Native Language Identification Using a Linear SVM Classifier

...

Brian Ward

The Task

- Given a set of training data from the TOEFL11 corpus, develop a model to accurately classify the native languages (L1) of authors from a set of unseen test data
- Inspired by the Association for Computational Linguistics' (ACL) shared task

Methods

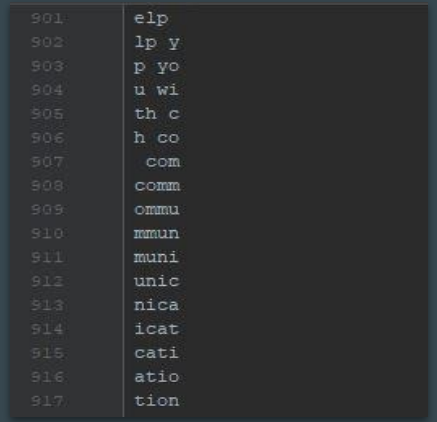
- ACL 2017 submissions
  - Many machine learning approaches found success
  - SVMs very common
  - Differed in choosing of features

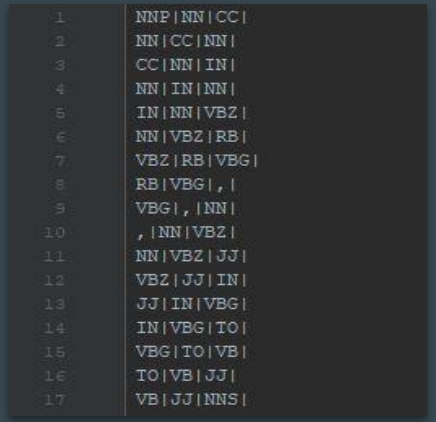
My system

- Linear SVM classifier
  - *n*-gram features
    - Parts-of-speech
    - Characters
  - Penalty value tuning using TOEFL-DEV set

Feature extraction

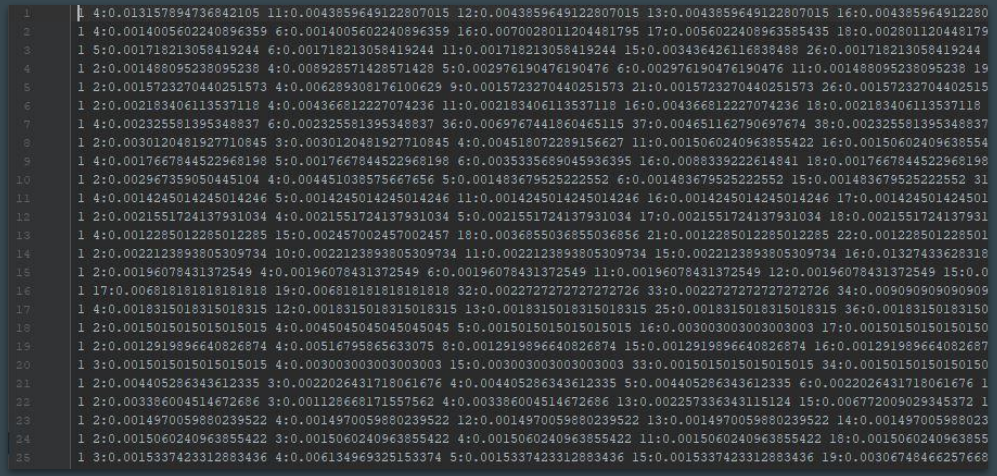
- Determine which features could be relevant to training
  - Get *n*-gram counts across all documents in test set
  - Normalize counts between 0-1
  - Drop *n*-gram features which occur in only 1 document
- Save relevant features





