10-602: Statistical Approaches to Learning and Discovery

Problem Set 3 Solutions

---

1. The first model is the multivariate Bernoulli model, which only concerns the presence or absence of a word in a document. A word w is present with fixed probability b_w, independent of all other words. If we represent a document by the vector $[x_1, ..., x_W]$, where $x_w = 1$ if word w is present, then the probability of that document is

$$p(\mathbf{x}|\mathbf{b}) = \prod_{w=1}^{W} b_w^{x_w}(1 - b_w)^{1-x_w} \tag{1}$$

Given several documents $D = [\mathbf{x}_1, ..., \mathbf{x}_n]$ in the same class, the authors use the estimate

$$\hat{b}_w = \frac{1 + \sum_{i=1}^{n} x_{iw}}{2 + n} \tag{2}$$

This is done separately for each class since the $\mathbf{b}$ parameters are assumed independent across classes.

This estimate could be explained in several ways, but some explanations are better than others. We could, for example, explain it as a MAP estimate under a Beta $(2, 2)$ prior for b_w. But there is little justification for this. MAP arises from a zero-one loss function, which doesn't make sense in this context, and there is no prior information which would suggest a Beta $(2, 2)$ distribution. The authors say they use a "Laplacean prior" which implies a uniform Beta $(1, 1)$ distribution. Another explanation is that the estimate is the posterior mean. This implies a squared error function on b, which is more plausible than zero-one but not great. Estimating $\hat{b} = 0.11$ when the true value is $b = 0.01$ is much worse than estimating $\hat{b} = 0.55$ when the true value is $b = 0.45$, even though the squared errors are the same. The best explanation is that the estimate is the predictive density for a new document: $\hat{b}_w = p(x_w = 1|D)$. This estimate minimizes the expected loss under the loss function $l(\hat{b}, b) = b \log \frac{b}{\hat{b}}$, which measures differences on a logarithmic scale.

The second model is the multinomial model, which concerns the number of times a word occurs in a document. The kth word of a document is word w with fixed probability p_w, independent of the rest of the document. If we represent a document by the vector $[x_1, ..., x_W]$, where x_w is the count of word w, then the probability of that document is

$$p(\mathbf{x}|\mathbf{p}) = \prod_{w=1}^{W} p_w^{x_w} \tag{3}$$

Given several documents $D = [\mathbf{x}_1, ..., \mathbf{x}_n]$ in the same class, the authors use the estimate

$$\hat{p}_w = \frac{1 + \sum_{i=1}^{n} x_{iw}}{W + \sum_{iw} x_{iw}} \tag{4}$$

This is done separately for each class since the $\mathbf{p}$ parameters are assumed independent across classes. With a uniform Dirichlet $(1, ..., 1)$ prior, this is the predictive density for a new word.

For both models, there is also a probability that a random document would be in class c, which is written θ_c. The authors use a maximum-likelihood estimate and do not mention a prior. However, a reasonable choice would be Beta $(1, 1)$.

It is implied in the paper that the multivariate Bernoulli model is somehow a special case of the multinomial, because it uses less information about the document. That is not true. The multivariate Bernoulli can capture the idea that a document must contain some occurrence of w, without requiring that w occurs many times. It also implicitly models length, since appearance probability increases with a longer document.

2. (a) Given parameter estimates b_{cw} for the multivariate Bernoulli model, a document is classified according to the class which gives it the highest probability, or equivalently highest log-probability:

$$\log p(\mathbf{x}|\mathbf{b}_c) = \sum_{w=1}^{W} x_w \log \frac{b_{cw}}{1 - b_{cw}} + \log(1 - b_{cw}) \tag{5}$$

Since this function is linear in x_w, it is a linear classifier. Word w will be ignored if and only if it has equal weight in each class, i.e. $\log \frac{b_{cw}}{1-b_{cw}}$ is the same for all c, which means the appearance probability b_{cw} is the same for all c. Note that it is not sufficient to make the weight zero, i.e. $b_{cw} = 0.5$, in just one class, nor is it necessary to make $b_{cw} = 0.5$ in all classes.

For the multinomial model, the log-probability is

$$\log p(\mathbf{x}|\mathbf{p}_c) = \sum_{w=1}^{W} x_w \log p_{cw} \tag{6}$$

which also leads to a linear classifier. Word w will be ignored if and only if p_{cw} is the same for all c. It is not sufficient if $p_{cw} = 1$ in one class, nor is it necessary to make $p_{cw} = 1$ in all classes.

