Professional Certificate in Machine Learning and Artificial Intelligence

Module 11: Naïve Bayes

Learning outcomes:

- Use Bayes' theorem to calculate conditional probabilities.
- Discuss examples of situations where you might want to predict a probability rather than anactual value.
- Describe important components of Bayes' theorem.
- Build a simple Naïve Bayes classifier.
- Convert numbers into categorical predictors.

Predicting probabilities using the Naive Bayes algorithm

Naive Bayes is one of the most successful machine learning methods when it comes toanalysing texts. It is used for:

- email spam filters,
- document classification algorithms
- sentiment analysis

Exact Bayes algorithm

We can use the 'Exact Bayes' algorithm, a concept concocted to understand why certainassumptions are made in the Naïve Bayes algorithm.

Steps

1. Consider that you're given *n* training samples, where all the input and output variables are categorical. Let this be the standing assumption. You have a new data point where you onlyknow the values of the input variables and would like to predict the value of the output variable.
   → N training samples ($X_{i1}, \ldots, X_{ip}$),i = 1, . . . , n
   → Categorical predictors: ($X_{i1}, \ldots, X_{ip}$)
   → Categorical outcome: Y_i
   → New data point ($X_1, \ldots, X_p$) outcome to be found

2. Just as in the *k*-nearest neighbour classifier, you will look at the *k* data points in the trainingdata whose input variable values are closest to the new sample point in question.

3. Now, use this imaginary classifier to predict the outcome variable value of a new sample point as well as to predict the probability of the outcome variable of the new sample point tobe of a specific category.

Predict category

$$C_j = \underset{j=1,m}{arg\,max} |\{i : X_{i1} = X_1, \ldots, X_{ip} = X_p \text{ and } Y_i = C_j\}|$$

Predict probability

$$P(Y = C_j) = \frac{|\{i: X_{i1} = X_1,\ldots,X_{ip} = X_p \text{ and } Y_i = C_j\}|}{|\{i: X_{i1} = X_1,\ldots,X_{ip} = X_p\}|}$$

For example, you mention you have five sample points in the training data whose input variablevalues coincide exactly with the input variable values of our new data point. Out of these five samples, three of category one and two of category two. In that case, your prediction for the probability of the new sample being of category one would be three over five, which is 60 per cent. And our prediction of the probability of the new sample point being of category two would be two over five. In other words, 40 per cent.

Bayes' theorem

Bayes' theorem allows us to calculate the probability of an event A happening under theadditional information that the event B has happened. This is also called a conditional probability.

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

Naive Bayes algorithm

n training samples (X_{i1}... X_{ip}; Y_1), *i*=1,...,*n* with X_{i1},..., X_{ip} categorical predictors, and Y_1as thecategorical outcome.

New data point (X_1,..., X_p), whose outcome should be found.

Algorithm 1: predict category

To each possible category C_1,..., C_m for Y, assign the estimate

$$BP(Y = C_j \mid X_1 = x_1, \ldots, X_p = x_p) \propto BP(Y = C_j) \prod_{i=1}^{p} BP(X_1 = x_1 \mid Y = C_j)$$

Assign to the new data point the category with the highest probability.

Input

(n) training samples (X_{i1}... X_{ip}; Y_i), (i = 1,...,n) with (X_{i1},..., X_{ip}) categorical predictors and (Y_i) asthe categorical outcome.

New data point (X_1... X_p), whose outcome should be found.

Algorithm 2: predict probability

Estimate the probability of *Y* being category C_1,..., C_m as

$$P(Y=C_j \mid X_1=x_1,\ldots, X_p=x_p) = \frac{1}{Z} P(Y=C_j) \prod_{i=1}^{p} P(X_i = x_i \mid Y = C_j)$$

where (Z = BP(X_1 = x_1,..., X_p = x_p)) and (BP(Y = C_j)) and (BP(X_i = x_i | Y = C_j)) are estimatedfrom frequency tables.

Converting features from numerical to categorical

There are two methods to convert numerical features to categorical ones:

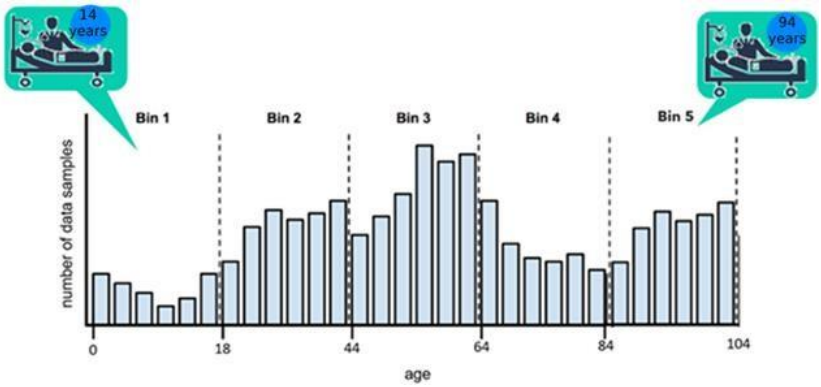
- Manual binning: specifies the cut-off points between consecutive bins manually
- Automated binning: specifies the cut-off points between consecutive bins automatically

Manual binning

Consider a medical application that has a numerical feature called age that you

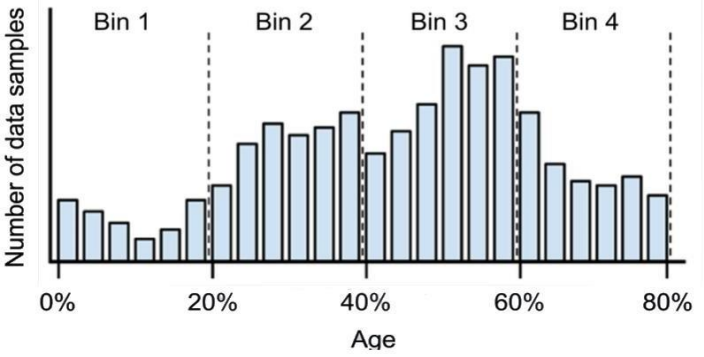
want to convertinto a categorical one. Following the medical literature, specify five age brackets: 0-18 years, 19-44 years, 45-64 years, 65-84 years, and 85 years or above.

Each data sample of a patient with age 14 will get the new feature value 0-18, because that'sthe appropriate age bracket for it. Likewise, a 94-year-old person will get the feature value 85 and above.



Automated binning

The automated binning method is based on quantiles. Using the previous medical example, youonce again have five age-brackets. You can compute the 20%, 40%, 60%, and the 80% quantiles of ages in the dataset. The first bin would contain all the ages that are less than equal to 20% quantile. The second bin would contain all the ages between the 20% and the 40% , andso on.



Choosing the right approach

- Follow the commonly accepted method (if there is one) to bin a numerical data.
- If you have the main knowledge that helps you define natural cut-off values between thebins, use manual binning.
- If neither is the case, use automated binning.

Choosing the right number of bins

- Too few bins can cause information loss
- Too many bins give small count in the frequency table and makes the dataset susceptible tonoise