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,

Ⓐ $X^{(1)}, \dots, X^{(d)}$ 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 that we only Ⓑ 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,

Ⓒ $$g_k(x) =$$

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

Ⓓ

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

Another advantage of NB is that it easily handles the case where some $\mathbf{X}^{(j)}$ 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 $z_1, \dots, z_L$. Denote

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

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

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

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

That is,

$$\hat{g}_k^{(j)}(z_\ell) = \frac{n_{k\ell}^{(j)}}{n_k}.$$

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 for some class. Then $\hat{g}_k^{(j)}(z') = 0$ and hence $\hat{g}_k = 0$, so that class will never be predicted. This is undesirable.

<u>Example</u> Document classification

Suppose you wish to classify documents, e.g. 1 = politics, 2 = sports, 3 = finance, etc. One simple representation of a document is $X = [X^{(1)} \cdots X^{(d)}]^T$ where d is the number of words in the vocabulary, and

$$X^{(j)} = \begin{cases} 1 & \text{if } j\text{th word occurs in document} \\ 0 & \text{otherwise.} \end{cases}$$

It could happen that all training documents contain the word "ball." But we don't want to exclude the possibility that some other sports article does not contain this word. //

To avoid this problem, it is common to estimate

$$\hat{g}_k^{(j)}(z_\ell) =$$

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

E. $$n_{k\ell}^{(j)} / n_k$$

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