(b) Let "very" be word 1. If the appearance probabilities for class 1 are $\{b_1, ..., b_W\}$, then the appearance probabilities for class 2 will be $\{b'_1, b_2, ..., b_W\}$ where $b'_1 \geq b_1$. According to the condition derived in part (a), every word except "very" will be ignored by a multivariate Bernoulli classifier using the true parameter values. In the special case where all documents to which "very" was added already had an occurrence of "very", then the appearance probabilities are unchanged and "very" would be ignored too.

For the multinomial model, suppose class 2 has a total of m extra counts of "very". If the word probabilities for class 1 are

$$p_w = \frac{\sum_{i=1}^{n} x_{iw}}{\sum_{iw} x_{iw}} \tag{7}$$

Then class 2 has word probabilities

$$p'_1 = \frac{m + \sum_{i=1}^{n} x_{i1}}{m + \sum_{iw} x_{iw}} \tag{8}$$

$$p'_w = \frac{\sum_{i=1}^{n} x_{iw}}{m + \sum_{iw} x_{iw}} \qquad w \neq 1 \tag{9}$$

If we define $a = 1/(m + \sum_{iw} x_{iw})$, then we can write the class 2 probabilities succinctly as $\{ap_1 + (1 - a), ap_2, ..., ap_W\}$. Since all probabilities have changed, no words will be ignored by a multinomial classifier using the true parameter values.

(c) Let class 1 have n_1 documents and class 2 have n_2 documents. The "Bayes-optimal" estimate for class 1 is

$$b_{1w} = \frac{1+n_1}{2+n_1} \tag{10}$$

and for class 2 is

$$b_{2w} = \frac{1+n_2}{2+n_2} \tag{11}$$

Since $n_1 \neq n_2$, these are not equal and word w will not be ignored.

(d) For the multivariate Bernoulli model, a word is ignored if its appearance probability is the same in all classes. It is typical for document classes to differ only in the introduction or omission of certain words. The remaining words therefore are likely to have the same appearance probability in all classes, and should be ignored. However, the estimated probabilities will typically not be equal between the two classes, even if the appearance counts are the same (part c). By omitting a word from the classifier we force the estimate of b_{cw} to be the same for all classes. Therefore feature selection helps by encouraging the estimated parameter values to resemble likely true parameter values. The same effect could be achieved by using a prior distribution on b_{cw} which favors equality across c—in fact that is one way to interpret feature selection.

If you do not believe that any words have $b_{1w} = b_{2w}$, then it is hard to motivate using feature selection. If $b_{1w} \neq b_{2w}$, then no matter how small the difference is, it helps to use word w. The only positive effect feature selection could have is to avoid situations where the estimated difference is larger than it should be—essentially trading one kind of estimation error with another. In this case, you should use a prior distribution which reflects your belief that $b_{1w} \approx b_{2w}$.

For the multinomial model, a word is ignored if its probability is the same in all classes. This is not as likely to be true, because when words are added or removed all word probabilities change. For the probability of one word to remain the same, these additions and removals would have to exactly balance. For this reason, feature selection is less likely to be useful for the multinomial model.

3. (a) For the multivariate Bernoulli model, the paper evaluates a feature x_w by its mutual information with the class variable c. This is another case where some explanations are better than others. One interpretation is a test for independence in the following 2×2 table (for 2 classes):

	$x_w = 1$	$x_w = 0$
c = 1	n_{1w}	$n_1 - n_{1w}$
c = 2	n_{2w}	$n_2 - n_{2w}$

where n_{cw} is the number of times word w appears in a document in class c, and n_c is the number of documents in class c. The mutual information score can be replaced by an exact test. If we use a uniform prior on the cell probabilities, then the model likelihood for dependence between c and x_w is

$$\frac{\Gamma(4)}{\Gamma(4+\sum_c n_c)} \prod_c \Gamma(n_{cw}+1)\Gamma(n_c - n_{cw}+1) \tag{12}$$

Under independence of c and x_w, the parameter b_w is the same for both classes. With a uniform prior on the marginal probabilities, the model likelihood for independence is

$$\frac{\Gamma(2)^2}{\Gamma(2+\sum_c n_c)^2}\Gamma(n_1+1)\Gamma(n_2+1)\Gamma(n_{1w}+n_{2w}+1)\Gamma(n_1-n_{1w}+n_2-n_{2w}+1) \tag{13}$$

The score for the word would be the ratio of these two.

