BMEG 3105
Lec11

Clustering and classification performance evaluation

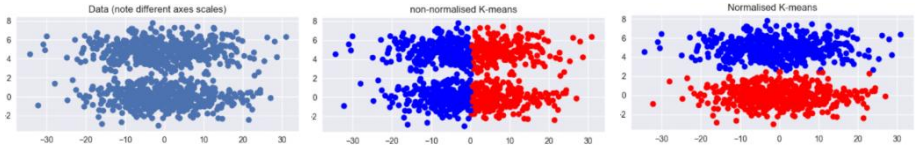
Yu LI
Friday, 10 October 2025

Purpose of model evaluation

- Characterize the performance of a model
  - Pinpoint the strong points and weak points of a method
  -
  - Method selection/Model selection
    - For clustering: normalization methods, distance measurements, distance between different clusters, …
    - For classification: normalization methods, distance measurements, K, …

Clustering performance evaluation

- Intra-cluster distances: small
- Inter-cluster distances: large



Classification evaluation

- Requires quantitative values to summarize the performance of different methods

Person	Height(m)	Weight(kg)	Gender
P1	1.79	75	M
P2	1.64	54	F
P3	1.70	63	M
P4	1.88	78	M
P5	1.75	70	??

Assign the class accurately

Person	Method 1	Method 2	Method 3
P1	M	F	M
P2	M	F	F
P3	M	M	M
P4	M	M	M
P5	F	F	M

1. Confusion matrix

		Predicted class	
		Class=Yes	Class=No
Actual class	Class=Yes	a(TP)	b(FN)
	Class=No	c(FP)	d(TN)

TP: True Positive
TN: True Negative
FP: False Positive
FN: False Negative

Accuracy

$$= \frac{a+d}{a+b+c+d} = \frac{TP+TN}{TP+TN+FP+FN}$$

Precision *(how many of the predicted positive samples are correct)*

$$= \frac{a}{a+c}$$

Recall *(how many actual positive samples are predicted to be positive)*

$$= \frac{a}{a+b}$$

F1 score *(the weighted average of precision and recall)*

$$= \frac{2 * precision * recall}{presicion + recall}$$

➔ Limitation: misleading for imbalanced data

Balanced Accuracy *(metric that accounts for class imbalance)*

$$= 0.5 * \left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP}\right)$$

***** No metric value is absolute. Context matters. *****

Example:

		Predicted class	
		Class=Yes	Class=No
Actual class	Class=Yes	2(TP)	0(FN)
	Class=No	50(FP)	50(TN)

accuracy = 51%
precision = 4%
recall = 100%

➔ Terrible model
➔ But because it misses zero cancer cases (0FN) it might be excellent for rare cancer pre-screening

KNN

Standard procedure:
1. Normalization
2. Compute distances
3. Identify the K most similar data
4. Take their class out and find the mode class

How to choose K when the data is all we have?
- A good K: good prediction accuracy
- Problem: we don't have the label for testing data
- Solution: use part of the training data as the testing data
    - User each part one by one -> calculate the average over the parts

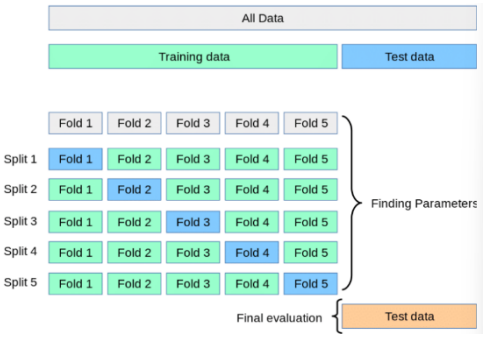
Procedure:
1. Hide label of one data point; the remaining points as training set for predicting the hidden label
2. Use a specific K
3. Record prediction
4. Repeat steps 1-3 for every data point
5. Calculate and compare accuracy of the trials

Cross-fold validation (/rotation estimation)

- To assess how the results of a machine learning analysis will generalize to an independent data set
- A procedure to measure the performance of models
- One round of cross-validation involves partitioning a set of data into complementary subsets, performing the analysis on one subset(training set), and validating the analysis on the other subset(testing set)

n-fold validation

- train multiple times, leaving out a disjoint subset of data each time for validation, then averaging the validation set accuracies



