MODULE 10
Bayes Classifier

LESSON 19

Naive Bayes Classifier

Keywords: Class Conditional Independence, Parameter Estimation

Naive Bayes Classifier

- A naive bayes classifier is based on applying Bayes theorem to find the class of a pattern.
- The assumption made here is that every feature is class conditionally independent.
- Due to this assumption, the probabilistic classifier is simple.
- In other words, it is assumed that the effect of each feature on a given class is independent of the value of other features.
- Since this simplifies the computation, though it may not be always true, it is considered to be a naive classifier.
- Even though this assumption is made, the Naive Bayes Classifier is found to give results comparable in performance to other classifiers like neural network classifiers and classification trees.
- Since the calculations are simple, this classifier can be used for large databases where the results are obtained fast with reasonable accuracy.
- Using the minimum error rate classifier, we classify the pattern X to the class with the maximum posterior probability $P(c \mid X)$. In the naive bayes classifier, this can be written as

  $P(C \mid f_1, ..., f_d)$.

  where $f_1, ..., f_d$ are the features.
- Using Bayes theorem, this can be written as

  $P(C \mid f_1, ..., f_d) = \frac{P(C)\ P(f_1,...,f_d)|C}{p(f_1,...,f_d)}$

  Since every feature f_i is independent of every other feature f_j, for $j \neq i$, given the class

$P(f_i, f_j \mid C) = P(f_i \mid C)P(f_j \mid C)$

So we get,

$P(C, f_1, ..., f_d) = P(C)\ P(f_1 \mid C)\ P(f_2 \mid C)\ \cdots p(f_d \mid C)$

$=$

$$p(C) \prod_{i=1}^{d} p(f_i|C).$$

The conditional distribution over the class variable C is

$$p(C|f_1, \dots, f_n) = \frac{1}{Z} p(C) \prod_{i=1}^{n} p(f_i|C)$$

where Z is a scaling factor.

- The Naive Bayes classification uses only the prior probabilities of classes P(C) and the independent probability distributions $p(f_i \mid C)$.

Parameter Estimation

- In supervised learning, a training set is given. Using the training set, all the parameters of the bayes model can be computed.
- If n_C of the training examples out of n belong to Class C, then the prior probability of Class C will be

  $P(C) = \frac{n_C}{n}$

- In a class C, if n_1 samples take a range of values (or a single value) out of a total of n_C samples in the class, then the prior probability of the feature being in this range in this class will be

$P(f_1$ is in range (a,b)$) = \frac{n_1}{n_C}$

In case of the feature taking a small number of integer values, this can be calculated for each of these values. For example, it would be

$P(f_1$ is 6$) = \frac{n_2}{n_C}$

if n_2 of the n_C patterns of Class c take on the value 6.

- If some class and feature never occur together, then that probability will be zero. When this is multiplied by other probabilities, it may make some probabilities zero. To prevent this, it is necessary to give a small value of probability to every probability estimate.

- Let us estimate the parameters for a training set which has 100 patterns of Class 1, 90 patterns of Class 2, 140 patterns of Class 3 and 100 patterns of Class 4. The prior probability of each class can be calculated.
  The prior probability of Class 1 is

  $P(C_1) = \frac{100}{100+90+140+100} = 0.233$

  The prior probability of Class 2 is

  $P(C_2) = \frac{90}{100+90+140+100} = 0.210$

  The prior probability of Class 3 is

  $P(C_2) = \frac{140}{100+90+140+100} = 0.326$

  The prior probability of Class 4 is

$P(C_2) = \frac{100}{100+90+140+100} = 0.233$

Out of the 100 examples of Class 1, if we consider a particular feature f_1 and if 30 patterns take on the value 0, 45 take on the value 1 and 25 take on the value 2, then the prior probability that in Class 1 the feature f_1 is 0 is

$P(f_1 \text{ is } 0) = \frac{30}{100} = 0.03$

The prior probability that in Class 1 the feature f_1 is 1 is

$P(f_1 \text{ is } 1) = \frac{45}{100} = 0.45$

The prior probability that in Class 1 the feature f_1 is 2 is

$P(f_1 \text{ is } 2) = \frac{25}{100} = 0.25$

Example for Naive Bayes Classifier

Let us take an example dataset.

Consider the example given in decision trees given in the Table 1. We have a new pattern

money = 90, has-exams=yes, and weather=fine

We need to classify this pattern as either belonging to goes-to-movie=yes or goes-to-movie=no.
There are four examples out of 11 belonging to goes-to-movie=yes.
The prior probability of P(goes-to-movie=yes)= $\frac{4}{11}$= 0.364

The prior probability of P(goes-to-movie=no) = $\frac{7}{11}$ = 0.636

There are 4 examples with $money 50 - 150$ and goes-to-movie=no and 1 examples with $money < 50$ and goes-to-movie=yes. Therefore,

Money	Has-exams	weather	Goes-to-movie
25	no	fine	no
200	no	hot	yes
100	no	rainy	no
125	yes	rainy	no
30	yes	rainy	no
300	yes	fine	yes
55	yes	hot	no
140	no	hot	no
20	yes	fine	no
175	yes	fine	yes
110	no	fine	yes

Table 1: Example training data set

$P(money50 - 150 \mid goes - to - movie = yes) = \frac{1}{4} = 0.25$ and

$P(money50 - 150 \mid goes - to - movie = no) = \frac{4}{7} = 0.429$

There are 4 examples with has-exams=yes and goes-to-movie=no and 2 examples with has-exams=yes and goes-to-movie=yes. Therefore,

$P(has - exams \mid goes - to - movie = yes) = \frac{2}{4} = 0.5$

$P(has - exams \mid goes - to - movie = no) = \frac{4}{7} = 0.429$

There are 2 examples with weather=fine and goes-to-movie=no and 2 examples with weather=fine and goes-to-movie=yes. Therefore,

$P(weather = fine \mid goes - to - movie = yes) = \frac{2}{4} = 0.5$

$P(weather = fine \mid goes - to - movie = no) = \frac{2}{7} = 0.286$

Therefore

$$P(goes - to - movie = yes \mid X) = 0.364 * 0.25 * 0.5 * 0.5 = 0.023$$

$$P(goes - to - movie = no \mid X) = 0.636 * 0.429 * 0.429 * 0.286 = 0.033$$

Since $P(goes - to - movie = no \mid X)$ is larger, the new pattern is classified as belonging to the class goes-to-movie=no.