This explanation has the slight problem that it uses priors which are inconsistent with problem 1. In problem 1, we said that b_w had a uniform prior and θ_c had a uniform prior. The probability of $(c = 1, x_w = 1)$, being in class 1 and drawing word w given class 1, is $\theta_1 b_{1w}$, whose distribution is the product of two uniform random variables. Therefore the cell probabilities cannot be uniform. Changing the test to use the correct prior is easy. The probability of a labeled document set $D = \{(\mathbf{x}_1, c_2), ..., (\mathbf{x}_n, c_n)\}$ is

$$p(D|\mathbf{b}_c, \theta_c) = \prod_{i=1}^{n} p(c_i)p(\mathbf{x}_i|\mathbf{b}_{c_i}) = \prod_i \theta_{c_i} \prod_w b_{c_iw}^{x_{iw}}(1 - b_{c_iw})^{1-x_{iw}} \tag{14}$$

Focus on word w. Under the assumption that b_{cw} is different for the two classes, we get the likelihood (for some constant Z)

$$L(b_{1w}, b_{2w}) = Z\ b_{1w}^{n_{1w}}(1 - b_{1w})^{n_1 - n_{1w}} b_{2w}^{n_{2w}}(1 - b_{2w})^{n_2 - n_{2w}} \tag{15}$$

Integrating out (b_{1w}, b_{2w}) under the uniform prior gives the likelihood of dependence:

$$p(D|\text{dep}) = Z\ \frac{\Gamma(2)}{\Gamma(2 + n_1)}\Gamma(n_{1w} + 1)\Gamma(n_1 - n_{1w} + 1)\frac{\Gamma(2)}{\Gamma(2 + n_2)}\Gamma(n_{2w} + 1)\Gamma(n_2 - n_{2w} + 1) \tag{16}$$

which is similar to, but not identical to (12). Under the assumption that $b_{1w} = b_{2w}$, we get the likelihood

$$L(b_w) = Z\ b_w^{n_{1w}+n_{2w}}(1 - b_w)^{n_1 - n_{1w} + n_2 - n_{2w}} \tag{17}$$

whose integral is the likelihood of independence:

$$p(D|\text{indep}) = Z\ \frac{\Gamma(2)}{\Gamma(2 + n_1 + n_2)}\Gamma(n_{1w} + n_{2w} + 1)\Gamma(n_1 - n_{1w} + n_2 - n_{2w} + 1) \tag{18}$$

In the ratio, Z cancels out which means the test is irrespective of the parameter values for other words. Despite the different prior, this ratio test differs from the above ratio test only by a constant, so they are the same as far as word ordering is concerned.

(b) For the multinomial model, the mutual information score described in the paper could be interpreted as a test for independence in a similar table:

	w	not w
$c = 1$	n_{1w}	$n_1 - n_{1w}$
$c = 2$	n_{2w}	$n_2 - n_{2w}$

Here we redefine n_{cw} to be the number of times word w occurs in all text of class c and n_c to be the length in words of class c. But again we have to be careful about priors: it is not correct to assume uniformly distributed cell probabilities. For one thing, the probability of "not w" is much higher than the probability of w. Secondly, the marginals (n_1, n_2) are not necessarily uniformly distributed—it depends on the document lengths in the two classes, which we are not modeling. To get a correct test, start with the probability of the data set:

$$p(D|\mathbf{p}_c, \theta_c) = \prod_{i=1}^{n} p(c_i)p(\mathbf{x}_i|\mathbf{p}_{c_i}) = \prod_i \theta_{c_i} \prod_w p_{c_iw}^{x_{iw}} \tag{19}$$

The tricky part is that we want to integrate out p_{cw} for one word only, not the entire $\mathbf{p}$ vector, because we want the test to work irrespective of the other word probabilities. We can do this by reparameterizing the multinomial as follows:

$$p(w|c) = \begin{cases} p_{c1} & \text{if } w = 1 \\ (1-p_{c1})q_{cw} & \text{otherwise} \end{cases} \tag{20}$$

$$\text{where } \sum_{w=2}^{W} q_{cw} = 1 \tag{21}$$

Here q_{cw} is a probability distribution on the words other than word 1. According to problem 1, we have a Dirichlet $(1, ..., 1)$ distribution on the probability vector $p_{c1}, ..., p_{cW}$, which means the marginal distribution for any single p_{cw} is Beta $(1, W-1)$. Thus we should take

$$p_{c1} \sim \text{Beta}(1, W-1) \tag{22}$$

$$q_{cw} \sim \text{Dirichlet}(1, ..., 1) \tag{23}$$

The likelihood in this new parameterization is: (for some constant Z)

$$L(p_{11}, p_{21}) = Z\, p_{11}^{n_{11}}(1-p_{11})^{n_1-n_{11}} p_{21}^{n_{21}}(1-p_{21})^{n_2-n_{21}} \tag{24}$$

