We will start by examining the mathematics that justify the standard Multinomial Naïve Bayes model.

Before moving forward, we will define a series of variables and notations. Let x be the vector of training data points, y be the vector of possible classes to which data points can belong, z be an observed data point whose class we do not know, and $\phi(z)$ be the feature vector of binary values associated with data point z. Furthermore, let P(a) represent the probability of event a, and c(a) represent the count of a: the number of times a was observed in training. Lastly, let λ represent our Laplace Smoothing parameter.

We will start by examining simple probabilities:

$$\mathrm{c}(y_j) = \sum_{i=1}^{|x|} 1[x_i \in y_j] = \sum_{x_i:\, x_i \in y_j} 1$$

$$\mathrm{P}(y_j) = \frac{\mathrm{c}(y_j)}{|x|}$$

$$\mathrm{c}(y_j, \phi_k) = \sum_{i=1}^{|x|} 1[x_i \in \mathrm{y_j}, \phi(x_i)_\mathrm{k} = 1] = \sum_{x_i:\, x_i \in \mathrm{y_j}, \phi(x_i)_\mathrm{k}=1} 1$$

$$\mathrm{P}(y_j, \phi_k) = \frac{\mathrm{c}(\, y_j, \phi_k) + \lambda}{|x| + \lambda * |\phi|}$$

$$\mathrm{P}(\phi_k = \phi(z)_k | y_j) = \frac{\mathrm{P}(y_j, \phi_k = \phi(z)_k)}{\mathrm{P}(y_j)} = \frac{\left(\frac{(\mathrm{c}(\, y_j, \phi_k = \phi(z)_k) + \lambda)}{(|x| + \lambda * |\phi|)}\right)}{\left(\frac{\mathrm{c}(y_j)}{|x|}\right)}$$

$$= \frac{(\mathrm{c}(y_j, \phi_k = \phi(z)_k) + \lambda) * |x|}{(|x| + \lambda * |\phi|) * \mathrm{c}(y_j)}$$

By the Law of Total Probability:

$$\mathrm{P}(\phi(z)) = \sum_{j=1}^{|y|} \mathrm{P}(y_j) * \mathrm{P}(\phi(z)|y_j)$$

Finally, we will calculate the conditional probability of class y_j given observed data point z:

$$\mathrm{P}(y_j|z) \approx \mathrm{P}\left(y_j|\phi(z)\right) = \frac{\mathrm{P}(y_j) * \mathrm{P}(\phi(z)|y_j)}{\mathrm{P}(\phi(z))} = \frac{\mathrm{P}(y_j) * \mathrm{P}(\phi(z)|y_j)}{\sum_{\ell=1}^{|y|} \mathrm{P}(y_\ell) * \mathrm{P}(\phi(z)|y_\ell)}$$

By the Naïve Bayes assumption, assuming feature values are independent when conditioned upon classes y_j:

$$\mathrm{P}\big(\phi(z)|y_j\big) = \prod_{k=1}^{|\phi|} \mathrm{P}\big(\phi(z)_k|y_j\big)$$

$$\Longrightarrow \frac{\mathrm{P}\big(y_j\big) * \mathrm{P}\big(\phi(z)|y_j\big)}{\sum_{\ell=1}^{|y|} \mathrm{P}(y_\ell) * \mathrm{P}(\phi(z)|y_\ell)} \approx \frac{\mathrm{P}\big(y_j\big) * \prod_{k=1}^{|\phi|} \mathrm{P}\big(\phi_k = \phi(z)_k|y_j\big)}{\sum_{\ell=1}^{|y|} \mathrm{P}(y_\ell) * \prod_{k=1}^{|\phi|} \mathrm{P}\big(\phi_k = \phi(z)_k|y_\ell\big)}$$

