Simple derivations to accompany Keith et. al 2018

Katherine A. Keith

January 31, 2019

1 Marginal log likelihood (MLL) for Multinomial naive bayes (MNB)

We will derive the marginal log likelihood for the single-membership Multinomial Naive Bayes (MNB).[1]

For each document i, we define the following generative process.

1. Draw document label $y_i \sim Cat(\theta)$
2. Draw vector of word counts $x_i|y_i \sim Mult(\phi_{y_i})$.

From training, empirically have $\phi_y = p(x|y)$ and for our purposes, we want to infer θ, the proportion of documents in each class. The joint distribution of the word counts and the labels for one document i is

$$\begin{aligned} p(x_i, y_i = k) &= p(x_i|y_i = k)p(y_i = k) && (1)\\ &= Mult(x_i|\phi_{y_i=k})Cat(y_i = k|\theta) && (2)\\ &= B(x)\prod_j^V \phi_{j,k}^{x_{ij}}\theta_k. && (3) \end{aligned}$$

The log joint probability is

$$\log p(x_i, y_i = k) = \log B(x) + \sum_j^V x_{ij} \log \phi_{j,k} + \log \theta_k. \tag{4}$$

Now what we actually want is the marginal distribution on x,

$$p(x_i) = \sum_{k'=1}^K p(x_i, y_i = k'). \tag{5}$$

To find the marginal log likelihood (as a function of θ) we have

$$\begin{aligned} L(\theta) &= \prod_i^D p(x_i) && (6)\\ LL(\theta) &= \sum_i^D \log p(x_i) && (7)\\ MLL(\theta) &= \sum_i^D \log \sum_{k'=1}^K p(x_i, y_i = k'). && (8) \end{aligned}$$

[1] For further reading see Ch 2.2 in Jacob Eisenstein's "Natural Language Processing" `https://github.com/jacobeisenstein/gt-nlp-class/blob/master/notes/eisenstein-nlp-notes.pdf`

For numerical stability we use the logsumexp trick here and substitute Eq.5 and remove the constant $B(x)$

$$MLL(\theta) = \sum_{i}^{D} \log \sum_{k'=1}^{K} \exp\Big(\log p(x_i, y_i = k') \Big) \tag{9}$$

$$MLL(\theta) = \sum_{i}^{D} \log \sum_{k'=1}^{K} \exp\Big(\sum_{j}^{V} x_{ij} \log \phi_{j,k} + \log \theta_k \Big). \tag{10}$$

2 Derivative of MLL

Now we will find the derivative of the MNB log likelihood for the binary case (expanding the sum over k). Let $\theta = P(y_i = \text{positive})$. We can rewrite Eq.9 (without using the logsumexp trick) as

$$MLL(\theta) = \sum_{i}^{D} \log \Big((\prod_{j}^{V} \phi_{j,k=1}^{x_{ij}})(1-\theta) + (\prod_{j}^{V} \phi_{j,k=2}^{x_{ij}})\theta \Big) \tag{11}$$

Let $C_{1i} = (\prod_j^V \phi_{j,k=1}^{x_{ij}})$ and $C_{2i} = (\prod_j^V \phi_{j,k=2}^{x_{ij}})$. Then we have

$$MLL(\theta) = \sum_{i}^{D} \log \Big(C_{1i}(1-\theta) + C_{2i}\theta \Big) \tag{12}$$

$$= \sum_{i}^{D} \log \Big(C_{1i} - C_{1i}\theta + C_{2i}\theta \Big) \tag{13}$$

$$= \sum_{i}^{D} \log \Big(C_{1i} + \theta(C_{2i} - C_{1i}) \Big) \tag{14}$$

Taking the derivative with respect to θ we have

$$\frac{\partial}{\partial \theta} MLL(\theta) = \sum_{i}^{D} \frac{(C_{2i} - C_{1i})}{C_{1i} + \theta(C_{2i} - C_{1i})} \tag{15}$$

$$= \sum_{i}^{D} \frac{1}{\frac{C_{1i}}{(C_{2i} - C_{1i})} + \theta} \tag{16}$$

Let's call the constant $B_i = \frac{C_{1i}}{(C_{2i} - C_{1i})}$ Then

$$\frac{\partial}{\partial \theta} MLL(\theta) = \sum_{i}^{D} \frac{1}{B_i + \theta} \tag{17}$$

$$= \sum_{i}^{D} \frac{\prod_j (B_{j \neq i} + \theta)}{\prod_{i'} (B_i + \theta)} \tag{18}$$

Setting this equal to 0 we have

$$\sum_{i}^{D} \prod_{j} (B_{j \neq i} + \theta) = 0 \tag{19}$$

The **Second derivative** is thus

$$\frac{\partial^2}{\partial\theta^2} MLL(\theta) = -\sum_{i=1}^{D} \frac{1}{(B_i + \theta)^2} \tag{20}$$

3 Implicit likelihood with a generative re-interpretation

We want to combine (1) the CI coverage of the MLL model with (2) the better posteriors coming out the discriminative model like logistic regression.

Let's examine the odds ratio between the two class posteriors.

$$\text{odds} = \frac{P(y=1|x)}{P(y=0|x)}.$$

Using Bayes we obtain

$$\text{odds} = \frac{P(y=1|x)}{P(y=0|x)} = \frac{P_{\text{impl}}(x|y=1)P(y=1)}{P_{\text{impl}}(x|y=0)P(y=0)}.$$

We can estimate $P(y|x)$ from discriminative models like logistic regression, and we can empirically estimate $P(y=1) = \pi$ and $P(y=0) = (1-\pi)$ from the training true prevalence.

We set $P_{\text{impl}}(x|y=0) = 1$ and solve to obtain

$$P_{\text{impl}}(x|y=1) = \frac{P(y=1|x)}{P(y=0|x)} \frac{1-\pi}{\pi}.$$

Substituting into the MLL formula we get

$$\begin{aligned} MLL(\theta) &= \sum_i \log\Big(\theta P(x_i|y_i=1) + (1-\theta)P(x_i|y_i=0)\Big) \\ &= \sum_i \log\left(\theta \frac{P(y_i=1|x_i)}{P(y_i=0|x_i)} \frac{1-\pi}{\pi} + (1-\theta)\right). \end{aligned}$$