An Introduction to Naive Bayes Classification

Colan F. Biemer

In this I want to show the naive bayes classificaiton model and how it can be used for qualitative and quantitative data. I think it is easier to start with just qualitative data, so we'll start with that and then move to quantitative data once we understand the model; the start of which can be seen directly below.

$$P[Y|X] \ = \frac{P[X|Y]\, P[Y]}{P[X]}$$

Y is a vector of correct classifications in the dataset and X is the matrix of related data; each row in X maps to an answer in the same row of Y. In the equation above, we apply Bayes Rule [1] to get the right-hand-side. To classify, we're going to calculate this for $Y_i \in G \ \forall i$ where G is the set of all possible classifications. $P[X]$ is a constant so we can actually drop it from the denominator since it will affect all calculations equally. To show this in the equation we use the symbol $\propto$ which means "proportional to".

$$P[Y|X] \propto P[X|Y]\, P[Y]$$

The model is not naive yet. To make it naive, we add an assumption/bias. We assume that all columns in X are independent. We do this because it greatly simplifies the training process and reduces the required amount of data. To understand, first let's look at the new model below. Note, going forward P will represent the number of columns in X.

$$P[Y|X] \propto P[Y] \prod_{i=1}^{P} P[X_i|Y]$$

Now examine the number of parameters required. Again, assume that all columns in X are categorical data (e.g. sizes). For both classifiers, we need to learn the probability of each class. For naive bayes, we need to learn $P[X_i|Y]$ for each Y. Say that the size of G is K and we have L_i categories for X_i. This gives us $K + \sum L_i K$ parameters. If we don't have the assumption of independence, then we need to learn $P[X_1, ..., X_P|Y_i] \ \forall i \in G$. We have to learn $K \prod_{i=1}^{P} L_i$ parameters. The amount of parameters is vastly more and what if the dataset doesn't cover a case? Also, note that the number of parameters can be slightly reduced by using $P(A) = 1 - P(\neg A)$. The full model is described in the equation below.

$$\text{NaiveBayes}(X) = \underset{Y \in G}{\arg\max}\, P[Y] \prod_{i=1}^{P} P[X_i|Y]$$

This model is great for categorical data but what if we receive continuous data (i.e. 3.14)? This is the part that I think is actually pretty cool. We can learn a one-dimensional distribution for continuous columns. For simplicity, let's only think about the Gaussian distribution [2]—NOTE: If you aren't familiar with the shape of the distribution, please view the footnote.

$$f_g(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}$$

[1] `https://en.wikipedia.org/wiki/Bayes%27_theorem`

[2] `https://en.wikipedia.org/wiki/Normal_distribution`

μ is the mean of the continuous data in the column and σ is the standard deviation. We can calculate both of these with the column vector in X easily. The distribution will output a number which be the largest at μ and 0 at $-\infty$ and ∞. The exact value at μ is harder to say but it will be larger given a small σ and smaller given a larger σ. The unfortunate result of this is that we lose the probability. Though, I suppose you could take the integral to find the area under the curve but that isn't necessary. Simply output whatever the distribution outputs.

$$\text{NaiveBayes}(X) = \arg\max_{Y in G} P[Y] \prod_{i=1}^{P} f_i[X_i|Y]$$

f_i is the function associated with each column i, meaning it will either be probabilistic or based on a distribution. And that is the naive bayes classifier for qualitative and quantitative data. It is a great baseline model that is simple to implement. It does not rely on a lot of data and scales well. It does have a problem that we touched on before for a bayes model without the assumption of independence: unseen input. One method to resolve this is add-one smoothing, though, there are many other smoothing options. We add one to the numerator and add the total number of categories tha there should be to the denominator. In our simple example of sizes, this is unlikely to be necessary. However, imagine using English words as input and it is far more likely to come across unseen input.

$$P[X|Y] = \frac{count(Y|X) + 1}{count(Y) + L_i}$$