$$= \frac{\left(\frac{\mathrm{c}\big(y_j\big)}{|x|}\right) * \prod_{k=1}^{|\phi|} \left(\frac{\big(\mathrm{c}\big(y_j, \phi_k = \phi(z)_k\big) + \lambda\big) * |x|}{(|x| + \lambda * |\phi|) * \mathrm{c}\big(y_j\big)}\right)}{\sum_{\ell=1}^{|y|} \left(\frac{\mathrm{c}(y_\ell)}{|x|}\right) * \prod_{k=1}^{|\phi|} \left(\frac{\big(\mathrm{c}\big(y_\ell, \phi_k = \phi(z)_k\big) + \lambda\big) * |x|}{(|x| + \lambda * |\phi|) * \mathrm{c}(y_\ell)}\right)}$$

$$= \frac{\mathrm{c}\big(y_j\big)^{|\phi|-1} * \prod_{k=1}^{|\phi|} \big(\mathrm{c}\big(y_j, \phi(z)_k\big) + \lambda\big)}{\sum_{\ell=1}^{|y|} \mathrm{c}(y_\ell)^{|\phi|-1} * \prod_{k=1}^{|\phi|} (\mathrm{c}(y_\ell, \phi(z)_k) + \lambda)}$$

$$\approx \mathrm{P}\big(y_j|z\big)$$

Now, we will move on to our examination of the slightly more complex Gaussian Naïve Bayes model. Let $\mathcal{N}_{jk}(\phi(z)_k)$ be the Gaussian probability density function (PDF) of ϕ_k given y_j. In other words:

$$\mathcal{N}_{jk}(\phi(z)_k) = \text{Gaussian PDF of } \phi_k|\, y_j$$

For an infinitesimally small ϵ, the probability that normally distributed feature $\phi(z)_k$ associated with observed data point z takes on a certain value given z belongs to class y_j is the following:

$$\mathrm{P}\big(\phi_k = \phi(z)_k\big|y_j\big) = \int_{\phi(z)_k - \epsilon/2}^{\phi(z)_k + \epsilon/2} \mathcal{N}_{jk}(\phi_k)\, d\phi_k = \epsilon * \mathcal{N}_{jk}(\phi(z)_k)$$

Note that because we chose ϵ to be infinitesimally small, this quantity will and should be zero, as is the probability of any continuous variable taking on a single value. However, as will become clear in a moment, leaving the expression in terms of ϵ will allow us to simplify expressions considered in the future to well-defined, non-zero values.

We will remember from our calculations regarding the Multinomial Naïve Bayes model that the following is true, given our model assumptions:

$$\mathrm{P}\big(y_j|z\big) \approx \mathrm{P}\left(y_j|\phi(z)\right) \approx \frac{\mathrm{P}\big(y_j\big) * \prod_{k=1}^{|\phi|} \mathrm{P}\big(\phi_k = \phi(z)_k|y_j\big)}{\sum_{\ell=1}^{|y|} \mathrm{P}(y_\ell) * \prod_{k=1}^{|\phi|} \mathrm{P}(\phi_k = \phi(z)_k|y_\ell)}$$

However, we now define $\mathrm{P}\big(\phi_k = \phi(z)_k | y_j\big)$ differently than before. As a result, our expression for the conditional probability of class y_j given observed data point z will simplify to the following.

$$\frac{\mathrm{P}\big(y_j\big) * \prod_{k=1}^{|\phi|} \mathrm{P}\big(\phi_k = \phi(z)_k | y_j\big)}{\sum_{\ell=1}^{|y|} \mathrm{P}(y_\ell) * \prod_{k=1}^{|\phi|} \mathrm{P}(\phi_k = \phi(z)_k | y_\ell)}$$

$$= \frac{\left(\frac{\mathrm{c}\big(y_j\big)}{|x|}\right) * \prod_{k=1}^{|\phi|} \Big(\epsilon * \ \mathcal{N}_{jk}(\phi(z)_k)\Big)}{\sum_{\ell=1}^{|y|} \left(\frac{\mathrm{c}(y_\ell)}{|x|}\right) * \prod_{k=1}^{|\phi|} \big(\epsilon * \ \mathcal{N}_{\ell k}(\phi(z)_k)\big)}$$

