Notes & Comments for Predictive Methods for Text Mining

xxing@ucsc.edu

Xing

1. A detailed presentation and all steps for processing have some issues mention.
2. A general framework to do text mining.

Min

1. How to learn similarity measure?
   Lots of literatures work on similarity measure, e.g. different weight for different words and learn the different weight
2. Why they optimize $(w^TX-y)^2$ as the loss function instead of precision?
   Due to the central limit theorem, the distribution for large quantity of data tends to be normal distribution. And maximizing the probability of the log-likelihood of the normal distribution equals to minimize the $(w^TX-y)^2$. Therefore we use the least square error as the loss function. In the real applications, the ratings from users are distributed like Gaussian distribution.

Jadiel

1. For the hierarchical classification, how do they do pruning?
   The pruning process depends on what kind of algorithms used, and its objective is not to model the whole space. In practice, the pruning may be based on how many data points fall into the category. Some work doesn't do pruning and just use the known categories, like Yahoo categories. Path pruning is not to check all the paths due to the efficiency and prune some paths, like A* search in artificial intelligence field.

Yize

1. When to use local modeling approach and when to use global modeling approach? Are they task dependent?
   Local modeling approach is easy and common, often used for segmentation task, like logistic model. Global modeling is used for sequence labeling, using CRF method. Random walk is over the graph like HMM and more like a local modeling method.

Sarah

1. How to know whether the model works or the feature works? How about good model with poor features or vice versa?
   They work together for the final performance. To compare models, we need to fix the features and compare different methods, like some published papers. As the example of the email spam detection in the presentation, features are same.

Jian

1. How to model user behavior as the feature?
   It depends on the algorithm. Some algorithms are sensitive and some are not. For sensitive ones, like NB methods, it's better to use feature selection and remove the irrelevant or noisy features. For non-sensitive ones, it is ok to use all the features. Feature selection can be performed as a machine learning problem, e.g. clustering results can work as a feature.

Jessica

1. How to incorporate non-text features with text features?
   One way is to assume the features are independent and train different models over different features. And then we can use voting/weighted voting to get the result. This is a simple method. Another way is to handle both features. It works well if the features are complementing each other. We can change the function formula or just play with the output using weighted learning.

Yi

1. Consistency in tokenization is very important.
2. Truncated term frequency has different methods to implement, like modified term frequency, map term frequency to log or RM 25. For language model or NB model, they also use truncated term frequency.
3. From Zipf's law, it can be used as the prior probability in the Bayesian method.
4. Computation complexity is an important issue in text categorization when there are hundreds of classes. Some methods in personalization, like having millions of users, can be used in this scenario.

Jiazhong

1. How does sparse regularization work at page 50?
   In the formula, Norm 1 has the feature selection effect. w in the formula is sparse and the 0 value works equal to the feature is deleted. There is one book, statistical learning and data mining, talking about this question in detail.