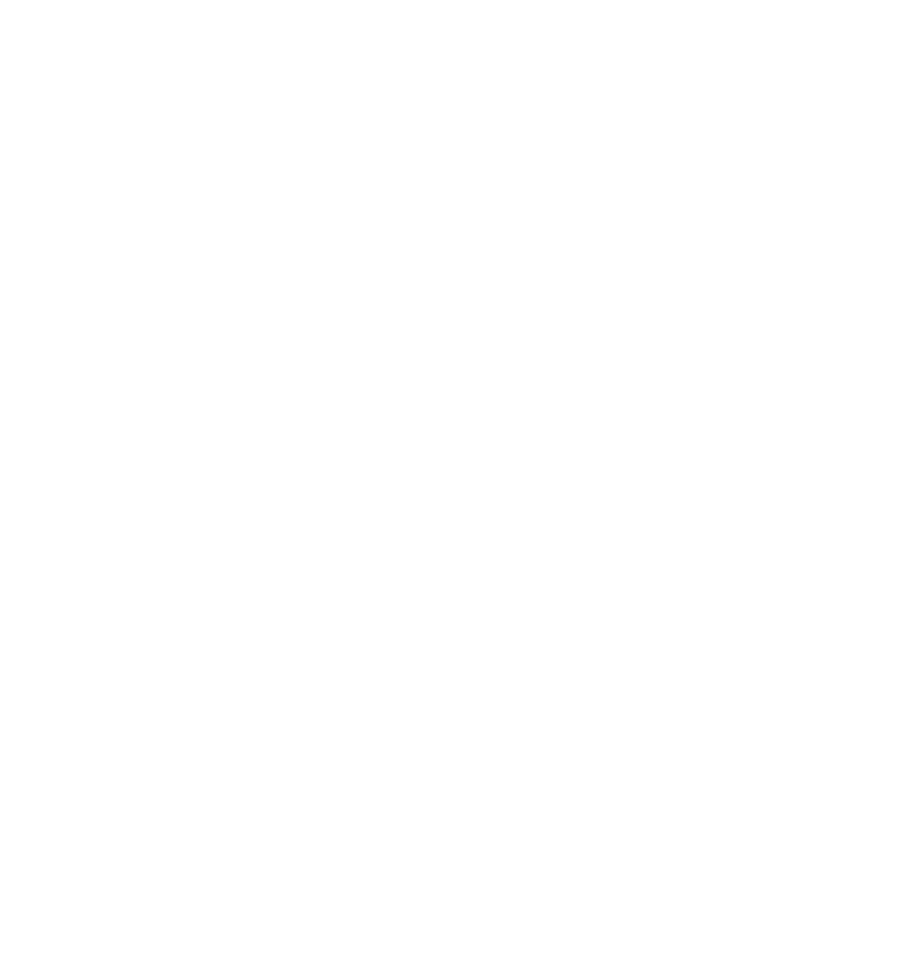
### Discriminant Functions

An alternative possibility would be to learn a function which can map directly into a decision. Such a function is called a **discriminant function**.

For example, the function , passed through a sigmoid function, can be directly mapped to a class based on whether the output is positive or negative.

### Graphical Models

**Graphical Models** are used to represent a neural network visually. They consist of a bunch of **nodes**, with each node corresponding to a random variable , having the value . An **arc** between two nodes, say and , indicates that has an influence on the value of , i.e. the value of is .



## Naïve Bayes Classifiers

A **Naïve Bayes Classifier** is a classifier based on the Bayes theorem. It is said to be naïve because it makes a single assumption, that **every feature is independent**. This makes our work a lot easier, in exchange for making our model less accurate.

For example, consider that we have some outcome which depends on the input features and . According to the **chain rule**,

If we assume that and are independent, we can say that , i.e. regardless of the value of , the value of is not affected.