$$= \frac{\mathrm{c}\big(y_j\big) * \prod_{k=1}^{|\phi|} \mathcal{N}_{jk}(\phi(z)_k)}{\sum_{\ell=1}^{|y|} \mathrm{c}(y_\ell) * \prod_{k=1}^{|\phi|} \mathcal{N}_{\ell k}(\phi(z)_k)}$$

$$\approx \mathrm{P}\big(y_j | z\big)$$

Finally, we will examine the most generalized representation of the Naïve Bayes model possible. We will simultaneously incorporate categorical discrete features, ordered discrete features, normally distributed continuous features and arbitrarily distributed continuous features.

For discrete features, the conditional probability of class y_j given feature value $\phi(z)_k$ associated with observed data point z is simply the value of the probability mass function we model ϕ_k with. We denote the probability mass function of feature ϕ_k given class y_j as p_{jk}. As was previously discussed, this probability mass function takes on the following form for discrete binary features:

$$\mathrm{P}\big(\phi_k = \phi(z)_k | y_j\big) = p_{jk}(\phi(z)_k) = \frac{\big(\mathrm{c}\big(y_j, \phi_k = \phi(z)_k\big) + \lambda\big) * |x|}{(|x| + \lambda * |\phi|) * \mathrm{c}\big(y_j\big)}$$

For discrete features that do not take on binary values, we must learn the probability mass functions p_{jk} they conform to during training. First, we must choose the probability mass function we will model feature ϕ_k with. Subsequently, one typically uses maximum likelihood estimation (MLE) to learn the most likely parameterization the probability mass function modeling ϕ_k given the training data available. For example, the maximum likelihood estimate for the commonly used Poisson distribution's single parameter λ, not to be confused with the Laplace Smoother parameter λ, takes on the following form:

$$\lambda_{jk,MLE} = \frac{\sum_{x_i \in y_j} \phi(x_i)_k}{\mathrm{c}\big(y_j\big)}$$

For continuous valued features, the conditional probability of class y_j given feature value $\phi(z)_k$ associated with observed data point z is slightly more complicated. As mentioned

during our discussion of the Gaussian Naïve Bayes model, the probability that a normally distributed feature value takes on a given value is the following:

$$\mathrm{P}\left(\phi_k = \phi(z)_k \middle| y_j\right) = \int_{\phi(z)_k - \epsilon/2}^{\phi(z)_k + \epsilon/2} \mathcal{N}_{jk}(\phi_k)\, d\phi_k = \epsilon * \mathcal{N}_{jk}(\phi(z)_k)$$

Fortunately, this math remains identical regardless of the specific probability density function involved. We denote the probability density function of feature ϕ_k given class y_j as f_{jk}. Thus the probability that any continuous feature with probability density function f_{jk} takes on a given value is the following:

$$\mathrm{P}\left(\phi_k = \phi(z)_k \middle| y_j\right) = \int_{\phi(z)_k - \epsilon/2}^{\phi(z)_k + \epsilon/2} f_{jk}(\phi_k)\, d\phi_k = \epsilon * f_{jk}(\phi(z)_k)$$

As with discrete features, we must learn the probability density functions continuous features conform to during training, typically using maximum likelihood estimation. To provide an example of learning one such distribution from training data, we will examine the most common distribution used to model continuous feature values in the context of the Naïve Bayes model: the Gaussian distribution. The maximum likelihood estimate for its parameters μ and σ are simply the mean and standard deviation of the appropriate feature values encountered in training:

$$\mu_{jk,MLE} = \frac{\sum_{x_i \in y_j} \phi(x_i)_k}{\mathrm{c}(y_j)}$$

$$\sigma_{jk,MLE} = \frac{\sum_{x_i \in y_j} \left(\phi(x_i)_k - \mu_{jk,MLE}\right)^2}{\mathrm{c}(y_j)}$$

