We would like to show that Multinomial Naive Bayes and the one-vs-all version are identical (assuming that the parameters are known). We can show this for binary classification, but we give a counter example to prove that they are not identical when the number of classes is three or greater.

Documents are generated from one of a set of classes, $C = \{1, 2, \ldots, m\}$. Given a class, a document is generated as a multinomial. The likelihood of a document in class $c$ is

$$p(d|c) = \prod_i \theta_{ci}^{f_i} = e^{\sum_i f_i \log \theta_{ci}}. \quad (1)$$

We assign to a document the label with the maximum likelihood.

$$l_{mnb}(d) = \arg \max_c \left[ \sum_i f_i \log \theta_{ci} \right] \quad (2)$$

The one-versus-all classifier, which we will denote as $l_{ova}$, uses the notion of the “complement class,” which we denote by $\hat{c}$. The complement class is a fictitious class, effectively a composite class of all classes but $c$. The multinomial parameter for word $i$ in the complement class $\hat{c}$ is the average of the parameters in the classes other than $c$ $(C \setminus c)$,

$$\theta_{ci} = \frac{1}{m-1} \sum_{k \in C \setminus c} \theta_{ki} \quad (3)$$

The classification rule for the one-vs-all classifier is

$$l_{ova}(d) = \arg \max_c \left[ \sum_i f_i (\log \theta_{ci} - \log \theta_{\hat{c}i}) \right]. \quad (4)$$

In the case of binary classification $(m = 2)$, we can show that $l_{mnb}$ and $l_{ova}$
are identical.

\[ l_{\text{ova}}(d) = \arg \max_{c \in \{1, 2\}} \left[ \sum_i f_i \left( \log \theta_{ci} - \log \theta_{(3-c)i} \right) \right] \]

\[ = \arg \max_{c \in \{1, 2\}} \left[ \sum_i f_i \left( \log \theta_{ci} - \log \theta_{(3-c)i} \right) + \sum_i f_i \left( \log \theta_{1i} + \log \theta_{2i} \right) \right] \]

\[ = \arg \max_{c \in \{1, 2\}} \left[ 2 \sum_i f_i \log \theta_{ci} \right] = l_{\text{mnb}}(d). \]

We can show that it is impossible to show the equivalence for multiple class classification \((m \geq 3)\) by producing an example where \(l_{\text{mnb}}(d) \neq l_{\text{ova}}(d)\). Consider a three class example with three words. Let

\[ \tilde{\theta}_1 = (\theta_{11}, \theta_{12}, \theta_{13}) = (13/28, 13/28, 2/28) \]

\[ \tilde{\theta}_2 = (6/28, 6/28, 16/28) \]

\[ \tilde{\theta}_3 = (2/28, 2/28, 24/28) \]

Let the document be composed of one each of the three words. In other words, let \(f = (f_1, f_2, f_3) = (1, 1, 1)\). The MNB scores are

\[ \text{mnb-score}_1 = 2 \log 13 + \log 2 - 3 \log 28 \approx -4.17, \]

\[ \text{mnb-score}_2 = 2 \log 6 + \log 16 - 3 \log 28 \approx -3.64, \]

\[ \text{mnb-score}_3 = 2 \log 2 + \log 24 - 3 \log 28 \approx -5.43. \]

The OVA scores are

\[ \text{ova-score}_1 = 2 \log 13 + \log 2 - 2 \log 4 - \log 20 \approx 0.05, \]

\[ \text{ova-score}_2 = 2 \log 6 + \log 16 - 2 \log 7.5 - \log 13 \approx -0.24, \]

\[ \text{ova-score}_3 = 2 \log 2 + \log 24 - 2 \log 9.5 - \log 9 \approx -2.14. \]

Hence, \(l_{\text{mnb}}(d) = 2\) but \(l_{\text{ova}}(d) = 1\). MNB and the one-versus-all version of it are not identical\(^1\).

\(^1\)Thanks to Jonathan Gough for pointing out a miscalculation in my originally-published example.