The Naive Bayes Method

Suppose we're interested in predicting a target, y, that has possible values $0, \ldots, C-1$, based on inputs $x_1, \ldots, x_p$. To do this, we would like a model for $P(y = c \,|\, \mathbf{x}_1, \ldots, x_p)$. One way to get one is to apply Bayes' Rule:

$$P(y = c \,|\, x_1, \ldots, x_p) \;\;=\;\; \frac{P(y = c)\, P(x_1, \ldots, x_p \,|\, y = c)}{\sum_{c'} P(y = c')\, P(x_1, \ldots, x_p \,|\, y = c')}$$

This is always valid — it's like using the results of unsupervised learning of the joint distribution, $P(y,\, x_1, \ldots, x_p)$.

Now, however, we do something that is very "naive", and likely to be wrong — we assume that given a value for y, all of $x_1, \ldots, x_p$ are mutually independent. With this assumption, we can write

$$P(x_1, \ldots, x_p \,|\, y = c) \;\;=\;\; \prod_{j=1}^{p} P(x_j \,|\, y = c)$$

It's sometimes useful to look at this model in terms of log odds:

$$\log\, \frac{P(y = c \,|\, x_1, \ldots, x_p)}{P(y = c' \,|\, x_1, \ldots, x_p)} \;\;=\;\; \log\, \frac{P(y = c)}{P(y = c')} \;+\; \sum_{j=1}^{p} \log\, \frac{P(x_j \,|\, y = c)}{P(x_j \,|\, y = c')}$$

Estimating the Univariate Conditional Distributions

Estimating $P(y=c)$ is easy — just count occurrences in the training set.

Estimating $P(x_j \,|\, y=c)$ for $c = 0, \ldots C{-}1$ and $j = 1, \ldots, p$ is also easy — at least compared to estimating a joint distribution like $P(y,\, x_1, \ldots, x_p)$.

If x_j is discrete, we can just count how often each value for x_j occurs in training cases where $y = c$. Adjusting these to avoid any zero probabilities may be desirable, perhaps as follows:

$$P(x_j = a \,|\, y = c) \;\;=\;\; \frac{\alpha \;+\; \#\{i \;:\; y^{(i)} = c \text{ and } x_j^{(i)} = a\}}{M\alpha \;+\; \#\{i \;:\; y^{(i)} = c\}}$$

where M is the number of possible values for x_j.

If x_j is continuous, we might model its conditional distribution given $y = c$ as Gaussian, with mean and variance estimated from the training cases where $y = c$.

If the conditional distributions are clearly not Gaussian, more elaborate techniques might be necessary. Doing this for each variable separately is still a lot easier than estimating joint distributions of many variables.

Will the Independence Assumption be Correct?

Suppose we're predicting whether or not someone applying for a loan would default on the payments ($y = 1$ means default, $y = 0$ means they pay).

We have available the following inputs:

x_1	1 if they have ever been declared bankrupt, 0 if not
x_2	1 if they have been late on loan payments in the past, 0 if not
x_3	Education: 0 = no HS diploma, 1 = HS, 2 = Bachelors, 3 = grad degree
x_4	Income: 0 = below \$25000, 1 = \$25000 – \$50000, 2 = over \$50000
x_5	Age: 0 = below 18, 1 = 18 – 30, 2 = over 30

Among people who *do not* default ($y = 0$), is it reasonable to think that x_1 through x_5 are mutually independent? Ask yourself, if I know some of these inputs, will that tell me something about some of the others?

Among people who *do* default ($y = 1$), is it reasonable to think that x_1 through x_5 are mutually independent?

Note: It's *not* necessary for x_1 to x_5 to be independent unconditionally (ie, if you don't know y).

Using Wrong Models — Naively and Otherwise

For most problem, the inputs are not independent given y, though people may "naively" assume independence, thinking it's true.

But we might assume independence even if we know it's not true — a model that is not right can still be useful.

Look again at the log odds form of Naive Bayes:

$$\log \frac{P(y = c \,|\, x_1, \ldots, x_p)}{P(y = c' \,|\, x_1, \ldots, x_p)} \;=\; \log \frac{P(y = c)}{P(y = c')} + \sum_{j=1}^{p} \log \frac{P(x_j \,|\, y = c)}{P(x_j \,|\, y = c')}$$

Each term on the right represents evidence regarding $y = c$ versus $y = c'$. If inputs aren't really independent, we may count some evidence twice. This might be important if the log odds are close to our threshold for predicting one way or another. But often, the evidence is so clear that a bit of error doesn't matter.

Experience shows that Naive Bayes often has a pretty good error rate, *even if* the probabilities $P(y = c \,|\, x_1, \ldots, x_p)$ that it produces are far from being correct. But this isn't always so, and sometimes we really need good estimates of probabilities.