Explanation of `dirichlet-topic.pl`

Graham Neubig

August 3, 2010

1 Introduction

This is a description of the math behind the script `dirichlet-topic.pl`.

2 Problem Definition

Let's say we have a collection of topics $T = t_1, \ldots, t_I$, each described by its word counts $W_i = w_{i1}, \ldots, w_{iJ}$, where $w_{ij} = k$ indicates that word j appears in topic i k times. We will call the collection of all topic counts (the corpus) $\mathcal{W}$.

Our goal is to estimate reasonable posterior probabilities of topics given each word $P(t_i|w_j)$. The most obvious way to do so is maximum likelihood estimation

$$P_{ML}(t_i|w_j) = \frac{w_{ij}}{w_{\cdot j}},$$

where $w_{\cdot j}$ indicates the sum

$$w_{\cdot j} = \sum_{\tilde{i}} w_{\tilde{i}j}.$$

However, in most cases, there will be many low-frequency words where $w_{ij} = 0$, which will unreasonably cause $P(t_i|w_j)$ to be equal to zero. In order to fix this problem, we can use smoothing. Here we will use additive smoothing:

$$P_S(t_i|w_j) = \frac{w_{ij} + \alpha P(t_i)}{w_{\cdot j} + \alpha}. \tag{1}$$

It can be shown that additive smoothing is equivalent to using a Dirichlet distribution as a prior probability for the multinomial distribution of $P(t_i|w_j)$ [1].

The α we have introduced here is essentially a way to control the strength of the prior. If α is very large, the prior will be very strong, and it will take many observations to shift the topic distribution. If α is very small, the equation will act similarly to maximum likelihood estimation, and may cause over-fitting in the topic distribution of infrequent words. The next section describes how `dirichlet-topic.pl` automatically determines this α using Dirichlet processes.

3 Dirichlet Processes

A Dirichlet process is a method in non-parametric Bayesian statistics that allows for reasonable estimation of probabilities for relatively complicated distributions using only a single hyperparameter α. A good summary of Dirichlet processes can be found in Teh [2], but we will cover only the relevant details here (very informally).

A Dirichlet process is a type of generative distribution for discrete spaces. The most interesting characteristic of "processes" compared to other generative distributions is that that distribution changes every time it generates a new element. Let's say we have a Dirichlet process for w_1 that generates the following topic sequence $t_1t_1t_2t_1$.

First, before generating any elements, the distribution is equal to Equation (1) where $w_ij = 0$ for all w_j:

$$P_0(t_1|w_j) = \frac{0 + \alpha P(t_1)}{0 + \alpha} = P(t_1).$$

In other words, before generating any elements, the distribution is equal to the prior topic distribution, as should be expected.

The first topic t_1 is generated based on this prior, and we set $w_{1j} = 1$. We generate the second topic, also t_1, according to the new distribution with $w_{1j} = 1$:

$$P_1(t_1|w_j) = \frac{1 + \alpha P(t_1)}{1 + \alpha}.$$

We continue to generate the last two topics,

$$P_2(t_2|w_j) = \frac{0 + \alpha P(t_2)}{2 + \alpha},$$
$$P_3(t_1|w_j) = \frac{2 + \alpha P(t_1)}{3 + \alpha}.$$

If we put these together we get

$$P(t_1t_1t_2t_1|w_j) = \frac{0 + \alpha P(t_1)}{0 + \alpha} \frac{1 + \alpha P(t_1)}{1 + \alpha} \frac{0 + \alpha P(t_2)}{2 + \alpha} \frac{2 + \alpha P(t_1)}{3 + \alpha},$$

which can be further simplified into

$$P(t_1t_1t_2t_1|w_j) = \frac{\prod_{k=0}^{2}(k + \alpha P(t_1)) \prod_{k=0}^{0}(k + \alpha P(t_2))}{\prod_{k=0}^{3}(k + \alpha)}.$$

