Evaluation Metrics

Bonan Min

bonanmin@gmail.com

Some slides are based on class materials from Thien Huu Nguyen, Dan Jurafsky, James H. Martin

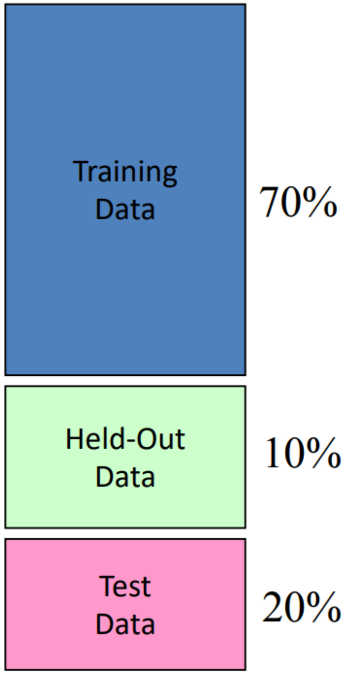
Evaluation

Data: labeled examples, e.g. emails marked spam/not-spam

- Training set
- Held out /Development (dev) set
- Test set
- Can also do cross-validation over multiple splits
  - Pool results over each split
  - Compute average dev/test set result

Features: attribute-value pairs which characterize each *X*

These sets are disjoint!



Evaluation

Accuracy: fraction of instances predicted correctly

Accuracy can be Misleading

- For tasks where one tag predominates, accuracy can overstate performance

> - Task: classify emails as spam or not-spam
> - Accuracy: the fraction of emails in the test set that are correctly predicted
> - It’s easy to build a high-accuracy “majority class” classifier when non-spam emails dominate the dataset
> - But we don’t really care about the ham emails. We want
>   - An evaluation measures that focus directly on the spam emails.

So, we use the confusion matrix:

- Accuracy = (TN + TP) / total = (50+100)/165 = .91
- Precision (P) = % predicted examples that are correct
  = TP / (TP + FP) = 100 / (100 + 10) = .91
- Recall (R) = % of correct examples that are selected
  = TP / (TP + FN) = 100 / (100 + 5) = .95
- F1 = 2PR/(P+R) – geometric mean of P and R

n=165	Predicted: NO	Predicted: YES	
Actual: NO	TN = 50	FP = 10	60
Actual: YES	FN = 5	TP = 100	105
	55	110	

Evaluation with More Than Two Classes

Confusion matrix: for each pair of classes <c_1, c_2>, how many documents from c_1 were incorrectly assigned to c_2?

Docs in test set	Assigned UK	Assigned poultry	Assigned wheat	Assigned coffee	Assigned interest	Assigned trade
True UK	95	1	13	0	1	0
True poultry	0	1	0	0	0	0
True wheat	10	90	0	1	0	0
True coffee	0	0	0	34	3	7
True interest	-	1	2	13	26	5
True trade	0	0	2	14	5	10

- Macroaveraging: compute performance for each class, then average (classes are equal)
- Microaveraging: collect decisions for all classes, compute confusion table, evaluate (more preferable if classes are imbalanced)

Recall:

Fraction of docs in class *i* classified correctly: $\frac{c_{ii}}{\sum_j c_{ij}}$

Precision:

Fraction of docs assigned class *i* that are actually about class *i*: $\frac{c_{ii}}{\sum_j c_{ji}}$

Accuracy: (1 - error rate)

Fraction of docs classified correctly: $\frac{\sum_i c_{ii}}{\sum_j \sum_i c_{ij}}$

Micro- vs. Macro-Averaging: An Example

Class 1

	Truth: yes	Truth: no
Classifier: yes	10	10
Classifier: no	10	970

Class 2

	Truth: yes	Truth: no
Classifier: yes	90	10
Classifier: no	10	890

Micro Ave. Table

	Truth: yes	Truth: no
Classifier: yes	100	20
Classifier: no	20	1860

- Macroaveraged precision: $(0.5 + 0.9)/2 = 0.7$
- Microaveraged precision: $100/120 = .83$
- Microaveraged score is dominated by score on common classes

Some Datasets for Text Classification

Reuters-21578 (http://disi.unitn.it/moschitti/corpora.htm)

20Newsgroups (http://disi.unitn.it/moschitti/corpora.htm)

Yelp reviews 2013, 2014, 2015 (http://ir.hit.edu.cn/~dytang/paper/emnlp2015/emnlp-2015-data.7z)

.....