Scribing: Classification perf –lecture 10

Name: Lai Jacobi Wing Ki SID: 1155159737

1 Find a good K for KNN

Standard KNN have chosen the distance metric and K, but we don’t know if the k we choose is the best. Thus, we need to find a good K.

1.1 Good K

- The K can give us good prediction accuracy

1.2 How to find a good k

- First use part of the training data as the testing data
- Choose k=1
- Thirdly use each part one by one
- After that calculate the average over all the parts
- Then try other k value
- Lastly select the k with highest accuracy

1.2.1 Example

Person	Height	Weight	Gender
P1	0.625	0.875	M
P2	0	0	F
P3	0.25	0.375	M
P4	1	1	M
P5	0.4583	0.6667	??

	P1	P2	P3	P4
P1	0	0.875	0.5	0.[illegible]
P2	0.875	0	0.375	
P3	0.5	0.375	0	[illegible]
p4	0.375	1	0.75	

Compute distances (here is Euclidean distance)

K=1	**P1: P4—M** **P2: P3—M** **P3: P2—F** **P4: P1—M**	Accuracy = 0.5
K=2	**P1:P3 P4-M** **P2: P1 P3-M** **P3: P1 P2-F/M** **P4: P1 P3-M**	Accuracy = 0.5/ 0.75
K=3	**P1: M** **P2: M** **P3: M** **P4: M**	Accuracy = 0.75

Thus, we choose K=3

2 Cross-fold validation

- it is a technique for assessing how the results of a machine learning analysis will generalize to an independent data set
- A procedure to measure the performance of models
- Involves partitioning a set of data into complementary subsets - performing the analysis (training set), and validating the analysis (testing set)

2.1 n-fold cross-validation

◊ train multiple times
◊ leaving out a disjoint subset of data each time for validation
◊ average the validation set accuracies

2.1.1Process of n-fold cross-validation

❖ **Process:**

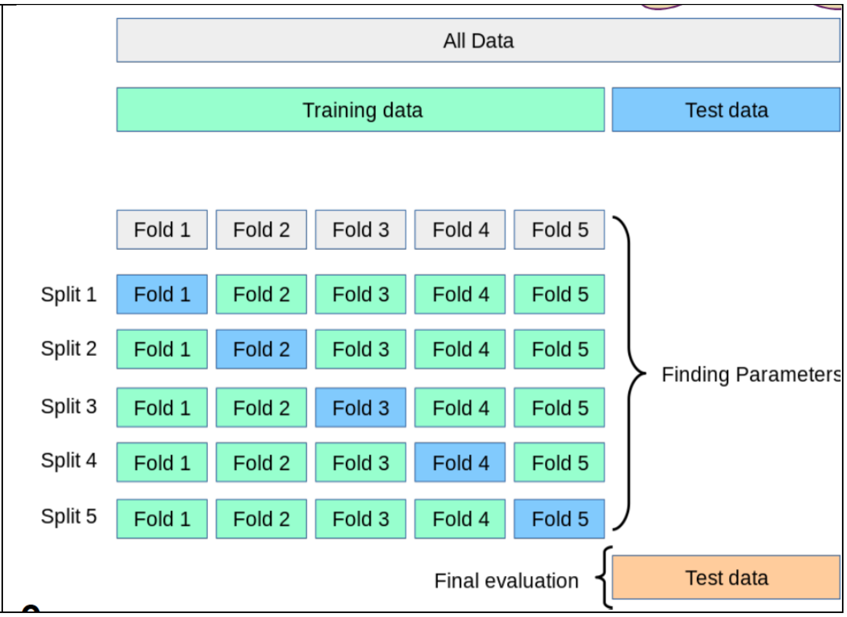
- ➢ Randomly partition data into n disjoint subsets
- ➢ For i = 1 to n
  - Validation Data = i-th subset
  - h <- classifier trained on all data except for Validation Data
  - Accuracy(i) = accuracy of h on Validation Data
- ➢ Final Accuracy = mean of the n recorded accuracies

2.1.2Example -5-fold cross-validation

10 data • P1-P10 5-fold • P1-2, P3-4, P5-6, P7-8, P9-10 • Can random group	

Procedure

- P1-2's results based on the model from P3-10
- P3-4's results based on the model from P5-10 and P1-2
- …
- P9-10's results based on the model from P1-8
- Averaging



2.2 Leave-one-out cross-validation

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

Process:

- Partition data into N disjoint subsets, each containing one data point
- For i = 1 to N
  - Validation Data = i-th subset
  - h <- classifier trained on all data except for Validation Data
  - Accuracy(i) = accuracy of h on Validation Data
- Final Accuracy = mean of the N recorded accuracies

3 Multi-class classification

KNN **logistic regression**

• it is trivial • No need to change the algorithm	• need some changes • When predicting, we assign class with highest value • When training, we train 3*6=18 parameters

❖ Using accuracy, precision, recall, F1 score
❖ Considering each class as a binary classification problem

3.1 Ways to aggregate multiple values into one value

$\rightarrow$ Macro – average
$\rightarrow$ Micro – average

3.1.1 Example

Class	Accuracy	Cells
1	0.9	150
2	0.95	50
3	0.85	100
4	0.8	40
5	0.7	20
6	0.2	10

Macro – average

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

Micro – average

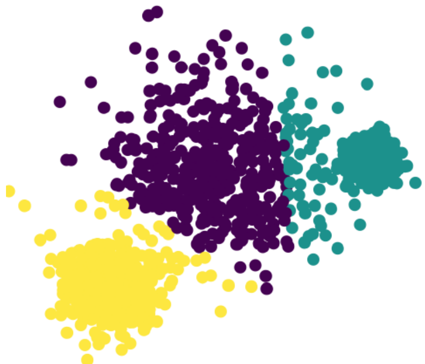
$$\frac{0.9 * 150 + \cdots + 0.2 * 10}{150 + \cdots + 10} = 0.85$$

★ The low performance of small classes will show up in Macro-average

4 Clustering evaluation

4.1 Clustering vs Classification

Clustering	Classification

we are correct as long as similar cells are in the same cluster	we are correct for a cancer cell only if we predict it as cancer cell
 True label	 Predicted label
good clustering	messy classification

4.2 How to evaluate clustering?

1 First evaluate a pair of cells

2 Second made a confusion matrix

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

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

3 Third, calculate the Rand index and Pairs

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

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

4 Finally, the Rand index closer to 1 ,the cells are more likely in the same cluster.

4.2.1 Example

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

sample data

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	√

After pairing and comparison

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

Rand index =

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