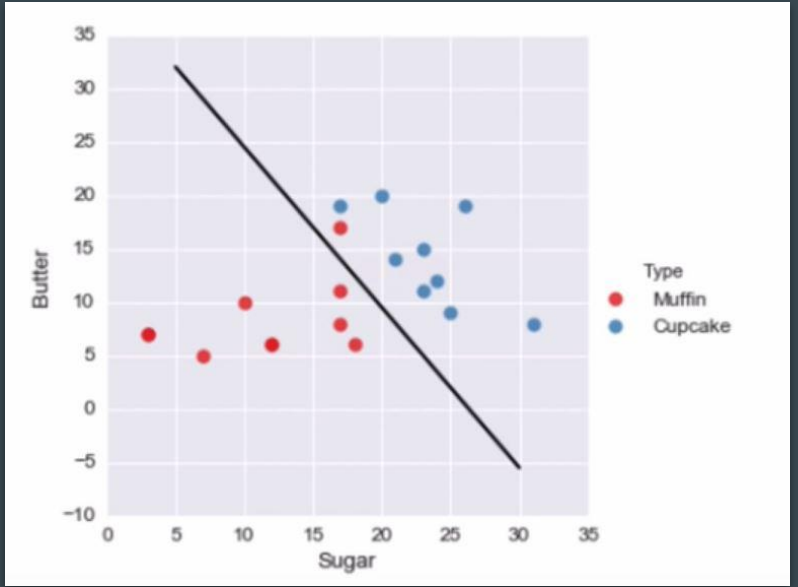
Feature extraction (cont.)

- Generate feature vectors for every document
  - Depending on value of *n* and features used, vector length spanned from around 20,000 features to over 150,000
- Output to LIBLINEAR SVM format



Support Vector Machines (SVM)

- Given training data and a set of classes, can we create a division in *n*-dimensional space which can separate vectors into their own classes
  - Easy to visualize in up to 3 dimensions
  - Hard to visualize when $n > 3$
- Multiple classes
  - One-vs-rest method

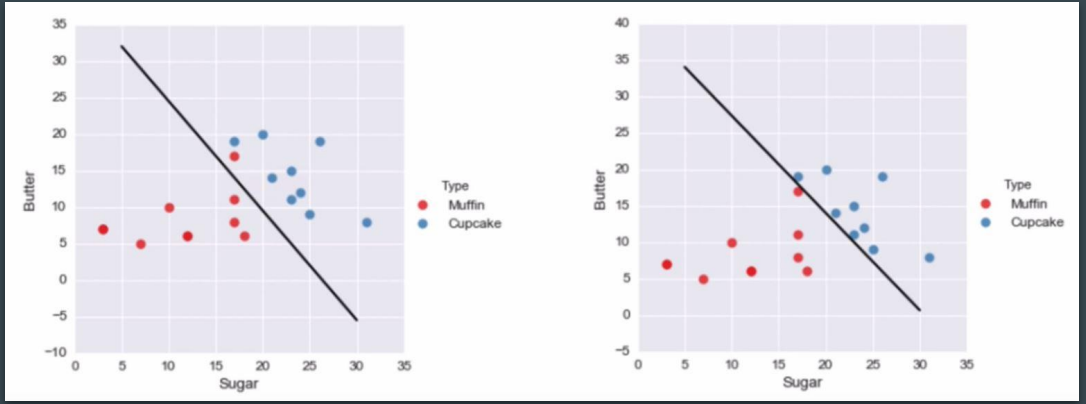


Training the SVM

- Tuning C-parameter
  - SVM trained with TOEFL11-TRAIN
  - Tested accuracy of different C values
  - Use C value which yields highest accuracy on DEV set
- What is C-parameter?
  - Penalty incurred for misclassified vectors

c-val	Overall Accuracy on DEV set
2	48.2%
4	50.5%
8	54.5%
16	58.8%
32	62.3%
64	65.4%
128	67.3%
256	68.8%
512	68.8%
1024	68.5%

C-Val Accuracy using 3-gram POS and Character *n*-grams



Results

- POS only feature model achieves best results with $n = 3$, C = 128
  - Overall accuracy 53.5%
  - Best at classifying native German speakers
    - Precision: 73.8%
    - Recall: 90%
    - F1 score: 81.1%
- Combined character and POS achieves best results with $n = 4$ and $n = 3$ respectively, C = 512
  - Overall accuracy 69.2%

Results (cont.)