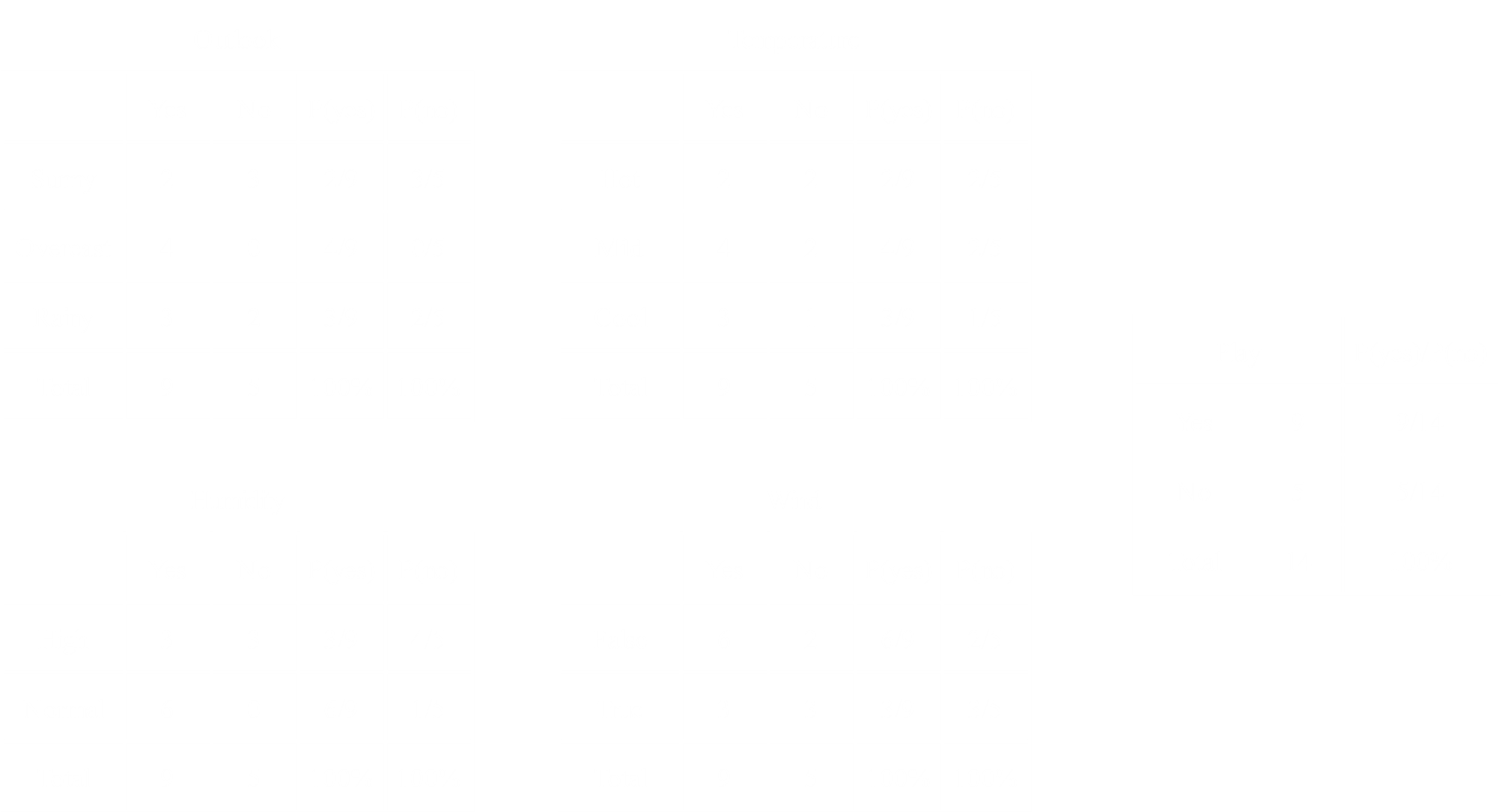
In our case, we want to predict some class from a set of input features , i.e. . From the Bayes rule, we saw

As mentioned previously, we can ignore the denominator. In the numerator, we mentioned that calculating is difficult. Using a Naïve Bayes Classified can solve this problem for us quite easily.

For possible classes, we want to find the class which gives the maximum value for , i.e.

Example

Suppose out of 14 days we go out to play for 9 days and we do not on 5 days. Whether or not we go out to play on a particular day is influenced by 4 factors, the weather outlook, the temperature, the humidity and the wind.



The probability that we will play (class ) given the set of features sunny outlook, hot temperature, high humidity and no wind is

### Applications

* Disease Prediction in real time since the process is very fast
* Multi-Class Prediction
* Text Classification, Spam Filtering and Sentiment Analysis, due to being able to deal with multiple classes well
* Recommendation Systems

### Advantages and Disadvantages

* Easy and fast predictions
* Performs well in multi-class problems
* If features are independent, performs better than other models while also needing less training data
* Performs well in case of categorical input compared to numerical variables
* Cannot make predictions for classes not seen in the training data
* Bad estimator, so probability outputs should not be taken seriously
* In real life, features are unlikely to be independent