Bayesian Classifiers

Naive Bayes classifier

Naive Bayes classifiers are a family of simple probabilistic classifiers based on applying Bayes' theorem with *strong (naive) independence assumptions* between the features.

Probabilistic model

Abstractly, naive Bayes is a conditional probability model: given a problem instance to be classified, represented by a vector $\boldsymbol{x} = (x_1, \cdots, x_n)$ representing some n features (independent variables), it assigns to this instance probabilities

$$p\left(C_k|x_1, \cdots, x_n\right)$$

for each of k possible outcomes or classes C_k.

The problem with the above formulation is that if the number of features n is large or if a feature can take on a large number of values, then basing such a model on probability tables is infeasible. We therefore reformulate the model to make it more tractable. Using Bayes' theorem, the conditional probability can be decomposed as

$$p\left(C_k|x\right) = \frac{p\left(C_k\right) p\left(x|C_k\right)}{p\left(x\right)}$$

Here $p\left(x_i|x_{i+1}, \cdots, x_n, C_k\right) = p\left(x_i|C_k\right)$ and the joint model can be written following the chain rules as:

$$\begin{aligned} p\left(C_k|x_1, \cdots, x_n\right) &= p\left(x_1|x_2, \cdots, x_n, C_k\right) p\left(x_2|x_3, \cdots, x_n, C_k\right) \cdots p\left(x_n|C_k\right) p\left(C_k\right) \\ &= \frac{p\left(C_k\right)}{p\left(x\right)} \prod_{i=1}^{n} p\left(x_i|C_k\right) \end{aligned}$$

or

$$p\left(x|C_k\right) = \prod_{i=1}^{n} p\left(x_i|C_k\right)$$

Pros and Cons

- pros
  - It is easy and fast to predict class of test data set. It also perform well in multi class prediction
  - When assumption of independence holds, a Naive Bayes classifier performs better compare to other models like logistic regression and you need less training data.
  - It perform well in case of categorical input variables compared to numerical variable(s). For numerical variable, normal distribution is assumed (bell curve, which is a strong assumption).

- cons
  - If categorical variable has a category (in test data set), which was not observed in training data set, then model will assign a 0 (zero) probability and will be unable to make a prediction. This is often known as "Zero Frequency". To solve this, we can use the smoothing technique. One of the simplest smoothing techniques is called Laplace estimation.
  - On the other side naive Bayes is also known as a bad estimator, so the probability outputs from predict_proba are not to be taken too seriously.
  - Another limitation of Naive Bayes is the assumption of independent predictors. In real life, it is almost impossible that we get a set of predictors which are completely independent.