Lecture notes: "Naive" Bayes classifier

(c) Marcin Sydow

Naive Bayes

We assume in this lecture that all the attributes are nominal (categorical).
The training set T consists of N observations, each being a n-dimensional vector of (nominal) attributes.

We treat each attribute X_i and the decision attribute Y as **random variables**.

The goal is to classify a vector $x = (x_1, x_2, ..., x_n)$

We apply the Bayes formula:

$$P(Y = y|X = x) = \frac{P(X = x|Y = y)P(Y = y)}{P(X = x)}$$

(interpretation: the probability that the decision attribute Y is equal to y, conditioned on the fact that the attribute vector (to be classified) is represented by the vector x)

Bayes classification rule

We classify the vector x to that class y (the value of the decision attribute), for which the Bayes probability above is maximal.

Thus, we compute the above probability for all the possible classes/categories y (values of the variable Y) and select the value y giving the maximal value of the probability $P(Y = y|X = x)$

Due to the fact that all the compared probabilities have the same denominator ($P(X = x)$), it is possible to omit it in computations.

"Naive" Bayes classifier

The key assumption for the **naive** Bayes classification is the "naive" assumption that all the attributes are **independent** random variables, so that:

$$P(X = (x_1, ..., x_n)|Y = y) = P(X_1 = x_1|Y = y) * ... * P(X_n = x_n|Y = y)$$

Thus, due to independence we obtain:
$P(Y = y|X = (x_1, ..., x_n)) \propto P(X_1 = x_1|Y = y) * ... * P(X_n = x_n|Y = y) * P(Y = y)$
where, the estimations of the probabilities can be made directly from the training set:

- $P(X_i = x_i|Y = y)$ (ratio of the observations in the training set T that have the value of the attribute $X_i = x_i$ among all the observations that have the value of the decision attribute $Y = y$)
- $P(Y = y)$ (ratio of the observations in the training set that have the value of the decision attribute $Y = y$)

Smoothing

It may happen that in the training set T there is no observation that satisfies $X_j = x_j$ and $Y = y$ for some attribute j.

In such case, the estimation of the probability $P(X_j = x_j|Y = y)$ from the training set T would be equal to 0 and would make the whole product of probabilities being zero, independently on the values of all the other probabilities $P(X_i = x_i|Y = y)$.

To avoid this problem, the technique of **smoothing** can be applied. It consists of assuring that even in such case the probability will be non-zero, i.e. it will be substituted by some small, positive value. This is achieved by "borrowing" (decreasing) part of value from all the other non-zero values of probabilities for this attribute.

Simple smoothing

A simple implementation of the idea of smoothing is as follows. We modify the ratio representing the probability so that we add 1 to the numerator and add the number of different values of this attribute to the denominator.
In this way, all the conditional probabilities for this attribute sum up to 1, and the zero estimation of the probability is avoided even if such case is not present in the training set T.

Thank you for the attention.