I. Naive Bayes Classifier

Preparation Video:

https://www.youtube.com/watch?v=mlumJPFvooQ&list=PLZoTAELRMXVPkl7oRvzyNnyj1HS4wt2K-&index=2&ab_channel=KrishNaik

Theoretical Understanding:

1. Tutorial 48th : https://www.youtube.com/watch?v=jS1CKhALUBQ
2. Tutorial 49th: https://www.youtube.com/watch?v=temQ8mHpe3k

1. What Are the Basic Assumptions?

Features Are Independent

2. Advantages

1. Work Very well with many number of features

   Explanation: Because of the class independence assumption, naive Bayes classifiers can quickly learn to use high dimensional features with limited training data compared to more sophisticated methods. This can be useful in situations where the dataset is small compared to the number of features, such as images or texts. Naive Bayes implicitly treats all features as being independent of one another, and therefore the sorts of curse-of-dimensionality problems which typically rear their head when dealing with high-dimensional data do not apply.

   If your data has k dimensions, then a fully general ML algorithm which attempts to learn all possible correlations between these features has to deal with 2^k possible feature interactions, and therefore needs on the order of 2^k many data points to be performant. However because Naive Bayes assumes independence between features, it only needs on the order of k many data points, exponentially fewer.

   However this comes at the cost of only being able to capture much simpler mappings between the input variables and the output class, and as such Naive Bayes could never compete with something like a large neural network trained on a large dataset when it comes to tasks like image recognition, although it might perform better on *very* small datasets.

2. Works Well with Large training Dataset
3. It converges faster when we are training the model

   Explanation: Unlike other machine learning models, naive bayes require little to no training. When trying to make a prediction that involves multiple features, we simply use the maths by making the *naive* assumption that the features are independent.

4. It also performs well with categorical features

   Explanation: For a Naive Bayes classifier, categorical values are the easiest to deal with. All you are really after is P(Feature | Class). This should be easy for the days of the week. Compute P(Monday | Class=Yes) and so on.

3. Disadvantages

1. Correlated features affects performance

4. Whether Feature Scaling is required?

No. *In fact, any Algorithm which is NOT distance based, is not affected by Feature Scaling.* As Naive Bayes algorithm is based on probability not on distance, so it doesn't require feature scaling.

5. Impact of Missing Values?

Naive Bayes can handle missing data. Attributes are handled separately by the algorithm at both model construction time and prediction time. As such, if a data instance has a missing value for an attribute, it can be ignored while preparing the model, and ignored when a probability is calculated for a class value tutorial : https://www.youtube.com/watch?v=EqjyLfpv5oA

6. Impact of outliers?

It is usually robust to outliers.

One potential issue with outliers is that unseen observations can lead to 0 probabilities. And we know in naive bayes we multiply probab of words lying in that particular class and results zero. For example, Bernoulli Naive Bayes applied to word features will always produce 0 probabilities when it encounters a word that wasn't seen in the training data. Outliers in this sense can be a problem.

However, all these and similar issues of Naive Bayes have well-known solutions (like Laplace smoothing, i.e. adding an artificial count for every word) and are routinely implemented. In Gaussian Naive Bayes, outliers will affect the shape of the Gaussian distribution and have the usual effects on the mean etc. So depending on your use case, it still makes sense to remove outliers.

Different Problem statement you can solve using Naive Bayes

1. Sentiment Analysis
2. Spam classification
3. twitter sentiment analysis
4. document categorization