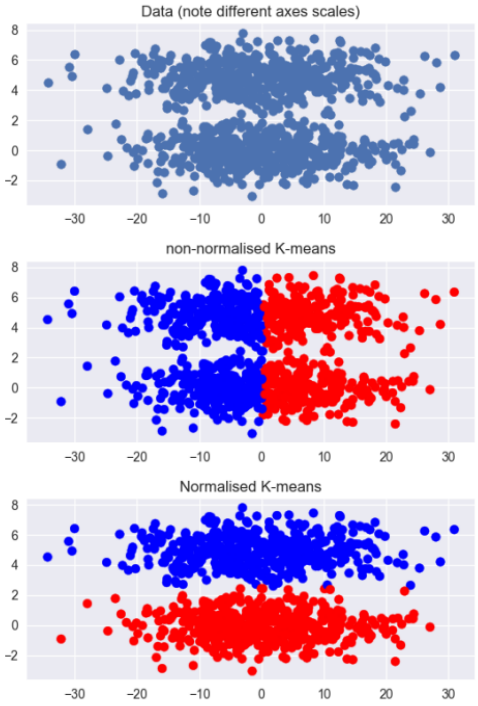
Course Code : BMEG 3105
Course Title : Data Analytics for Personalized Genomics and Precision medicine
Lecture Topic: Performance Evaluation
Scribed by Kyaw Saw Lin (1155173213@link.cuhk.edu.hk)

Contents

1. Binary classification evaluation
2. Cross Validation
3. Multi-class classification
4. Clustering evaluation

We do performance evaluation to analyze which clustering and classification methods are suitable for particular purposes.
In clustering - the goal is to make the distances between <u>inter-cluster classes to be large</u> while <u>intra-cluster classes to be small</u>.


Data (note different axes scales)

non-normalised K-means

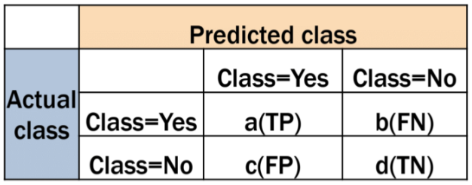
Normalised K-means

In this example: different clustering methods are applied to data. Normalized K-means method achieves the goal of clustering more than Non-normalised K-means. But how can we get the quantitative values to summarize the performance of different methods?

What can get varied between different clustering methods? ● Normalized methods ● Distance measurements ● Choose of K	What can get varied between different clustering methods? ● The way you normalization ● distance measurements

1. Binary classification evaluation

With the use of Confusion matrix



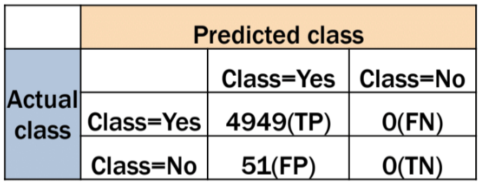
		Predicted class	
		Class=Yes	Class=No
Actual class	Class=Yes	a(TP)	b(FN)
	Class=No	c(FP)	d(TN)

Confusion matrix

You can calculate the accuracy of the classification method by

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

One drawback is this evaluation formula is prone to imbalanced data.



		Predicted class	
		Class=Yes	Class=No
Actual class	Class=Yes	4949(TP)	0(FN)
	Class=No	51(FP)	0(TN)

$$Accuracy = \frac{TP+TN}{TP+TN+\mathrm{FP}+TN} = \frac{4949}{4949+51} = 0.99$$

In the above table, it is most likely that this is a bad classifier and predicts everything as YES but the accuracy formula still believes the accuracy is very high. So, this formula is misleading for the imbalance data. There are other evaluation formulas such as

$$Precision = \frac{a}{a+c} \qquad Recall = \frac{a}{a+b} \qquad F1\ score = \frac{2*precision*recall}{presicion+recall} = 0.995$$

But if you use those, all of them give misleading results for imbalance data.

We can solve this imbalanced data accuracy by a new formula which successfully points out the bad classifier.

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

To conclude, aside from applying these formulas to assess the classifier performance, it is most important to take into account the context of the data.

