Navie Bayes Classifier

Andersen Ang

First created: 2014. Last update : 2017

- Consider a supervised learning problem : given data, the aim is try to construct a parametric probabilistic models, assume the model is static.
- When all th attributes of the data (feature vector for example) are independent of each other.
- For a data vector $\mathbf{x} = [x_1\, x_2\, ..., x_n]^T$ with class label C.
- Using Bayes rule, the probability of class label C given the features $\mathbf{x}$ is :

$$\mathbb{P}\left(C \,|\, x_1, x_2, ..., x_n\right) = \frac{\mathbb{P}\left(x_1,\, ...\,, x_n|C\right)\mathbb{P}\left(C\right)}{\mathbb{P}\left(x_1, ..., x_n\right)}$$

- Because now all the parameters are assumed to be independent of each other

$$\mathbb{P}\left(x_1,\, ...\,, x_n|C\right) = \mathbb{P}\left(x_1|C\right)\mathbb{P}\left(x_2|C\right)...\mathbb{P}\left(x_n|C\right)$$

- Using short hand notation

$$\mathbb{P}\left(C \,|\, x_1, x_2, ..., x_n\right) = \frac{\prod_i \mathbb{P}\left(x_i|C\right)\mathbb{P}\left(C\right)}{\mathbb{P}\left(x_1, ..., x_n\right)}$$

- $\mathbb{P}\left(C\right)$: Prior probability distribution, which can be estimated from data.
- $\mathbb{P}\left(x_i|C\right)$: likelihood function, class conditional probability distribution, which can be estimated from data.
- $\mathbb{P}\left(x_1,\, ...,\, x_n\right)$: __ distribution, which is a constant

In this case, the class label C is choosen such that the $\mathbb{P}\left(C \,|\, x_1, x_2, ..., x_n\right)$ is maximized.

$$\text{maximize}\mathbb{P}\left(C \,|\, x_1, x_2, ..., x_n\right) \qquad \Longleftrightarrow \qquad \text{maximize}\frac{\prod_i \mathbb{P}\left(x_i|C\right)\mathbb{P}\left(C\right)}{\mathbb{P}\left(x_1, ..., x_n\right)}$$

Since $\mathbb{P}(x_1, ..., x_n)$ and $\mathbb{P}(C)$ is a constant.

$$\text{maximize}\mathbb{P}(C \mid x_1, x_2, ..., x_n) \qquad \Longleftrightarrow \qquad \text{maximize}\mathbb{P}(C) \prod_i \mathbb{P}(x_i|C)$$

Therefore, in Naive Bayes Classifier, we need to estimate the Prior probability and conditional probability.

Example. Normal distribution

$$\mathcal{N}(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$$

Based on Bayes assumption, all the parameters are independent, so the pdf can be decomposed as

$$\mathbb{P}(x_1, x_2, ..., x_n|\mu, \sigma) = \prod_{i=1}^{N} \mathbb{P}(x_i|\mu, \sigma)$$

$$\mathbb{P}(x_1, x_2, ..., x_n|\mu, \sigma) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2\right)$$

Log likelihood

$$\log \mathbb{P}(x_1, x_2, ..., x_n|\mu, \sigma) = \sum_{i=1}^{N} \log \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2\right)$$

$$L = \log \mathbb{P}(x_1, x_2, ..., x_n|\mu, \sigma) = N(-\sqrt{2\pi} - \log\sigma) + \sum_{i=1}^{N} \frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2$$

Take derivative

$$\frac{\partial L}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{N}(x_i - \mu) = 0 \qquad \Rightarrow \qquad \mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

$$\frac{\partial L}{\partial \sigma} = -\frac{N}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{N}(x_i - \mu)^2 = 0 \qquad \Rightarrow \qquad \sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$