- Best results with character 4-grams, C = 1024
  - Overall accuracy 74.1%

2-gram POS features only with c-val of 128											
	ARA	**DEU**	**FRA**	**HIN**	**ITA**	**JPN**	**KOR**	**SPA**	**TEL**	**TUR**	**ZHO**
Precision	46.6%	58.4%	50.5%	44.9%	50.9%	51.4%	50.0%	48.1%	57.9%	43.0%	58.3%
Recall	41.0%	66.0%	48.0%	40.0%	59.0%	56.0%	47.0%	39.0%	66.0%	40.0%	63.0%
F1	43.6%	62.0%	49.2%	42.3%	54.6%	53.6%	48.5%	43.1%	61.7%	41.5%	60.6%
Overall Accuracy = 51.4%											
3-gram POS features only with c-val of 128											
	ARA	**DEU**	**FRA**	**HIN**	**ITA**	**JPN**	**KOR**	**SPA**	**TEL**	**TUR**	**ZHO**
Precision	69.4%	73.8%	67.7%	57.3%	71.2%	55.3%	55.3%	43.4%	50.9%	51.6%	61.4%
Recall	59.0%	90.0%	63.0%	59.0%	74.0%	57.0%	52.0%	36.0%	59.0%	47.0%	70.0%
F1	63.8%	81.1%	65.3%	58.1%	72.5%	56.2%	53.6%	39.3%	54.6%	49.2%	65.4%
Overall Accuracy = 53.5%											
3-gram character features only with c-val of 512											
	ARA	**DEU**	**FRA**	**HIN**	**ITA**	**JPN**	**KOR**	**SPA**	**TEL**	**TUR**	**ZHO**
Precision	61.6%	63.1%	68.7%	55.3%	76.4%	75.5%	73.9%	65.6%	69.9%	69.2%	73.0%
Recall	61.0%	77.0%	68.0%	63.0%	75.5%	74.0%	65.0%	61.0%	72.0%	63.0%	73.0%
F1	61.3%	69.4%	68.3%	58.9%	78.6%	74.7%	69.1%	63.2%	70.9%	66.0%	73.0%
Overall Accuracy = 70.2%											
4-gram character features only with c-val of 1024											
	ARA	**DEU**	**FRA**	**HIN**	**ITA**	**JPN**	**KOR**	**SPA**	**TEL**	**TUR**	**ZHO**
Precision	76.6%	80.6%	78.9%	59.0%	76.2%	79.8%	72.7%	67.0%	70.8%	81.8%	73.6%
Recall	72.0%	87.0%	75.0%	62.0%	80.0%	75.0%	72.0%	67.0%	75.0%	72.0%	78.0%
F1	74.2%	83.7%	76.9%	60.5%	78.0%	77.3%	72.4%	67.0%	72.8%	76.6%	75.7%
Overall Accuracy = 74.1%											
3-gram combined POS and character with c-val of 1024											
	ARA	**DEU**	**FRA**	**HIN**	**ITA**	**JPN**	**KOR**	**SPA**	**TEL**	**TUR**	**ZHO**
Precision	69.0%	72.8%	69.1%	51.9%	73.2%	73.2%	63.6%	59.6%	65.4%	67.7%	76.8%
Recall	60.0%	91.0%	65.0%	56.0%	71.0%	71.0%	63.0%	56.0%	68.0%	65.0%	76.0%
F1	64.2%	80.9%	67.0%	53.8%	72.1%	72.1%	63.3%	57.7%	66.7%	66.3%	76.4%
Overall Accuracy = 67.5%											
3-gram POS and 4-gram character with c-val of 1024											
	ARA	**DEU**	**FRA**	**HIN**	**ITA**	**JPN**	**KOR**	**SPA**	**TEL**	**TUR**	**ZHO**
Precision	73.2%	73.4%	68.0%	55.2%	75.8%	74.0%	65.0%	55.6%	66.0%	76.1%	79.8%
Recall	60.0%	91.0%	66.0%	58.0%	75.0%	74.0%	65.0%	55.0%	68.0%	70.0%	79.0%
F1	65.9%	81.2%	67.0%	56.6%	75.4%	74.0%	65.0%	55.3%	67.0%	72.9%	79.4%
Overall Accuracy = 74.1%											

SVM Predictions on Test Data

Questions?