Document classification with the multinomial model

Having numbered the words in English 1 to M, let X_i be the number of occurrences of word i in some document. The multinomial model for word counts is

$$p\,(\,X\,|\,\alpha\,) = \frac{(\sum_{i=1}^{M} X_i)!}{\prod_{i=1}^{M} X_i!} \prod_{i=1}^{M} \alpha_i^{X_i}$$

where $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_M)$ is a parameter vector satisfying $\sum_{i=1}^{M} \alpha_i = 1$.

To train the multinomial model on a document with word count vector $X = (X_1, X_2, \ldots, X_M)$, you can just take the maximum likelihood estimate for α. If you have several word count vectors $X^{(1)}, X^{(2)}, \ldots, X^{(N)}$ (i.e., several documents) to train on, take the maximum of the joint likelihood

$$p(X^{(1)}, X^{(2)}, \ldots, X^{(N)}|\alpha) = p(X^{(1)}|\alpha) \cdot p(X^{(2)}|\alpha) \cdot \ldots \cdot p(X^{(N)}|\alpha).$$

You can use multinomial models for classification by learning a different parameter vector $\hat{\alpha}^{[1]}, \hat{\alpha}^{[2]}, \ldots$ for each class. When given a testing document $X = (X_1, X_2, \ldots, X_M)$, you compute the likelihood

$$p(X|\hat{\alpha}^{[j]}) = \frac{(\sum_{i=1}^{M} X_i)!}{\prod_{i=1}^{M} X_i!} \prod_{i=1}^{M} \left(\hat{\alpha}_i^{[j]}\right)^{X_i}$$

for each class $j = 1, 2, \ldots$ and classify X to the class that gives highest likelihood.

Words that do not appear in the training set at all can cause problems. To avoid this, it customary to increase all word counts by one. This actually corresponds to using a Dirichlet prior on α.

Another possible issue is that the factorials might get far too big to compute on a computer. Note, however, that this combinatorial factor is independent of α.