THE NAIVE BAYES CLASSIFIER

The Naive Bayes Assumption

NB is a plug-in method. It could be generative or discriminative, and parametric or nonparametric, depending on design choices.

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

The NB classifier assumes that, given Y, $X^{(1)}, \dots, X^{(d)}$ are independent.

Main Use: Features with Finite Range

Let's assume each feature $X^{(j)}$ takes values $z_1, \dots, z_L$. In this setting, is NB generative or discriminative? Parametric or nonparametric?

Example Document classification

Suppose we wish to classify documents into categories like "business," "politics," "sports," etc. A simple yet popular feature representation is the bag-of-words representation. A document is represented as a vector

$$X = [X^{(1)} \ \cdots \ X^{(d)}]$$

where d is the # of words in the vocabulary, and

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

Let $g_k(x)$ be the pmf of $X \mid Y=k$. By the NB assumption,

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

marginal pmf of $X^{(j)} \mid Y=k$

Let $(x_1, y_1), \ldots, (x_n, y_n)$ be training data, and let

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

$$\hat{g}_k^{(j)} = \text{estimate of } \hat{g}_k^{(j)}$$

Then the NB classifier is

$$\hat{f}(x) = \arg\max_k \hat{\pi}_k \cdot \prod_{j=1}^{d} g_k^{(j)}(x)$$

So how should we estimate $g_k^{(j)}$? Denote

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

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

Then the natural (and maximum likelihood) estimate of

$$g_k^{(j)}(z_\ell) = \Pr\{X^{(j)} = z_\ell \mid Y = k\}$$

is

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

Does this seem reasonable?

What would happen if all sports documents in our training data contain the word "ball"?

Any document not containing ball will have a pmf of zero for the sports category. But this may be undesirable. An alternative is

$$g_k^{(j)}(z_\ell) = \frac{n_{k\ell}^{(j)} + 1}{n_k + L}$$

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

$\boxed{\text{Other Models}}$

NB can be used when $X^{(j)}$ are continuous RVs.

Then we could model $X^{(j)}$ as a univariate Gaussian and estimate the parameters via maximum likelihood. Alternatively, we could estimate the marginal densities $g_k^{(j)}$ with a nonparametric density estimator, such as the kernel density estimator.