Statistical Learning for Text Data Analytics
Text Categorization 2: Clustering

Yangqiu Song

Hong Kong University of Science and Technology

yqsong@cse.ust.hk

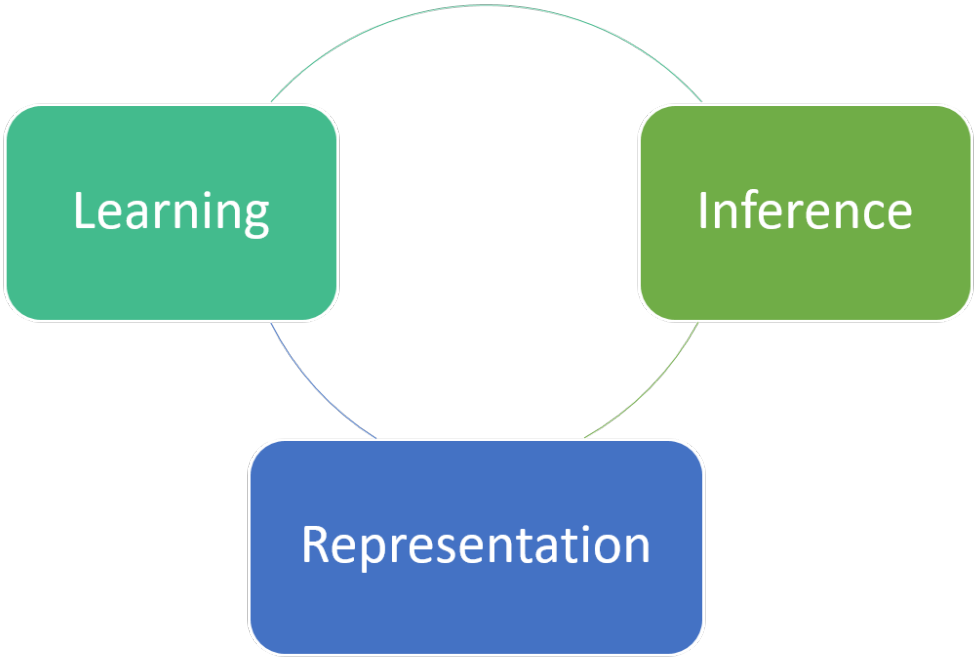
Spring 2018

∗Contents are based on materials created by Noah Smith, Xiaojin (Jerry) Zhu, Eric Xing, Vivek Srikumar, Dan Roth

Reference Content

- Noah Smith. CSE 517: Natural Language Processing https://courses.cs.washington.edu/courses/cse517/16wi/
- Xiaojin (Jerry) Zhu. CS 769: Advanced Natural Language Processing. http://pages.cs.wisc.edu/~jerryzhu/cs769.html
- Eric Xing. 10715 Advanced Introduction to Machine Learning. https://www.cs.cmu.edu/~epxing/Class/10715/lectures/lecture1.pdf
- Vivek Srikumar. CS 6355 Structured Prediction. https://svivek.com/teaching/structured-prediction/spring2018/
- Dan Roth. CS546: Machine Learning and Natural Language . http://l2r.cs.uiuc.edu/~danr/Teaching/CS546-16/

Course Topics



- Representation: language models, word embeddings, topic models
- Learning: supervised learning, unsupervised learning, semi-supervised learning, sequence models, deep learning, optimization techniques
- Inference: constraint modeling, joint inference, search algorithms

NLP applications: tasks introduced in Lecture 1

Overview

Evaluation

- Accuracy:

$$\begin{aligned} A(f) &= P(f(\mathbf{X}) = Y) \\ &= \sum_{\mathbf{x}\in\mathcal{V}, y\in\mathcal{Y}} P(\mathbf{X}=\mathbf{x}, Y=y) \cdot \begin{cases} 1 & \text{if } f(\mathbf{x}) = y \\ 0 & \text{otherwise} \end{cases} \\ &= \sum_{\mathbf{x}\in\mathcal{V}, y\in\mathcal{Y}} P(\mathbf{X}=\mathbf{x}, Y=y) I(f(\mathbf{x}) = y) \end{aligned}$$