Integrating out (p_{11}, p_{21}) under the Beta prior gives the likelihood of dependence:

$$p(D|\text{dep}) = Z\, \frac{\Gamma(W)}{\Gamma(W+n_1)}\Gamma(n_{11}+1)\frac{\Gamma(n_1-n_{11}+W-1)}{\Gamma(W-1)}\frac{\Gamma(W)}{\Gamma(W+n_2)}\Gamma(n_{21}+1)\frac{\Gamma(n_2-n_{21}+W-1)}{\Gamma(W-1)} \tag{25}$$

Under the assumption that $p_{11} = p_{21}$, we get the likelihood

$$L(p_1) = Z\, p_1^{n_{11}+n_{21}}(1-p_1)^{n_1-n_{11}+n_2-n_{21}} \tag{26}$$

whose integral is the likelihood of independence:

$$p(D|\text{indep}) = Z\, \frac{\Gamma(W)}{\Gamma(W+n_1+n_2)}\Gamma(n_{11}+n_{21}+1)\frac{\Gamma(n_1-n_{11}+n_2-n_{21}+W-1)}{\Gamma(W-1)} \tag{27}$$

The ratio of these is the score for word 1, and similarly for any word. Up to a constant, it resembles an independence test for the 2×2 table, using a nonuniform prior.

4. Using the exact test derived in the previous problem, the top words are:

Bernoulli	Multinomial
god	god
optilink	jesus
jesus	president
cramer	stephanopoulos
clayton	objective
christian	bible
dwyer	christian
bible	cramer
government	myers
christians	christ

The top ten ordering for the multinomial model is unchanged with a Beta $(1,1)$ prior instead of the correct $(1, W-1)$ prior. This may be because we are only using the 5586 most frequent words in the dataset; words with small count in both classes were already pruned.

Note that Clayton Cramer (`cramer@optilink.com`) was a frequent poster to talk.politics.misc, as was Frank O'Dwyer to talk.religion.misc. George Stephanopoulos and Dee Dee Myers were involved in most White House press briefings.

Since both models lead to a linear classifier, it is tempting to consider the linear weight on a word as a measure of relevance. However, this gives quite different results than the statistical test. In particular, it favors words which appear in only one class, regardless of how rare they are. For example, the word "hens" is ranked higher by this method than "god", because "hens" only appeared in religion whereas "god" was used in politics and religion.

5. The only requirement for this problem is that the proposal be a valid probability distribution different from the multivariate Bernoulli or multinomial. It is not sufficient to simply suggest a different classifier such as a neural network or suggest a new set of features like word pairs, if there is no associated probability model.

   One extension which was already hinted at by the authors is to model the dependence of document length on class, e.g. with a Poisson distribution:

$$p(L|c) \sim Po(\lambda_c) = e^{\lambda_c} \frac{\lambda_c^L}{L!} \tag{28}$$

   This does not help for politics vs. religion, however, because the distribution of document lengths is nearly equal.

   Another extension is to use a multivariate Poisson instead of multivariate Bernoulli:

$$p(\mathbf{x}|c) \sim \prod_{w=1}^{W} Po(x_w; \lambda_{cw}) = \prod_{w=1}^{W} e^{\lambda_{cw}} \frac{\lambda_{cw}^{x_w}}{x_w!} \tag{29}$$

   Like the multinomial, the full count of each word is used, but each document counts as a single event. Interestingly, this extension is the same as the first: it is equivalent to using a multinomial model coupled with a Poisson distribution on document length. If each word count is Poisson, then the total count is also Poisson, with mean $\lambda_c = \sum_w \lambda_{cw}$. Conditional on the total, the word counts have a multinomial distribution with probability vector λ_{cw}/λ_c. So this extension will also not help for politics vs. religion.

   A third approach is to model dependencies between words. This could be done in several ways. For example, we could retain the bag-of-words approach and model the count of word w conditional on the counts of related words. Or we could model the document as a first-order Markov chain, with the current word choice dependent on the previous word choice. The second approach entails a significant increase in the number of parameters of the model, which means special care will be needed in specifying a reasonable prior. A simple non-informative prior as used with the multinomial will probably not give good results.

   Finally, instead of modeling $p(\mathbf{x}|c)$, we could learn a direct model for $p(c|\mathbf{x})$. It was mentioned that the multivariate Bernoulli and multinomial models lead to linear classifiers. We can generalize them both by assuming $p(c = 1|\mathbf{x}) = \sigma(\mathbf{w}^{\mathrm{T}}\mathbf{x})$, for any weight vector $\mathbf{w}$. Here $\sigma(z)$ is any monotonic function which maps $\Re \to (0,1)$, such as $\sigma(z) = 1/(1+e^{-z})$. The parameter vector $\mathbf{w}$ can be learned by maximum likelihood or Bayesian techniques.