Naive Bayes classifier

Nathanaël Carraz Rakotonirina

Mathématiques Informatique et Statistique Appliquées (MISA)
Université d'Antananarivo

Model

It is a generative classifier which assumes that the features are conditionally independent given the class label (Naive Bayes assumption). The model is called "naive" since the features are generally not independent still it gives good results. The class conditional density has the form

$$p(\boldsymbol{x}|y=c;\boldsymbol{\theta}) = \prod_{i=1}^{N} p(x_i|y=c;\boldsymbol{\theta}_{dc})$$

where $\boldsymbol{\theta}$ are the parameters for class c and feature d. The posterior is given by

$$p(y=c|\boldsymbol{x};\boldsymbol{\theta}) = \frac{\pi_c \prod_{d=1}^{D} p(x_d|y=c;\boldsymbol{\theta}_{dc})}{\sum_{c'} \pi'_c \prod_{d=1}^{D} p(x_d|y=c';\boldsymbol{\theta}_{dc'})}$$

where $\pi_c = p(y=c)$ is the prior probability of class c

Model

The distribution depends on the nature of x_d

Bernoulli

Bernoulli distribution for Binary features, $x_d \in \{0, 1\}$:
$p(\boldsymbol{x}|y = c; \boldsymbol{\theta}) = \prod_{d=1}^{D} \text{Ber}(x_d; \theta_{dc})$ where θ_{dc} is the probability that $x_d = 1$ in class c.

Categorical

Categorical distribution for categorical features, $x_d \in \{1, ..., K\}$:
$p(\boldsymbol{x}|y = c; \boldsymbol{\theta}) = \prod_{d=1}^{D} \text{Cat}(x_d; \boldsymbol{\theta}_{dc})$ where θ_{dck} is the probability that $x_d = k$ in class c.

Univariate Gaussian

Univariate Gaussian distribution for real-valued features, $x_d \in \{1, ..., K\}$:
$p(\boldsymbol{x}|y = c; \boldsymbol{\theta}) = \prod_{d=1}^{D} \mathcal{N}(x_d; \mu_{dc}, \sigma_{dc}^2)$ where μ_{dck} and σ_{dc}^2 are the mean and variance of feature d when the class label is c (equivalent to Gaussian discriminant analysis using diagonal covariance matrices).

Model fitting

The MLE can be written as follows :

$$\prod_{i=1}^{N} p(y_i|\boldsymbol{x}_i;\boldsymbol{\theta}) = \prod_{i=1}^{N} \text{Cat}(y_i|\boldsymbol{\pi}) \prod_{d=1}^{D} p(x_{id}|y_i;\boldsymbol{\theta}_d)$$
$$= \prod_{i=1}^{N} \text{Cat}(y_i|\boldsymbol{\pi}) \prod_{d=1}^{D} \prod_{c=1}^{C} p(x_{id};\boldsymbol{\theta}_{dc})^{\mathbb{I}(y_i=c)}$$

The log-likelihood is

$$\left[\sum_{i=1}^{N}\sum_{c=1}^{C} \mathbb{I}(y_i = c) \log \pi_c\right] + \sum_{c=1}^{C}\sum_{d=1}^{D}\left[\sum_{i:y_i=c} \log p(x_{id};\boldsymbol{\theta}_{dc})\right]$$

We can separately estimate the parameters $\boldsymbol{\pi}$ and $\boldsymbol{\theta}_{dc}$. The MLE of $\boldsymbol{\pi}$ is

$$\hat{\pi}_c = \frac{N_c}{N}$$

The MLE of $\boldsymbol{\theta}_{dc}$ depends on the distribution of the class conditional density :

- Bernoulli :

$$\hat{\theta}_{dc} = \frac{N_{dc}}{N_c}$$

- Categorical :

$$\hat{\theta}_{dck} = \frac{N_{dck}}{\sum_{k'=1}^{K} N_{dck'}} = \frac{N_{dck}}{N_c}$$

  where N_{dck} is the number of times feature d had value k in examples of class c

- Univariate Gaussian :

$$\hat{\boldsymbol{\mu}}_{dc} = \frac{1}{N_c} \sum_{i:y_i=c} x_{id}$$

$$\hat{\sigma}^2_{dc} = \frac{1}{N_c} \sum_{i:y_i=c} (x_{id} - \hat{\mu}_{dc})^2$$

More

As usual, there is always more.

Explore further

- Bayesian naive Bayes
- Connection between naive Bayes and logistic regression