It should be noted that this is probability is *exchangable*, that is the probability will be the same regardless of the order of the sequence, as long as the topic counts are equal (it doesn't matter whether the sequence is $t_1t_1t_2t_1$ or $t_1t_1t_1t_2$). We can generalize this distribution using topic counts w_{ij} as follows:

$$P(w_{1j} = n_1, \ldots, w_{Ij} = n_I) = \frac{\prod_{i=1}^{I} \prod_{k=0}^{n_i - 1}(k + \alpha P(t_i))}{\prod_{k=0}^{n_\cdot}(k + \alpha)},$$

where

$$n_\cdot = \sum_{i=1}^{I} n_i.$$

Using this probability for a single word's topic counts, we can find the likelihood for the full corpus:

$$P(\mathcal{W}) = \prod_{j=1}^{J} \frac{\prod_{i=1}^{I} \prod_{k=0}^{w_{ij} - 1}(k + \alpha P(t_i))}{\prod_{k=0}^{w_{\cdot j}}(k + \alpha)}.$$

In the next section, we will find the α that maximizes this likelihood.

4 Parameter Estimation

We can estimate the parameters using Newton's method if we can calculate the derivative and second derivative of the likelihood for the entire corpus. In order to do this, we first take the log likelihood:

$$LL(\mathcal{W}) = \log P(\mathcal{W}) = \sum_{j=1}^{J}[\sum_{i=1}^{I}(\sum_{k=0}^{w_{ij}-1}\log(k+\alpha P(t_i)) - \sum_{k=0}^{w_{\cdot j}}\log(k+\alpha)].$$

As this is a single large sum, all we need to do is take the derivative of each component:

$$LL'(\mathcal{W}) = \sum_{j=1}^{J}[\sum_{i=1}^{I}\sum_{k=0}^{w_{ij}-1}\frac{P(t_i)}{k+\alpha P(t_i)} - \sum_{k=0}^{w_{\cdot j}}\frac{1}{k+\alpha}],$$
$$LL''(\mathcal{W}) = \sum_{j=1}^{J}[\sum_{i=1}^{I}\sum_{k=0}^{w_{ij}-1}\frac{-P(t_i)^2}{(k+\alpha P(t_i))^2} + \sum_{k=0}^{w_{\cdot j}}\frac{1}{(k+\alpha)^2}]$$

Updating α can be performed by calculating these first and second order derivatives and plugging in the values to Newton's method.

However, this equation requires a significant amount of repetitive calculation. Looking at the equation for the first order derivative, we can see that for every element $w_{ij} \geq k-1$, this equation calculates and adds $P(t_i)/(k+\alpha P(t_i))$ once, and for every element $w_{\cdot j} \geq k-1$, the equation calculates and subtracts $1/(k+\alpha)$ once. We can reduce this computation by first calculating a variable $m_i k$ that counts the number of elements of w_{ij} greater than or equal to k:

$$m_{ik} = \sum_{j=1}^{J}\mathrm{I}(w_{ij} \geq k).$$

Using this variable, we can calculate the derivatives in a much more efficient fashion:

$$LL'(\mathcal{W}) = \sum_{i=1}^{I}\sum_{k=1}^{K_i} m_{ik}\frac{P(t_i)}{k+\alpha P(t_i)} - \sum_{k=1}^{K_\cdot} m_{\cdot k}\frac{1}{k+\alpha},$$
$$LL''(\mathcal{W}) = \sum_{i=1}^{I}\sum_{k=1}^{K_i} -m_{ik}\frac{P(t_i)^2}{(k+\alpha P(t_i))^2} + \sum_{k=1}^{K_\cdot} m_{\cdot k}\frac{1}{(k+\alpha)^2},$$

where K_i is the largest value for which $m_{ik} > 0$. This is the method that is being used in `dirichlet-topic.pl`.

References

[1] David J. C. MacKay and Linda C. Peto. A hierarchical Dirichlet language model. *Natural language engineering*, 1(03):289–308, 1994.

[2] Y.W. Teh, M.I. Jordan, M.J. Beal, and D.M. Blei. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.