2. Cross Validation

Standard procedure of KNN

- We choose distance matrix and K
- Normalization
- Compute distances
- Identify the K most similar data
- Take their class out and find the mode class

- But we do not know what value K in KNN is most suitable for the particular application.

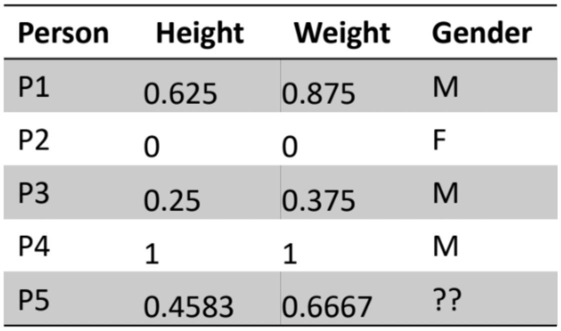
Therefore, we find out using cross validation.

- Good K value is the one that gives good prediction accuracy. But we do not have the testing data (dun have labels to check with) while we are training to see the accuracy.

We use part of training data as testing data by splitting training data to different parts and do the training on all parts except one and test it on that one part and get the accuracy. We then average all the accuracies. We will choose the K value which gives the highest prediction accuracy among all.

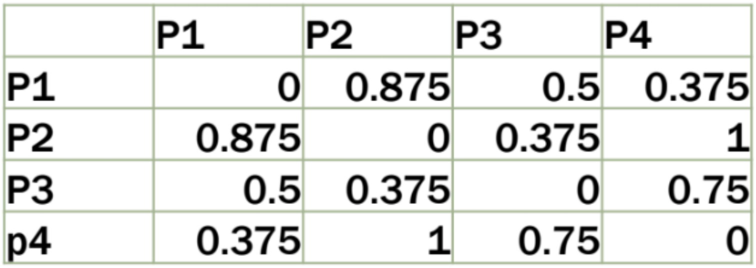
Working example:

We want to find the good K value for classification of following data with a data matrix.



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	??

Given Data



	P1	P2	P3	P4
P1	0	0.875	0.5	0.375
P2	0.875	0	0.375	1
P3	0.5	0.375	0	0.75
p4	0.375	1	0.75	0

Data matrix

P5 is the actual testing data and we have P1-5 as training data.

Suppose K=1	Suppose K=3
For P1, in the data matrix, the most similar one to P1 is P4 - having the smallest value close to 0. This means P1 gender is the same as P4. If you input P1 information to the current model, it will say gender = Male. which is correct and we noted accuracy as 1.	Since the nearest neighbor is 3. We will find the mode of gender of the nearest 3 neighbors. For P1, we will find the mode of P2,P3,P4 which is Male and actual gender of P1 is also Male, so accuracy = 1.
For P2, the data matrix says it is the same as P3 gender which is Male but actual P2 gender is female so, our model is wrongly predicting. So accuracy for P2 is 0.	For P2, mode of P1,P3,P4 is Male and actual P2 gender is Female. So accuracy = 0.
For P3, our data matrix says P3 is like P2 which is female but P3 actually is male, so it is wrongly predicting again. Accuray for P3 is 0.	For P3, the same goes like P1 and P2. accuracy=1
For P4, our data matrix says P4 is like P1 which is male and actual P4 gender is male. So accuray for P4 is 1.	Compute the same as above columns, accuracy=1
So total average accuracy of choosing K= 1 is (1+0+0+1)/4 = 0.5	Accuracy for choosing K=3 is (1+0+1+1)/4 = 0.75. **So we choose K=3 and Gender of P5 is Male.**

3. Multi-class classification

In reality, it is common to have more than two classes.

