Tutorial #3 for Assignment 1: Classification

Feb 3rd, 2017

TA: Aryan Arbabi
arbabi@cs.toronto.edu
Methodology

Tweet Corpus → Preprocessing → Normalized Tweets → Feature Extraction → Feature Vectors → Classification → Classes
Methodology

Tweet Corpus

Tweet Corpus → Preprocessing → Normalized Tweets → Feature Extraction → Feature Vectors → Classification → Classes

Today’s class
Create arff files

buildarff.py test.twt some.arff 50

input filename

output filename

Number of tweets per class
Using Weka

- Available in /u/cs401/WEKA

- Command:
  ```java -cp weka.jar *weka classifier* *classifier options*```
Test 3 different classifiers

- SVM
  - weka.classifiers.functions.SMO

- Naive Bayes
  - weka.classifiers.bayes.NaiveBayes

- Decision Tree
  - weka.classifiers.trees.J48
Training size

- Experiment with different training sizes
- 500 to 10k tweets for each class
- Increments of 500
Information Gain

- $\text{InfoGain(Class, Attribute)} = \text{H(Class)} - \text{H(Class | Attribute)}$
- `sh /u/cs401/WEKA/infogain.sh *arff file* > *output file*`
Cross-validation
Cross-validation

• 10-fold cross-validation on all 20,000 training tweets

• A vector of 10 accuracies for each classifier

• Find p-value of difference in classifier results
  • null hypothesis: identical means
Precision & Recall

Precision = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}

Recall = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}

Image from Wikipedia
Questions?