A Naive Bayes implementation for classifying news headlines text

AAAI Press
Association for the Advancement of Artificial Intelligence
2275 East Bayshore Road, Suite 160
Palo Alto, 94303

Abstract

A text classification algorithm is used to classify the incoming headlines into 4 broad categories:Business, Health, Entertainment, and Technology. The Naive Bayes is a popular technique for classifying text due to its relative simplicity and ease of implementation. In this project, we design a Naive Bayes classifier for classifying headlines based on only the text contained in them.

Introduction

Text classification is an intersting field of machine learning and its applications are diverse. Many algorithms exist for classifying text. However, a Naive Bayes classifier is a simple and effective method for classifying text. For text classification, a bag of words model is usually used. But, the number of potential words is much larger than the number of training documents available[1]**(An Extensive Empirical Study of Feature Selection Metrics for Text Classification)**. Once the features are selected, the classifier is trained and parameters of the model are estimated. These parameters are **P(X(i)/Y(j)) (forall) X and (forall) Y(j)**. Based on the distribution we assume, these parameters are estimated using maximum likelihood estimate. Once the classifie has been trained, the performance of the classifier has to be evalutaed. **(write about the peroformance metrics, confusion matrix from paper and cite)**.

Naive Bayes

Naive Bayes is a simple probabilistic classifier which exploits the Bayes theorem. A very strong assumption that the Naive Bayes makes is independece of features given a class label.

$$P(X1, X2, X3....XN/|Y) = P(X1|N).P(X1|N)...P(XN|N)$$

Feature selection

The feature selection procedure, in general, scores each feature according to a metric, and selects the best k features[1]**(An Extensive Empirical Study of Feature Selection Metrics for Text Classification)**. A pre-processing step consisting of stop-word removal and stemming were also carried out to reduce the feature size. Document Frequency thresholding (DF) has proven to be one of the best feature selection method text classification. Document frequency signifies the occurrence of particular term in whole collection of documents and by setting threshold, terms occurred multiple times are retained. Terms which are non informative for classification can be removed. By using this, tens of hundreds rare features can be removed before the step of feature selection[4]. As document frequency can be calculated in entire test set, selecting features based on repeated word increases probability of these feature can be present in future test cases. Text domain has number of features and most of these features are not useful for text classification. In such case document frequency plays important role and speed up the categorisation process. Due to its simplicity and effectiveness, DF is adopted in more and more text mining experiments[1][2][3]. **Write about document frequency and the two types of filtering methods from the papaer**

Classifier design

Once the bag-of-words features are selected, the conditional probability of each words given the class lablel has to be computed. The random variable here is asssumed to bernoulli distributed as the number of words in the headline is not of much importance. Just the presence of certain keywords will be used in the classification. As a resut, the parameter estimation using maximum likelihood estimate is just **(insert formula here)**. To handle precision, log probability.

Performance metric

Any classifier developed has to have a general metric to measure its performance and evaluate the model against them for it to be used for real world data. A generic performance measure called Multiclass Performance Measure(MPS) is developed for pattern recognition (**?**). For this multi-class classifier where the input is classified into one class out of given number of classes, different performance measures such as accuracy, precision and recall are used. Average accuracy, precision and recall are calculated as given in the Table [2]

which uses the concept of confusion matrix for binary classification as a micro- or macro- average of Table [1].

Table 1: Confusion Matrix for binary classification

Data class	Classified as positive	Classified as negative
Positive	True positive (tp)	True negative(tn)
Negative	False positive (fp)	False negative(fn)

Table 2: Measures for multi-class classification based on a generalization of the measures of Table 1 for many classes Ci:tpi are true positive for Ci , and fpi false positive, fni false negative, and tni true negative counts respectively. l and M indices represent micro- and macro-averaging.

Measure	Formula
Average Accuracy	$\frac{\sum_{i=1}^{l} \frac{(tp_i+tn_i)}{(tp_i+fn_i+fp_i+tn_i)}}{l}$
Precision	$\frac{\sum_{i=1}^{l} tp_i}{\sum_{i=1}^{l}(tp_i+fp_i)}$
Recall	$\frac{\sum_{i=1}^{l} tp_i}{\sum_{i=1}^{l}(tp_i+fn_i)}$

Bibliography

[1] Xu Y., Wang B., Li J., Jing H. (2008) An Extended Document Frequency Metric for Feature Selection in Text Categorization. In: Li H., Liu T., Ma WY., Sakai T., Wong KF., Zhou G. (eds) Information Retrieval Technology. AIRS 2008. Lecture Notes in Computer Science, vol 4993. Springer, Berlin, Heidelberg

[2] Dumais, S.T., Platt, J., Heckerman, D., Sahami, M.: Inductive learning algorithms and representations for text categorization. In: Proceedings of CIKM 1998, pp. 148 155 (1998)

[3] Li, Y.H., Jain, A.K.: Classification of text documents. Comput. J. 41(8), 537 546 (1998)

[4] B. Tang, S. Kay, H. He, "Toward Optimal Feature Selection in Naive Bayes for Text Categorization", IEEE Transactions on Knowledge and Data Engineering, vol. 28, no. 9, pp. 2508-2521, 2016

[5] Thomas Kautz, Bjoern M. Eskofire, Cristian F. Pasluosta, Generic performance measure for multiclass classifier, In Pattern Recognition, Volume 68, 2017, Pages 111-125

[6] Marina Sokovola, Guy Lapalme, A Systematic analysis of performance measures for classification tasks, In Information Processing and Management, volume 45, Issue 4, 2009, pages 427-437

Thank you for reading these instructions carefully. We look forward to receiving your electronic files!