NAIVE BAYES CLASSIFICATION

Naive Bayes Classification is a method used for binary or multinomial classification (but not a regression) that utilizes the Bayes' formula. Suppose there are k predictors $\mathbf{X} = (X_1, \ldots, X_k)$ which are binary, categorical, or continuous. And let Y denote the response variable. By the Bayes' formula

$$\mathbb{P}(Y|\mathbf{X}) = \frac{\mathbb{P}(\mathbf{X}|Y)\mathbb{P}(Y)}{\mathbb{P}(\mathbf{X})}.$$

The Naive Bayes classification method assumes that the predictors are conditionally independent, given Y, that is,

$$\mathbb{P}(Y|\mathbf{X}) = \frac{\mathbb{P}(Y)\prod_{i=1}^{k}\mathbb{P}(X_i|Y)}{\mathbb{P}(\mathbf{X})}.$$

This conditional independence assumption is rather naive, hence the name of the technique.

In classification problem (binary or multinomial), we compute the conditional (posterior) probability $\mathbb{P}(Y|\mathbf{X})$ of each class, and classify the record into the class with the highest probability. Since we compare the posterior probabilities and the denominator $\mathbb{P}(\mathbf{X})$ is present in each expression, it can be ignored. That is, $\mathbb{P}(Y|\mathbf{X})$ is proportional to $\mathbb{P}(Y)\prod_{i=1}^{k}\mathbb{P}(X_i|Y)$ up to a multiplicative constant.

To estimate the prior probability $\mathbb{P}(Y = y)$ of each class y, we compute the proportion of observations in each class in the training set. To compute the empirical conditional probabilities $\mathbb{P}(X_i = x|Y = y)$ for categorical predictors, we calculate the fraction of observations in the class $Y = y$ in the training set for which $X_i = x$. If a predictor is continuous, we assume that the underlying distribution is normal (Gaussian) with estimated mean $\hat{\mu} = \bar{x}$ and estimated variance $\hat{\sigma}^2 = s^2$.

Characteristics of Naive Bayes Classifiers

1. Robust to outliers because they average out when computing posterior probabilities.
2. Handles missing values by ignoring the missing data points in calculations.
3. Robust to irrelevant predictors since $\mathbb{P}(X_i|Y)$ is almost uniformly distributed and factors out in comparisons of posterior probabilities.
4. Correlated predictors can degrade the performance of the technique. The conditional independence assumption is the key.

Example. Suppose the training data are as given in the table below.

ID	Home Owner	Marital Status	Annual Income ($K)	Defaulted Borrower
1	yes	single	125	no
2	no	married	100	no
3	no	single	70	no
4	yes	married	120	no
5	no	divorced	95	yes
6	no	married	60	no
7	yes	divorced	220	no
8	no	single	85	yes
9	no	married	75	no
10	no	single	90	yes

The prior probabilities are $\mathbb{P}(default = no) = 7/10 = 0.7$, $\mathbb{P}(defaul = yes) = 3/10 = 0.3$. The conditional probabilities are:

$$\mathbb{P}(homeowner = yes \,|\, default = no) = 3/7,$$
$$\mathbb{P}(homeowner = yes \,|\, default = yes) = 0,$$
$$\mathbb{P}(homeowner = no \,|\, default = no) = 4/7,$$
$$\mathbb{P}(homeowner = no \,|\, default = yes) = 1,$$
$$\mathbb{P}(maritalstatus = single \,|\, default = no) = 2/7,$$
$$\mathbb{P}(maritalstatus = single \,|, default = yes) = 2/3,$$
$$\mathbb{P}(maritalstatus = married \,|\, default = no) = 4/7,$$
$$\mathbb{P}(maritalstatus = married \,|\, default = yes) = 0,$$
$$\mathbb{P}(maritalstatus = divorced \,|\, default = no) = 1/7,$$
$$\mathbb{P}(maritalstatus = divorced \,|\, default = yes) = 1/3.$$

The posterior density for annual income is normal with the estimated parameters for *default=no*, $\hat{\mu}$ =sample mean= $(125 + 100 + 70 + 120 + 60 + 220 + 75)/7 = 110$, $\hat{\sigma}^2 = s^2 = 2975$, and for *default=yes*, $\hat{\mu} = (95 + 85 + 90)/3 = 90$, and $\hat{\sigma}^2 = s^2 = 25$.

Suppose we would like to predict the default status for a person who is not a home owner, who is single, and whose annual income is $\$120K$. We write

$$\mathbb{P}(\mathbf{X} \,|\, default = no) = \mathbb{P}(homeowner = no \,|\, default = no) \times \mathbb{P}(maritalstatus = single \,|\, default = no) \times$$
$$\mathbb{P}(annualincome = \$120K \,|\, defaul = no) = (4/7)(2/7)\frac{1}{\sqrt{(2\pi)(2975)}}\, e^{-\frac{(120-110)^2}{(2)(2975)}} = 0.001215,$$

and

$$\mathbb{P}(\mathbf{X} \,|\, default = yes) = \mathbb{P}(homeowner = no \,|\, default = yes) \times \mathbb{P}(maritalstatus = single \,|\, default = yes) \times$$

$$\mathbb{P}(annualincome = \$120K \,|\, defaul = yes) = (1)(2/3)\frac{1}{\sqrt{(2\pi)(25)}}\, e^{-\frac{(120-90)^2}{(2)(25)}} = (8.1)(10)^{-10}.$$

Hence,

$$\begin{aligned}\mathbb{P}(default = no \,|\, \mathbf{X}) &= \mathbb{P}(default = no)\mathbb{P}(\mathbf{X} \,|\, default = no)/\mathbb{P}(\mathbf{X}) \\ &= (0.7)(0.001215)/\mathbb{P}(\mathbf{X}) = 0.000851/\mathbb{P}(\mathbf{X}),\end{aligned}$$

and

$$\begin{aligned}\mathbb{P}(default = yes \,|\, \mathbf{X}) &= \mathbb{P}(default = yes)\mathbb{P}(\mathbf{X} \,|\, default = yes)/\mathbb{P}(\mathbf{X}) \\ &= (0.3)(8.1)(10)^{-10}/\mathbb{P}(\mathbf{X}) = (2.43)(10)^{-10}/\mathbb{P}(\mathbf{X}).\end{aligned}$$

We can see that $\mathbb{P}(default = no \,|\, \mathbf{X}) > \mathbb{P}(default = yes \,|\, \mathbf{X})$ and so we predict *default=no* for this person. □