Bayesian vs. Frequentist Point Estimation II

Modeling text documents with the multinomial "Naive Bayes" model

One of the simplest but still usable models for text is to imagine that each word of a document $\mathcal{D}$ is chosen independently from some fixed distribution over all words of the English language characteristic of the author or the topic. Assuming that the probability of choosing the i'th word in the language is θ_i, the probability of having x_1 occurrences of word 1, x_2 occurrences of word 2, etc. in a given document is then

$$p(\boldsymbol{x}) = \frac{n!}{x_1!\,x_2!\ldots x_k!}\,\theta_1^{x_1}\,\theta_2^{x_2}\ldots\theta_k^{x_k}, \tag{1}$$

known as the **multinomial** distribution with parameters $\theta_1, \theta_2, \ldots, \theta_k$. The parameters must satisfy $\sum_{i=1}^k \theta_i = 1$.

Typically, to figure out whether some document was written by author A or author B, one would fit a multinomial model to documents that are certain to have been written by A (i.e., we would estimate the corresponding parameters $\theta_1, \theta_2, \ldots, \theta_k$), fit a different model with parameters $\theta'_1, \theta'_2, \ldots, \theta'_k$ to the works of B, and then classify a new document based on whether it has higher likelihood under model A or model B. The question is how to find the $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_k)$ parameters for both authors.

The frequentist way

Just as in estimating the speed of light, a natural place to start is to find the MLE for (1). Taking the logarithm and ignoring the normalizer, one can write the log-likelihood as

$$\log \ell(\boldsymbol{\theta}) = \text{const} + \sum_{i=1}^k x_i \log \theta_i.$$

We want to maximize this, subject to the constraints $\sum_{i=1}^k \theta_i = 1$.

The standard way to solve optimization problems of this form is to use **Lagrange multipliers**. This involves defining

$$\mathcal{L}(\boldsymbol{\theta}, \lambda) = \left[\sum_{i=1}^k x_i \log \theta_i\right] - \lambda \sum_{i=1}^k \theta_i,$$

and then minimizing this expression for fixed λ, giving

$$0 = \frac{\partial}{\partial \theta_i}\mathcal{L}(\boldsymbol{\theta}, \lambda) = \frac{x_i}{\theta_i} - \lambda \qquad \Longrightarrow \qquad \theta_i = \frac{x_i}{\lambda}.$$

Finally, λ must be set so as to satisfy the original constraint:

$$\sum_{i=1}^{k} \frac{x_i}{\lambda} = 1 \qquad \Longrightarrow \qquad \lambda = \sum_{i=1}^{k} x_i,$$

from which we conclude that the MLE for (1) is

$$\widehat{\boldsymbol{\theta}} = (\widehat{\theta}_1, \widehat{\theta}_2, \ldots, \widehat{\theta}_k) \qquad \text{where} \qquad \widehat{\theta}_i = \frac{x_i}{x_1 + x_2 + \ldots + x_k}. \tag{2}$$

One appealing feature of this estimator is that if we have a sequence of training documents $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_m$ and corresponding likelihood

$$\ell(\theta) = \prod_{t=1}^{m} p(\boldsymbol{x}_t|\boldsymbol{\theta}),$$

the MLE of the entire training set will be

$$\theta_i = \frac{\sum_{t=1}^{m} [\boldsymbol{x}_t]_i}{\sum_{j=1}^{k} \sum_{t=1}^{m} [\boldsymbol{x}_t]_j},$$

which is the same as if we had trained on a single document which is the concatenation of $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_m$.

The down side of the MLE is that if we plug it back into the multinomial and try and compute the likelihood of a new document $\mathcal{D}'$ (with word count vector $\boldsymbol{x}'$) under the resulting model, as soon as $\mathcal{D}'$ has at least one word that was not present in $\mathcal{D}$, the likelihood will be zero! This is clearly very dangerous.