  where P is the true distribution over data

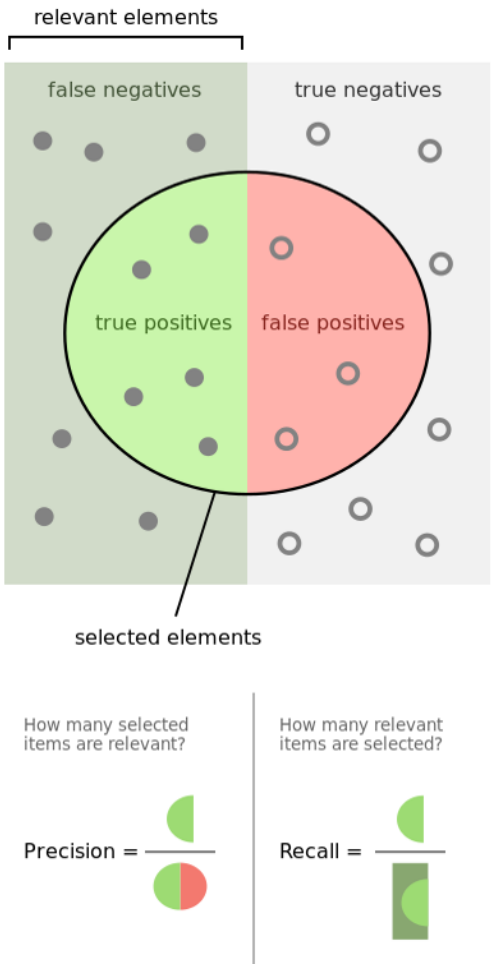
- Error is $1 - A(f)$
- This is estimated using a test dataset $\langle \bar{\mathbf{x}}_1, \bar{y}_1 \rangle, \ldots, \langle \bar{\mathbf{x}}_m, \bar{y}_m \rangle$:

$$\hat{A}(f) = \frac{1}{m} \sum_{i=1}^{m} I(f(\bar{\mathbf{x}}_i) = \bar{y}_i)$$

Issues with Test-Set Accuracy

- Class imbalance: if $P(L = \text{not spam}) = 0.99$, then you can get $\hat{A} \approx 0.99$ by always guessing "not spam"
- Relative importance of classes or cost of error types
- Variance due to the test data

Evaluation in the Two-Class Case



- Precision
  - Fraction of predicted positive documents that are indeed positive, i.e., P(human label $= 1 \mid$ prediction $= 1$)
- Recall
  - Fraction of positive documents that are predicted to be positive, i.e., P(prediction $= 1 \mid$ human label $= 1$)
- F-1 Score:

$$F_1 = 2 \cdot \frac{precesion \cdot recall}{precesion + recall}$$

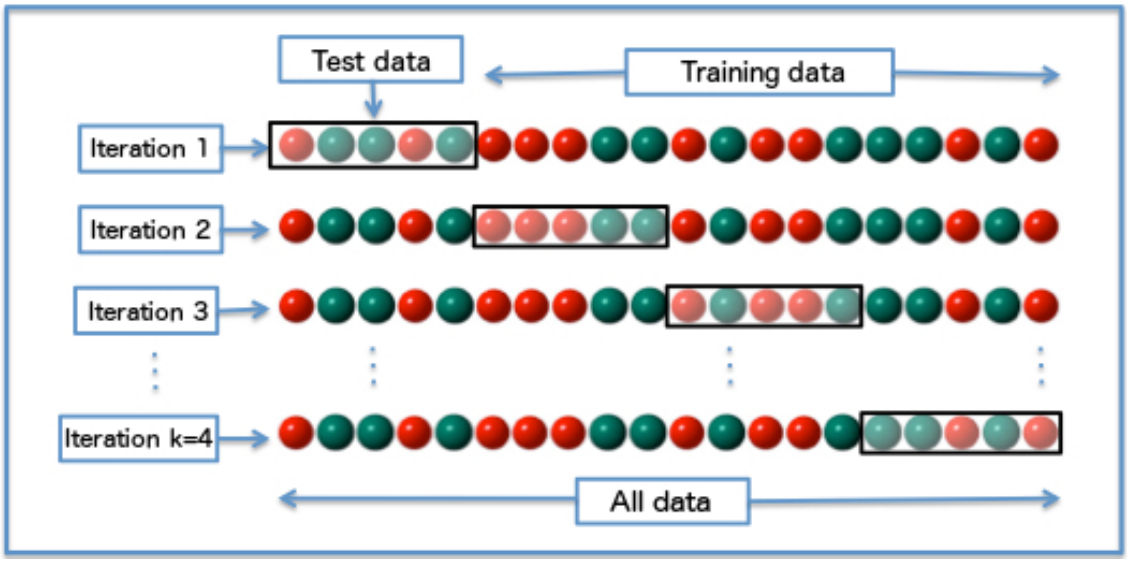
Evaluation in the Multi-Class Case

- Accuracy
- F1
  - Let TP_t, FP_t, FN_t denote the true-positives, false-positives, and false-negatives for the t-th label in label set $\mathcal{L}$ respectively
  - Micro-averaged $F_1 = \frac{2PR}{P+R}$ where $P = \frac{\sum_{t\in\mathcal{L}} TP_t}{\sum_{t\in\mathcal{L}} TP_t+FP_t}$ and $R = \frac{\sum_{t\in\mathcal{L}} TP_t}{\sum_{t\in\mathcal{L}} TP_t+FN_t}$
  - Macro-averaged $F_1 = \frac{1}{|\mathcal{L}|}\sum_{t\in\mathcal{L}} \frac{2P_tR_t}{P_t+R_t}$ where $P_t = \frac{TP_t}{TP_t+FP_t}$ and $R_t = \frac{TP_t}{TP_t+FN_t}$

Actual/ Predicted	Class 1	2	3	4	5	6	7	8	9	10	Total	Recall
Class 1	**9.06**		0.07	0.05	0.01	0.03	0.06	0.59	0.01	0.14	10	90.60
Class 2		**8.20**			0.52	0.04	0.30		0.53	0.42	10	82.00
Class 3	0.03		**9.52**	0.03	0.01	0.02	0.01	0.15	0.02	0.22	10	95.20
Class 4	0.01	0.01	0.01	**9.01**	0.13	0.12	0.52	0.10	0.05	0.06	10	90.10
Class 5		0.48	0.01	0.05	**2.67**	1.87	1.40		2.63	0.90	10	26.70
Class 6		0.11			0.86	**7.75**	0.56		0.10	0.62	10	77.50
Class 7	0.02	0.18		0.32	1.47	1.50	**3.66**	0.11	2.08	0.67	10	36.60
Class 8	0.20		0.05	0.01			0.02	**9.70**		0.03	10	97.00
Class 9		0.39	0.01		1.21	0.11	0.42		**6.84**	1.02	10	68.40
Class 10		0.24	0.13	0.01	0.95	1.01	0.43	0.01	1.85	**5.37**	10	53.70
Total	9.32	9.61	9.80	9.48	7.83	12.45	7.38	10.66	14.11	9.45	100	
Precision	97.21	85.33	97.14	95.04	34.10	62.25	49.59	90.99	48.48	56.83		

Model Estimation and Selection

- k-fold cross-validation
  - Partition all training data into k equal size disjoint subsets
  - Leave one subset for validation and the other k-1 for training
  - Repeat step (2) k times with each of the$k$subsets used exactly once as the validation data



Statistical Significance

- Suppose we have two classifiers f_1 and f_2
- Is f_1 better? The "null hypothesis," denoted H_0, is that it isn't. But if $\hat{A}(f_1) \gg \hat{A}(f_2)$, we are tempted to believe otherwise
- How much larger must $\hat{A}(f_1)$ be than $\hat{A}(f_2)$ to reject H_0?
- Frequentist view: how (im)probable is the observed difference, given $H_0 = true$?
- Caution: statistical significance is neither necessary nor sufficient for research significance or practical usefulness!

A Hypothesis Test for Text Classifiers

McNemar (1947)

- The null hypothesis: $A(f_1) = A(f_2)$
- Pick significance level α, an "acceptably" high probability of incorrectly rejecting H_0
- Calculate the test statistic, k (explained in the next slide)
- Calculate the probability of a more extreme value of k, assuming H_0 is true; this is the p-value
- Reject the null hypothesis if the p-value is less than α

The p-value is $P(\text{this observation} \,|H_0 \text{ is true})$, not the other way around

McNemar's Test: Details

- Assumptions: independent (test) samples and binary measurements. Count test set error patterns:
- The test is applied to a 2×2 contingency table, which tabulates the outcomes of two tests on a sample of n subjects

	f_1 is incorrect	f_1 is correct	
f_2 is incorrect	a	b	$a+b$
f_2 is correct	c	d	$n \cdot \hat{A}(f_1) = c+d$
	$a+c$	$n \cdot \hat{A}(f_1) = b+d$	n