- process:
  1. randomly partition data into n disjoint subsets
  2. for i=1 to n
     - validation data= i-th subsest
     - h <- classifier trained on all data except for validation data
     - accuracy(i)= accuracy on h on validation data
  3. final accuracy= mean of the n recorded accuracies

leave-one-out cross-validation

- a special case of n-fold cross-validation, where n=N

- process:
  1. partition data into N disjoint subsets, each containing one data point
  2. for i=1 to N
     - validation data= i-th subsest
     - h <- classifier trained on all data except for validation data
     - accuracy(i)= accuracy of h on validation data
  3. final accuracy= mean of the N recorded accuracies

Multi-class classification

KNN: algorithm change not needed
Logistic regression: changes needed

- build a logistic regression for each class
- when predicting, assign class with highest value
- when training, train 3*6=18 parameters

Multi-class evaluation

- still using accuracy, precision, recall, F1 score, … (consider each class as a binary classification problem)
- to aggregate multiple values into one

$$Macro - average = \frac{0.9 + 0.95 + \cdots + 0.7 + 0.2}{6} = 0.73$$

$$Micro - average = \frac{0.9 * 150 + \cdots + 0.2 * 10}{150 + \cdots + 10} = 0.85$$

The low-performance of small classes will show up in Macro-average
More criteria at: https://scikit-learn.org/stable/modules/model_evaluation.html

Clustering evaluation

- correct as long as similar cells are in the same cluster (as opposed to only being correct for a cancer cell only if we predict it as cancer cell in classification)

		Predicted clusters	
		The same	Not the same
Actual clusters	The same	a(TP)	b(FN)
	Not the same	c(FP)	d(TN)

For all the pairs in the dataset (how many do we have?):

a: the number of pairs are in the same cluster in the True clusters and also assigned to one cluster in the Predicted clusters

b: the number of pairs are in the same cluster in the True clusters and also assigned to different clusters in the Predicted clusters

c: the number of pairs are in different clusters in the True clusters and also assigned to one cluster in the Predicted clusters

d: the number of pairs are in different clusters in the True clusters and also assigned to different clusters in the Predicted clusters

Rand index

$$R = \frac{a+d}{a+b+c+d} = \frac{a+d}{Number\ of\ all\ the\ pair\ combinations}$$

$$Pairs = \binom{n}{2} = \frac{n*(n-1)}{2}$$ n: Total number of points

Example:

Pair	Real	Predicted	Results
C1, C2	Same	Same	✓
C1, C3	Same	Different	×
C1, C4	Different	Different	✓
C1, C5	Different	Different	✓
C2, C3	Same	Different	×
C2, C4	Different	Different	✓
C2, C5	Different	Different	✓
C3, C4	Different	Same	×
C3, C5	Different	Same	×
C4, C5	Same	Same	✓

Cell	C1	C2	C3	C4	C5
Real cluster	0	0	0	1	1
Predicted cluster	2	2	3	3	3

How many pairs?

$$Pairs = \binom{5}{2} = \frac{5*(5-1)}{2} = 10$$

Rand index?

$$R = \frac{a+d}{a+b+c+d} = \frac{6}{10} = 0.6$$

More clustering performance evaluation: https://scikit-learn.org/stable/modules/clustering.html#clustering-performance-evaluation

Potential project-2,3

Data preprocessing for the gene expression matrix

➢Data collecting and merging (if needed)
➢Exploration
➢Visualization
➢Data cleaning
➢Dimension reduction (next lecture)
➢Get distance matrix
➢Perform classification/clustering
➢Performance evaluation

Model evaluation in Python

https://scikit-learn.org/stable/modules/generated/sklearn.metrics.classification_report
https://scikit-learn.org/stable/modules/clustering.html#clustering-performance-evaluation
https://scikit-learn.org/stable/modules/cross_validation.html

Resources and uncovered topics

❖Introduction to data mining: Chapter 4.5 & 4.6 & 5.7 & 5.8 & 8.5
❖Bootstrap
❖Overfitting and generalization
❖Other clustering and classification methods
❖Comparison between different methods
    ➢Clustering
    ➢Classification