Summary: A series of papers about text classification by McCallum and Nigam et al

Saki Kuzushima

June 2019

1 "A Comparison of Event Models for Naive Bayes Text Classification" (1998)

1.1 Summary

- Clarify the difference of two basic models of text classification
- Both relies on Naive Bayes assumption. i.e. the ocurrence of one word is independent of the occurence of another word
- The first model represents a document as a vector of Bernoulli random variable (multi-variate Bernoulli)
- The second model represents a document as a vector of Multinomial random variable
- Multi-variate Bernoulli performs better for small vocarbulary size, but the multinomial model is generally better for a larger vocabulary size

1.2 Notation

- V: set of vocabulary, $|V|$: size of vocabulary (indexed by t or s)
- D: set of training documents, $|D|$: size of training documents (indexed by i)
- C: set of classes, $|C|$: size of classes ($|C| = 2$ for binary classification) (indexed by j)
- θ: parameter

1.3 Common Assumption

The likelihood of a document is expressed as a mixture model of classes.

$$p(d_i|\theta) = \sum_{j=1}^{|C|} p(c_j|\theta)p(d_i|c_j, \theta) \tag{1}$$

where $p(c_j|\theta)$ is a class prior, $p(d_i|c_j, \theta)$ is a class-conditional document likelihood. The class prior is estimated by

$$\hat{\theta}_{c_j} = p(c_j|\hat{\theta}) = \frac{\sum_{i=1}^{|D|} p(c_j|d_j)}{|D|} \tag{2}$$

i.e. the number of documents with class c_j divided by the number of total documents. We have different assumptions about the form of class-conditional document likelihood.

1.4 Model 1: Multi-variate Bernoulli

The key assumption is that a document is represented as a vector of Bernoulli random variables. i.e.

$$d_j = [B_{j1}, ..., B_{j|V|}] \quad \text{where } B_{jt} \overset{i.i.d}{\sim} Bernoulli(\theta_{w_t|c_j,\theta}) \tag{3}$$

i.e. a class-conditional document likelihood is

$$p(d_j|c_j,\theta) = \prod_{v=1}^{|V|}(B_{jt}p(w_t|c_j,\theta) + (1-B_{it})(1-p(w_t|c_j,\theta))) \tag{4}$$

The class-conditional word likelihood is estimated by the following.

$$\hat{\theta}_{w_t|c_j} = p(w_t|c_j,\hat{\theta}) = \frac{1+\sum_{i=1}^{|D|} B_{it}p(c_j|d_i)}{2+\sum_{i=1}^{|D|} p(c_j|d_i)} \tag{5}$$

where $p(c_j|d_i) \in \{0,1\}$ is a given class label. The numerator is the number of documents with class c_j that contains the word t, and the denominator is the number of documents with a class c_j plus 2.

1.5 Model 2: Multinomial

This model represent a document as a vector of Multinomial random variables.

$$d_j = [N_{j1}, ..., N_{j|V|}] \quad \text{where } N_{jt} \overset{i.i.d}{\sim} Multinomial(\theta_{w_t|c_j,\theta}) \tag{6}$$

Then, the class-conditional document likelihood is

$$p(d_j|c_j,\theta) = p(|d_i|)|d_i|! \prod_{v=1}^{|V|} \frac{p(w_t|c_j,\theta)^{N_{it}}}{N_{it}!} \tag{7}$$

$$\propto \prod_{v=1}^{|V|} p(w_t|c_j,\theta)^{N_{it}} \tag{8}$$

The class-conditional word likelihood is

$$\hat{\theta}_{w_t|c_j} = p(w_t|c_j,\hat{\theta}) = \frac{1+\sum_{i=1}^{|D|} N_{it}p(c_j|d_i)}{2+\sum_{s=1}^{|V|}\sum_{i=1}^{|D|} N_{is}p(c_j|d_i)} \tag{9}$$

The numerator is the number of word t in the documents with class c_j + 1 and the denominator is the number of all words in the documents with class c_j + the size of vocabulary.

1.6 Classification

We can plug in the estimator of the prior and the class-conditional document likelihood to the Bayes' rule

$$p(c_j|d_i,\hat{\theta}) = \frac{p(c_j|\hat{\theta})p(d_i|c_j,\hat{\theta})}{p(d_i|\hat{\theta})} \tag{10}$$

2 "Learning to Classify Text from Labeled and Unlabeled Documents"

2.1 summary

- Obtaining a lot of labeled data is costly while a large amount of unlabeled data is often available
- By augmenting labeled data with unlabeled data, and probabilistically assigning labels to unlabeled data, the classification accuracy was improved
- Inference is done through EM algorithm and Naive Bayes

2.2 why does it work better?

3 "Employing EM and Pool-Based Active Learning for Text Classification"

- use active