15: Catch up: Clustering Evaluation

Andreas Vlachos

How do we evaluate clustering?

Last lecture: **purity**

Given a dataset annotated with gold standard class labels:

- Assign the majority label of the datapoints in each cluster as the “label”
- Measure accuracy with respect to these “labels”

Problem?

- Clusterings with more clusters more likely to achieve higher purity
- Trivially, assigning each datapoint to its own cluster achieves perfect purity

Today

We will look at some other evaluation options

- Rand Index
- Adjustment for chance
- F-score
- Evaluation of soft clustering

Let’s look at an example situation

	Class1	Class2	Class3
cluster1	2	2	0
cluster2	2	4	1
cluster3	0	1	1
cluster4	0	0	4

How would you evaluate this clustering?

Purity?

A different view

Let’s measure clustering quality as accuracy over pairwise decisions: should a pair of datapoints be in the same cluster or not?

The Rand Index:

$$\text{RandIndex} = \frac{\text{correctly classified pairs as either in the same cluster or not}}{\text{all pairs}}$$

A form of pairwise accuracy

Let’s make this more concrete

Rand Index

Assume:

- datapoints $X = \{x_1, x_2, ...x_d\}$
- a clustering $\mathcal{K} = \{K_1, K_2, ...K_k\}$
- gold standard classes $\mathcal{C} = \{C_1, C_2, ...C_c\}$

Define:

- Pairs of points clustered together correctly

$$\mathcal{T} = \{(x_i, x_j) | x_i, x_j \in K_m, x_i, x_j \in C_n\}$$

- Pairs of points clustered separately correctly

$$\mathcal{S} = \{(x_i, x_j) | x_i \in K_{m1}, x_j \in K_{m2}, x_i \in C_{n1}, x_j \in C_{n2}\}$$

Finally: $$RandIndex(X, \mathcal{K}, \mathcal{C}) = \frac{|\mathcal{T}| + |\mathcal{S}|}{\binom{d}{2}}$$

Rand Index properties

Range of possible values?

In theory: 0 to 1

In practice: closer to 1 (most pairs of datapoints are clustered separately correctly)

Two solutions:

- Adjustment for chance
- F-score

Adjusted Rand Index

Need to take into account that a pair of datapoints might have been clustered together or separately by chance:

$$\mathrm{AdjustedRI} = \frac{RI - \mathrm{expected\_RI}}{1 - \mathrm{expected\_RI}}$$

Same concept as the Kappa score from lecture 6!

The eventual formula is a bit more complicated due to having to calculate the expectation, but it is available in many statistical/ML packages

Imbalance in clustering evaluation

Typically most datapoints should be clustered separately.

Rand Index is accuracy for pairs of datapoints, but the task is imbalanced.

We have true positives (clustered together correctly) and true negatives (clustered separately), we need the false positives: **clustered together incorrectly**:

$$\mathcal{FP} = \{(x_i, x_j) | x_i, x_j \in K_m, x_i \in C_{n1}, x_j \in C_{n2}\}$$

And the false negatives which are **clustered separately incorrectly**:

$$\mathcal{FN} = \{(x_i, x_j) | x_i \in K_{m1}, x_j \in K_{m2}, x_i, x_j \in C_n\}$$

And the definitions of recall, precision and F-score follow those from lecture 9.

Soft clustering

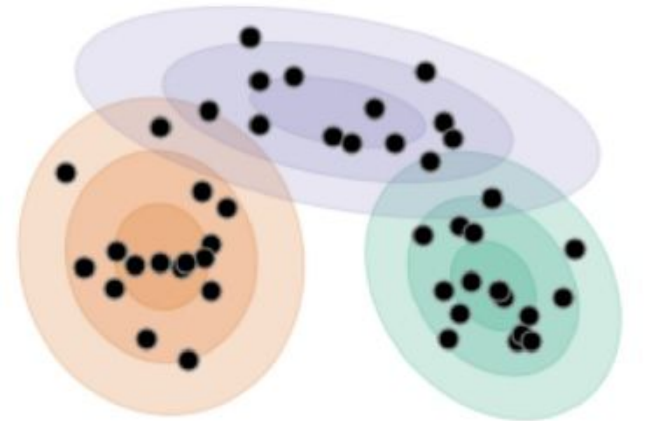
So far hard clustering: each datapoint is assigned to a single cluster.

What about soft clustering, where each datapoint can be assigned to multiple clusters with graded membership?

Many real-world applications:

- Document tagging
- Image labelling
- Social network analysis
- Word clustering

How do you evaluate them?



https://devopedia.org/text-clustering

An example

	Finance	Sports	Politics	History
document1	0	2	1	0
document2	2	4	1	0
document3	0	0.4	2	0
document4	0	0	4	1

How can obtain an appropriate mapping of clusters to classes?

How to compare the clustering proportions?

Stop thinking of them as classes, but as vectorial representations of text, a.k.a. **embeddings**, which have replaced the words as features for classification

Take-home messages

- The point of clustering is unsupervised knowledge discovery
  - Why evaluate it against labels?
- If you want labels in the output, train your models with labels as part of the input
  - Can train with a small number of training instances
  - You would need some labeled data anyway to map clusters to classes
- If you are interested in unsupervised structure discovery, extrinsic evaluation is a must
  - Need to have a downstream application in mind (Ulrike Von Luxburg)
  - Useful clusterings would serve many downstream applications

Bibliography

Introduction to Information Retrieval, Christopher D. Manning, Prabhakar Raghavan & Hinrich Schütze, 2008. Cambridge University Press. https://nlp.stanford.edu/IR-book/html/htmledition/evaluation-of-clustering-1.html (We didn't cover information-theoretic measures such as normalised/adjusted mutual information, check the references if interested)

Clustering: Science or Art? Ulrike von Luxburg, Robert C. Williamson, Isabelle Guyon Proceedings of ICML Workshop on Unsupervised and Transfer Learning, PMLR 27:65-79, 2012. https://proceedings.mlr.press/v27/luxburg12a.html

V-Measure: A Conditional Entropy-Based External Cluster Evaluation Measure. Andrew Rosenberg and Julia Hirschberg. 2007. https://aclanthology.org/D07-1043/

Information Theoretic Measures for Clustering Comparison: Is a Correction for Chance Necessary? Nguyen Xuan Vinh, Julien Epps and James Bailey (2009). https://www.jmlr.org/papers/volume11/vinh10a/vinh10a.pdf