However, the frequentist is unfazed: since in his book virtually anything can serve as an estimator, he suggest simply adding a constant γ to all our counts:

$$\tilde{\boldsymbol{\theta}} = (\tilde{\theta}_1, \tilde{\theta}_2, \ldots, \tilde{\theta}_k) \qquad \text{where} \qquad \tilde{\theta}_i = \frac{x_i + \gamma}{x_1 + x_2 + \ldots + x_k + k\gamma}. \tag{3}$$

Granted, $\tilde{\boldsymbol{\theta}}$ is not an unbiased estimator anymore, but that's not the end of the world. The resulting multinomial might still be perfectly fine for classifying documents.

In fact, this approach using **pseudocounts**, while aptly illustrating the arbitrariness of statistics, is not as crazy as it seems at first sight and is widely used in practical settings. Of course it leaves open the question of what γ should be, which will probably have to be set by cross-validation. A thorough frequentist would also compute confidence intervals for his estimators, but we shall forgo that for now.

The Bayesian way

As always, a Bayesian has to start with writing down a prior. For this problem it is not so intuitive what the prior should be, but from the purely pragmatic point of view of making the calculations easy it is very useful to choose the prior to be a Dirichlet distribution with parameters $\alpha_1, \alpha_2, \ldots, \alpha_k$:

$$p(\boldsymbol{\theta}) = \frac{\Gamma(\alpha_1 + \alpha_2 + \ldots + \alpha_k)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\ldots\Gamma(\alpha_k)}\, \theta_1^{\alpha_1 - 1}\theta_2^{\alpha_2 - 1}\ldots\theta_k^{\alpha_k - 1}, \tag{4}$$

where $\sum_{i=1}^k \theta_i = 1$ (otherwise $p(\boldsymbol{\theta}) = 0$). Here the Γ function is a kind of generalized factorial obeying $\Gamma(n) = (n-1)!$, but we don't need to worry about it much about, since the normalizer is not going to affect our calculations, anyway.

The important thing is that if we multiply a Dirichlet$(\alpha_1, \alpha_2, \ldots, \alpha_k)$ distribution (in the variables $\theta_1, \theta_2, \ldots, \theta_k$) with a Multinomial$(\theta_1, \theta_2, \ldots, \theta_k)$ distribution (in the variables $x_1, x_2, \ldots, x_k$) we get something which, up to normalization, looks just like a Dirichlet$(\alpha_1 + x_1, \alpha_2 + x_2, \ldots, \alpha_k + x_k)$ distribution:

$$\frac{n!}{x_1!\, x_2! \ldots x_k!}\, \theta_1^{x_1}\, \theta_2^{x_2} \ldots \theta_k^{x_k}\, \frac{\Gamma(\alpha_1 + \alpha_2 + \ldots + \alpha_k)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\ldots\Gamma(\alpha_k)}\, \theta_1^{\alpha_1 - 1}\theta_2^{\alpha_2 - 1}\ldots\theta_k^{\alpha_k - 1} \propto$$
$$\frac{\Gamma(\alpha_1 + x_1 + \alpha_2 + x_2 + \ldots + \alpha_k + x_k)}{\Gamma(\alpha_1 + x_1)\Gamma(\alpha_2 + x_2)\ldots\Gamma(\alpha_k + x_k)}\, \theta_1^{\alpha_1 + x_1 - 1}\theta_2^{\alpha_2 + x_2 - 1}\ldots\theta_k^{\alpha_k + x_k - 1}$$

In particular, if our prior $p(\boldsymbol{\theta})$ is Dirichlet, and the likelihood $p(\boldsymbol{x}|\boldsymbol{\theta})$ is multinomial, as in (1), then our posterior distribution

$$p(\boldsymbol{\theta}|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\boldsymbol{\theta})\, p(\boldsymbol{\theta})}{p(\boldsymbol{x})}$$

will be

$$p(\boldsymbol{\theta}|\boldsymbol{x}) = \text{Dirichlet}(\alpha_1 + x_1, \alpha_2 + x_2, \ldots, \alpha_k + x_k).$$

