THE NAIVE BAYES CLASSIFIER

The naïve Bayes classifier is another generative method for classification.

The Naive Bayes Assumption

Let $X = [X^{(1)} \dots X^{(d)}]^T \in \mathbb{R}^d$ denote the random feature vector in a classification problem, and Y the corresponding label.

The naive Bayes classifier assumes that, given Y,

(A) $X^{(1)}, \dots, X^{(d)}$ are _____.

Although this assumption is rarely met in practice, it can still lead to reasonable, and sometimes very good, classification performance.

The reason is that in classification, we really don't need to model the class-conditional densities well, just the decision boundary of the Bayes classifier.

Estimation

The major advantage of NB is the we only

(B) need to estimate _____ densities.

Let $g_k(x)$ be the probability law (density or mass function) of $X \mid Y = k$, $k = 1, \ldots, K$

By the NB assumption,

(C) $g_k(x) =$

Let $(x_i, y_i)_{i=1}^n$ be training data, and let

$$\hat{\pi}_k = \frac{|\{i : y_i = k\}|}{n}$$

$\hat{g}_k^{(j)}$ = ... estimate of $g_k^{(j)}$ based on $\{x_i^{(j)} : y_i = k\}$

Then the NB classifier is

(D)

So it remains to determine $\hat{g}_k^{(j)}$.

Another advantage of NB is that it easily handles the case where some $X^{(j)}$ are <u>continuous</u> and others are <u>discrete</u>.

Continuous $X^{(j)} \mid Y=k$

- Gaussian MLE
- kernel density estimate
- quantize to discrete variable

Discrete $X^{(j)} \mid Y=k$

Suppose that given $Y=k$, $X^{(j)}$ takes on the values $z_1, \ldots, z_L$. Denote

Note: z_ℓ depends on j, k, but this dependence is omitted for simplicity.

$$n_k = |\{i : y_i = k\}|$$

$$n_{kl}^{(j)} = |\{i : y_i = k \wedge x_i^{(j)} = z_\ell\}|$$

Then the natural (and maximum likelihood) estimate of $\Pr\{X^{(j)} = z_\ell \mid Y=k\}$ is

That is,

$$\hat{g}_k^{(j)}(z_e) = \frac{n_{ke}^{(j)}}{n_k}.$$

It is possible that when we apply NB to a test pattern X, $X^{(j)}$ may take on a value z' not observed in the training data. In this case, the MLE $\hat{g}_k$ is undefined, which is undesirable.

Example | Text classification

$X^{(j)}$ = # of times word j occurs in a document,
$j = 1, \ldots, d$; where d = total # of words in training data.
In general, $d <$ # words in the vocabulary.

To avoid this problem it is c to estimate

(F) $$\hat{g}_k^{(j)}(z_e) =$$

which corresponds to a Bayesian estimate (of multinomial parameters with a Dirichlet prior)

Key

A. independent

B. scalar / univariate

C. $g_k(x) = \prod_{j=1}^{d} g_k^{(j)}(x^{(j)})$

D. $\hat{f}(x) = \underset{k=1,\ldots,K}{\arg\max} \; \hat{\pi}_k \, \hat{g}_k(x)$

E. $n_{kl}^{(j)} / n_k$

F. $\dfrac{n_{kl}^{(j)} + 1}{n_k + L}$ (note L depends on j, k)