Having established our representations of the conditional probabilities of class y_j given both discrete and ordered feature values we can move on to calculating the conditional probability of class y_j given observed data point z. As noted previously:

$$\mathrm{P}(y_j|z) \approx \mathrm{P}\left(y_j|\phi(z)\right) \approx \frac{\mathrm{P}(y_j) * \prod_{k=1}^{|\phi|} \mathrm{P}\left(\phi_k = \phi(z)_k|y_j\right)}{\sum_{\ell=1}^{|y|} \mathrm{P}(y_\ell) * \prod_{k=1}^{|\phi|} \mathrm{P}(\phi_k = \phi(z)_k|y_\ell)}$$

Inserting our newly defined expressions for the conditional probabilities involved:

$$\frac{\mathrm{P}(y_j) * \prod_{k=1}^{|\phi|} \mathrm{P}\left(\phi_k = \phi(z)_k|y_j\right)}{\sum_{\ell=1}^{|y|} \mathrm{P}(y_\ell) * \prod_{k=1}^{|\phi|} \mathrm{P}(\phi_k = \phi(z)_k|y_\ell)}$$

$$= \frac{\mathrm{P}(y_j) * \prod_{k=1}^{|\phi|} \begin{cases} p_{jk}(\phi(z)_k|y_j) & if\ discrete \\ \epsilon * f_{jk}(\phi(z)_k|y_j) & otherwise \end{cases}}{\sum_{\ell=1}^{|y|} \mathrm{P}(y_\ell) * \prod_{k=1}^{|\phi|} \begin{cases} p_{jk}(\phi(z)_k|y_\ell) & if\ discrete \\ \epsilon * f_{jk}(\phi(z)_k|y_\ell) & otherwise \end{cases}}$$

$$= \frac{\mathrm{P}(y_j) * \left(\prod_{k:\phi_k \text{ discrete}} p_{jk}(\phi(z)_k|y_j)\right) * \left(\prod_{k:\phi_k \text{ continuous}} \epsilon * f_{jk}(\phi(z)_k|y_j)\right)}{\sum_{\ell=1}^{|y|} \mathrm{P}(y_\ell) * \left(\prod_{k:\phi_k \text{ discrete}} p_{jk}(\phi(z)_k|y_\ell)\right) * \left(\prod_{k:\phi_k \text{ continuous}} \epsilon * f_{jk}(\phi(z)_k|y_\ell)\right)}$$

$$= \frac{\mathrm{P}(y_j) * \left(\prod_{k:\phi_k \text{ discrete}} p_{jk}(\phi(z)_k|y_j)\right) * \left(\prod_{k:\phi_k \text{ continuous}} f_{jk}(\phi(z)_k|y_j)\right) * \epsilon^{\left(\sum_{k=1}^{|\phi|} 1[\phi_k \text{ continuous}]\right)}}{\sum_{\ell=1}^{|y|} \mathrm{P}(y_\ell) * \left(\prod_{k:\phi_k \text{ discrete}} p_{jk}(\phi(z)_k|y_\ell)\right) * \left(\prod_{k:\phi_k \text{ continuous}} f_{jk}(\phi(z)_k|y_\ell)\right) * \epsilon^{\left(\sum_{k=1}^{|\phi|} 1[\phi_k \text{ continuous}]\right)}}$$

$$= \frac{\mathrm{P}(y_j) * \left(\prod_{k:\phi_k \text{ discrete}} p_{jk}(\phi(z)_k|y_j)\right) * \left(\prod_{k:\phi_k \text{ continuous}} f_{jk}(\phi(z)_k|y_j)\right)}{\sum_{\ell=1}^{|y|} \mathrm{P}(y_\ell) * \left(\prod_{k:\phi_k \text{ discrete}} p_{jk}(\phi(z)_k|y_\ell)\right) * \left(\prod_{k:\phi_k \text{ continuous}} f_{jk}(\phi(z)_k|y_\ell)\right)}$$

$$\approx \mathrm{P}(y_j|z)$$

We have finally arrived at an expression for the conditional probability of class y_j given observed data point z that supports categorical, ordered discrete and ordered continuous feature values.