Naive Bayes and multi-class classification

We are given d predictor attributes $X_1, \ldots, X_d$. We assume each predictor attribute is *binary*. Suppose target attribute Y can take on one of k possible values $y_1, y_2, \ldots, y_k$. For example, if we are classifying spam/ham then $y_1 =$ spam and $y_2 =$ ham; if we are classifying digits, then $y_1 = 0$, $y_2 = 1$, and $y_{10} = 9$.

Let us define the *joint probability* that $Y = y_j$ and $X = x$ under the Naive Bayes assumption to be the following quantity:

$$P(Y = y_j, X = x) = P(Y = y_j) \prod_{i=1}^{d} P(X_i = x_i | Y = y_j)$$

A naive bayes classifier for computes the above quantity for each $j = 1, \ldots, k$.

The classification decision is to predict whichever y_j value maximizes the above quantity.

In class, we did exactly what is described above for the special case when Y takes on two values so we need only compute: $P(Y = y_1, X = x)$ and $P(Y = y_2, X = x)$ and take the larger of the two.

Log Joint Probabilities

To avoid numerical underflow (as discussed DSFS Ch. 13), it is better to compute the logarithm of the joint probability:

$$\begin{aligned} \log\left(P(Y = y_j, X = x)\right) &= \log\left(P(Y = y_j) \prod_{i=1}^{d} P(X_i = x_i | Y = y_j)\right) \\ &= \log\left(P(Y = y_j)\right) + \sum_{i=1}^{d} \log\left(P(X_i = x_i | Y = y_j)\right) \end{aligned}$$

Observe that taking the log doesn't affect which y_j maximizes the joint probability.

Maximum Odds Ratios

One diagnostic tool for looking at which features are predictive is to look at odds ratios. Let's assume that Y takes on two values, y_1 and y_2. Let X_i be a categorical attribute that takes on one value from the set $a_1, a_2, \ldots, a_m$.

Define the maximum odds ratio as:

$$\text{maxOdds}(X_i, y_1, y_2) = \max_{x \in \{a_1, a_2, \ldots, a_m\}} \frac{P(X_i = x | Y = y_1)}{P(X_i = x | Y = y_2)}$$

The maximum is over all possible values that X_i can take on. If X_i is a binary attribute, this equation simplifies to:

$$\text{maxOdds}(X_i, y_1, y_2) = \max \left\{ \frac{P(X_i = 1 | Y = y_1)}{P(X_i = 1 | Y = y_2)}, \frac{P(X_i = 0 | Y = y_1)}{P(X_i = 0 | Y = y_2)} \right\}$$

This ratio will be greater than one for features which cause belief in y_1 to increase relative to y_2.