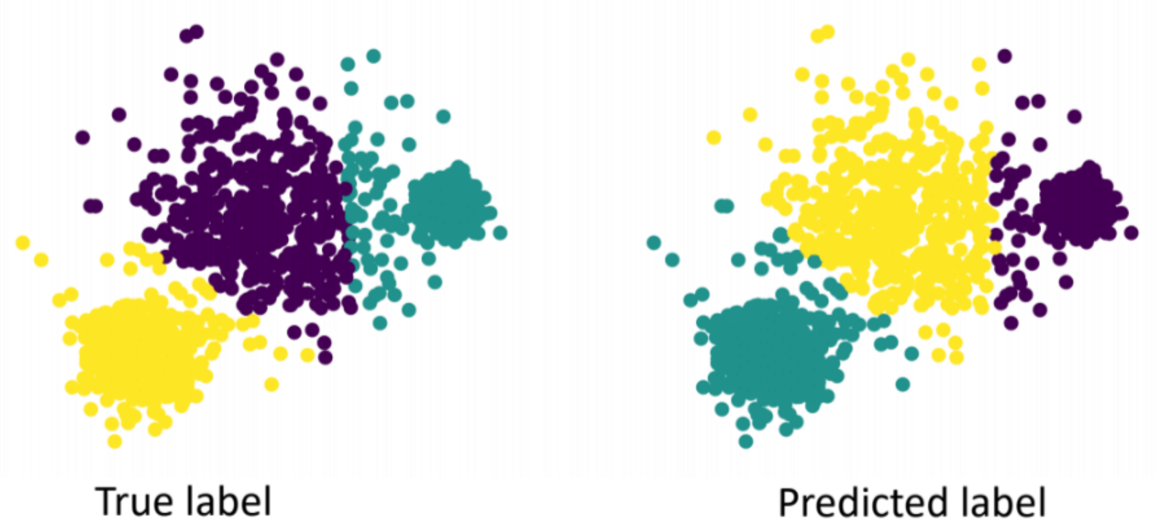
- The same as KNN algorithm
- But need logistic regression for each class
- Run all those classifiers and assign classes with highest value

For multi-class evaluation, considering each class as a binary classification problem, each class has their own accuracy so we have to find the aggregate and there are two ways for that.

Macro-averaging - performance of smaller classes	Micro-averaging - performance of larger classes
Total accuracy / number of classes	∑ (accuracy * S)/ ∑ S

4. Clustering evaluation

As long as two similar items are in the same cluster, that clustering method is correct. Classification has to point out whether this class belongs to this label or not exactly but clustering doesn't need to care about their label, just need the similar items in a group together.

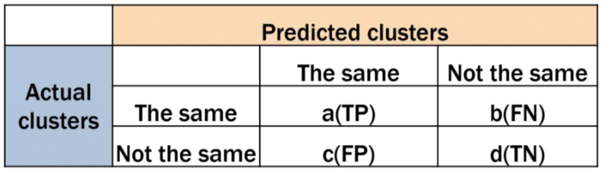


True label

Predicted label

Above prediction is very bad classification performance but quite good clustering accuracy.

To calculate clustering accuracy is similar to binary classification evaluation, but here you observe the items in pairs.

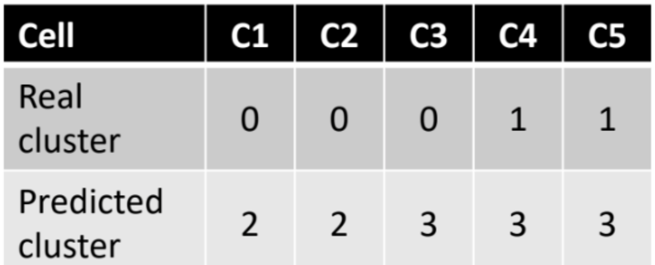


		Predicted clusters	
		The same	Not the same
Actual clusters	The same	a(TP)	b(FN)
	Not the same	c(FP)	d(TN)

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

- Rand index R tells how much accuracy the cluster is.
- Number of pair combinations can be calculated by -

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



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

For example, C1 and C2, both are 0 gp in the real cluster while both are in 2 gp in the predicted cluster. This condition satisfies TP. For C2 and C3, both are in the same gp in real but different gp in predicted, so this condition is FN. You continue doing this and finally, you count the number of TP and TN. and substituted it into the R formula above.