If $A(f_1) = A(f_2)$, then b and c are each distributed according to Binomial$(k, b+c, \frac{1}{2})$ (The probability of getting k successes in $b+c$ trials)

$$\text{test statistic } k = \min(b, c)$$

$$p - \text{value} = 2\sum_{0}^{k} \text{Binomial}(k; b+c, \frac{1}{2}) = \frac{1}{2^{b+c-1}} \sum_{j=0}^{k} \binom{b+c}{j}$$

Other Tests

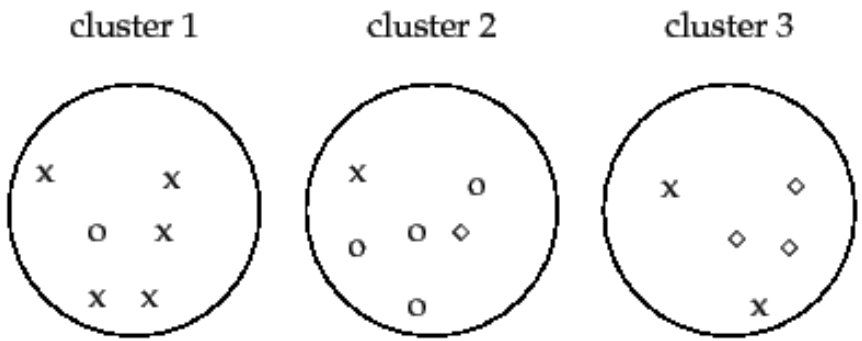
- Different tests make different assumptions
- Sometimes we calculate an interval that would be "unsurprising" under H_0 and test whether a test statistic falls in that interval (e.g., t-test and Wald test)
- In many cases, there is no closed form for estimating p-values, so we use random approximations (e.g., permutation test and paired bootstrap test)
- If you do lots of tests, you need to correct for that
- Read lots more in (Smith (2011)), appendix B

Metrics for Clustering

- Purity between two random variables CAT (category label) and CLS (cluster label) is defined as:

$$\text{Purity (CAT; CLS)} = \frac{1}{n}\sum_{j} \max n_{ij},$$

  - n is the number of documents
  - $n_{i,j}$ is the number of documents in category i as well as in cluster j



► **Figure 16.1** Purity as an external evaluation criterion for cluster quality. Majority class and number of members of the majority class for the three clusters are: x, 5 (cluster 1); o, 4 (cluster 2); and ⋄, 3 (cluster 3). Purity is $(1/17) \times (5+4+3) \approx 0.71$.

Sometimes Hungarian algorithm is used to match category and cluster $\frac{1}{n}\max\sum_i n_{i,f(i\to j)}$

Metrics for Clustering

- In probability theory and information theory, the mutual information (MI) of two random variables is a measure of the mutual dependence between the two variables.
- More specifically, it quantifies the "amount of information" (in units such as Shannons, more commonly called bits) obtained about one random variable, through the other random variable.
- NMI between two random variables CAT (category label) and CLS (cluster label) is defined as:

$$\text{NMI}(\text{CAT}; \text{CLS}) = \frac{\text{I}(\text{CAT}; \text{CLS})}{\sqrt{\text{H}(\text{CAT})\text{H}(\text{CLS})}},$$

where I(CAT; CLS) is the mutual information between CAT and CLS. The entropies H(CAT) and H(CLS) are used for normalizing the mutual information to be in the range of $[0, 1]$.

Metrics for Clustering

- In practice, we made use of the following formulation to estimate the NMI score (Strehl and Ghosh (2002)):

$$\text{NMI} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{k} n_{i,j} \log\left(\frac{n \cdot n_{i,j}}{n_i \cdot n_j}\right)}{\sqrt{\left(\sum_i n_i \log\frac{n_i}{n}\right)\left(\sum_j n_j \log\frac{n_j}{n}\right)}},$$

  - n is the number of documents
  - n_i and n_j denote the number of documents in category i and cluster j
  - $n_{i,j}$ is the number of documents in category i as well as in cluster j
- The NMI score is 1 if the clustering results perfectly match the category labels, and the score is 0 if data are randomly partitioned.
- The higher the NMI score, the better the clustering quality.

References I

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

Smith, N. A. (2011). *Linguistic Structure Prediction*. Synthesis Lectures on Human Language Technologies. Morgan and Claypool.

Strehl, A. and Ghosh, J. (2002). Cluster ensembles — A knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, 3:583–617.