If we are to be mercenary and look for the MAP estimator, using the same Lagrange multiplier technique as in the frequentist case, we find that

$$\widehat{\boldsymbol{\theta}} = (\theta_1, \theta_2, \ldots, \theta_k) \qquad \text{where} \qquad \widehat{\theta_i} = \frac{\alpha_i + \theta_i - 1}{\sum_{j=1}^k \alpha_j + \theta_j - 1}\,.$$

If, on the other hand, we are to look for the mean of the posterior, we get (calculation omitted)

$$\mathbb{E}\boldsymbol{\theta} = (\mathbb{E}\theta_1, \mathbb{E}\theta_2, \ldots, \mathbb{E}\theta_k) \qquad \text{where} \qquad \mathbb{E}\theta_i = \frac{\alpha_i + x_i}{\sum_{j=1}^k \alpha_j + x_j}\,.$$

The astute reader will note that these estimators are *exactly* the same as the frequentist's estimators if we just set $\alpha_i = \gamma + 1$ or $\alpha_i = \gamma$. This just proves that people who say very different things, when no one is looking, are often found to do exactly the same thing.

Of course a real Bayesian would not just use a point estimate for $\boldsymbol{\theta}$ to compute the probability of a new document $\mathcal{D}'$, but would use the full posterior:

$$p(\mathcal{D}'|\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_m) = \int_\Delta p(\mathcal{D}'|\boldsymbol{\theta})\, p(\boldsymbol{\theta}|\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_m)\, d\boldsymbol{\theta},$$

or, in terms of word count vectors,

$$p(\boldsymbol{x}'|\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m) = \int_\Delta p(\boldsymbol{x}'|\boldsymbol{\theta})\, p(\boldsymbol{\theta}|\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m)\, d\boldsymbol{\theta}, \tag{5}$$

where, in both cases Δ is to remind us that we are only integrating over the simplex given by

$$\theta_i \geq 0, \qquad \sum_{i=1}^{k} \theta_i = 1.$$

Fortunately, the integral in (5) is not as scary as it seems at first sight. The thing to note that if $p(\boldsymbol{\theta}|\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m) = \text{Dirichlet}(\beta_1, \beta_2, \ldots, \beta_k)$, then the integrand $p(\boldsymbol{x}'|\boldsymbol{\theta})\, p(\boldsymbol{\theta}|\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m)$ is

$$\frac{n!}{x_1'!\, x_2'! \ldots x_k'!} \frac{\Gamma(\beta_1 + \beta_2 + \ldots + \beta_k)}{\Gamma(\beta_1)\Gamma(\beta_2)\ldots\Gamma(\beta_k)}\, \theta_1^{\beta_1 + x_1' - 1} \theta_2^{\beta_2 + x_2' - 1} \ldots \theta_k^{\beta_k + x_k' - 1},$$

whereas we know that since the Dirichlet is well normalized,

$$\int_\Delta \frac{\Gamma(\beta_1 + \beta_2 + \ldots + \beta_k + x_1 + x_2 + \ldots + x_k)}{\Gamma(\beta_1 + x_1)\Gamma(\beta_2 + x_2)\ldots\Gamma(\beta_k + x_k)}\, \theta_1^{\beta_1 + x_1' - 1} \theta_2^{\beta_2 + x_2' - 1} \ldots \theta_k^{\beta_k + x_k' - 1} = 1.$$

Thus, we conclude that

$$p(\boldsymbol{x}'|\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m) = \frac{n!}{\prod_{i=1}^k x_i'!} \frac{\Gamma(\sum_{i=1}^k \beta_i)}{\prod_{i=1}^k \Gamma(\beta_i)} \frac{\prod_{i=1}^k \Gamma(\beta_i + x_i')}{\Gamma(\sum_{i=1}^k \beta_i + x_i')}\,.$$