Bayesian, Logistic Regression, Naive Bayesian

yongduek.seo@gmail.com

September 2018

Bayesian Formula:

$$P(l|d) = \frac{P(d|l)P(l)}{P(d)} = \frac{P(d|l)P(l)}{\sum_i P(d|l_i)P(l_i)} \tag{1}$$

1 Two Classes

For a two class problem, the posterior probability for class L_1 given data d can be written as

$$\begin{aligned} P(L_1|d) &= \frac{P(d|L_1)P(L_1)}{P(d|L_1)P(L_1) + P(d|L_2)P(L_2)} && (2) \\ &= \frac{1}{1 + \frac{P(d|L_2)P(L_2)}{P(d|L_1)P(L_1)}} && (3) \\ &= \frac{1}{1 + \exp\left(-\ln \frac{P(d|L_1)P(L_1)}{P(d|L_2)P(L_2)}\right)} && (4) \end{aligned}$$

Let's define the **logistic sigmoid** function $\sigma(z)$ by

$$\sigma(z) = \frac{1}{1 + \exp(-z)} \tag{5}$$

The inverse of the logistic sigmoid, known as the **logit** function is given by

$$z = \ln\left(\frac{\sigma}{1-\sigma}\right) \tag{6}$$

which represents the log of the ratio of probabilities $\ln(P(L_1|d)/P(L_2|d))$ for the two classes. This is known as the **log odds**.

Now, let's make a linear model for the log odds:

$$z = \ln \frac{P(d|L_1)P(L_1)}{P(d|L_2)P(L_2)} = \sum_i w_i f_i + w_0 = \boldsymbol{w}^\top \boldsymbol{f} \tag{7}$$

where f_i is i-th feature value of the data d, w_i and w_0 are parameters for linear modeling. This model is known as **logistic regression** in the terminology of statistics.

1.1 Problem Solving

For a data set $\{d_n, y_n\}$ where $y_n \in \{0, 1\}$, and $n = 1, ..., N$. The likelihood function can be written

$$P(y_1, ..., y_n | w_0, ..., w_n) = \prod_{n=1}^{N} p_n^{y_n} (1 - p_n)^{1 - y_n} \tag{8}$$

where $p_n = P(L_1 | d_n)$.

As usual, we can define an error function by taking the negative logarithm of the likelihood, which gives the **cross-entropy** error function in the form

$$E(\boldsymbol{w}) = -\ln P(\boldsymbol{y} | \boldsymbol{w}) = -\sum_{n=1}^{N} \left\{ y_n \ln p_n + (1 - y_n) \ln(1 - p_n) \right\} \tag{9}$$

where $p_n = P(L_1 | d_n) = \sigma(z_n)$ and $z_n = \boldsymbol{w}^\top \boldsymbol{f}$.

2 Naive Bayes

Conditional independence property:

$$p(f_1, f_2 | L) = p(f_1 | L) p(f_2 | L) \tag{10}$$

Then the posterior probability is given by

$$\begin{aligned} p(L_k | f_1, f_2) &= \frac{p(f_1, f_2 | L_k) p(L_k)}{\sum_i p(f_1, f_2 | L_i) p(L_i)} && (11) \\ &\propto p(f_1, f_2 | L_k) p(L_k) && (12) \\ &\propto p(f_1 | L_k) p(f_2 | L_k) p(L_k) && (13) \end{aligned}$$

This is utilized for Naive Bayes Models. That is, the posterior distribution over the class variable L under the conditional independence assumption is given by

$$p(L_k | f_1, ..., f_n) = \frac{1}{Z} p(L_k) \prod_{i=1}^{n} p(f_i | L_k) \tag{14}$$

where Z is the **evidence**

$$Z = p(f_1, ..., f_n) = \sum_i p(L_i) p(f_1, ..., f_n | L_i) \tag{15}$$

which is a scaling factor dependent only on $f_1, ..., f_n$. Note that Z is a constant if the values of the feature variables f_i are known and fixed.

2.1 Problem solving

$$\hat{L} = \underset{k \in \{1,...,K\}}{\arg\max} \quad p(L_k) \prod_{i=1}^{n} p(f_i|L_k) \tag{16}$$

The class probability L_k can be estimated from the data population, and the feature likelihood $p(f|L_k)$ is modeled based on the characteristics of the features.

- Gaussian naive Bayes

$$p(f = f|L_k) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left[-\frac{(f-\mu_k)^2}{2\sigma_k^2}\right] \tag{17}$$

- Bernoulli naive Bayes

$$p(f_1, ..., f_n) = \prod_{i=1}^{n} p_{ki}^{f_i}(1 - p_{ki})^{(1-f_i)} \tag{18}$$

  where f_i is a boolean variable expressing the occurrence or absence of the i-th feature from the feature vocabulary, and p_{ki} is the probability of class L_k generating the feature f_i. The probability p_{ki} must be learned before inference.

- Multinomial naive Bayes.

Bibliography

https://en.wikipedia.org/wiki/Naive_Bayes_classifier

Bishop, Pattern Recognition and Machine Learning, Springer 2006.