BMEG3105: Data analytics for personalized genomics and precision medicine

Lecture 11:

Clustering and classification performance evaluation

Lecturer: Prof. LI Yu *Scriber: ZHANG, Laibohan*

Outline of Lecture 11:

- Recap from last lecture
- Performance evaluation
- Cross-validation
- Multi-class classification
- Clustering evaluation

Part 1. Recap from last lecture

A. Logistic regression

Logistic regression is a statistical method used to predict the classification of test data. For example, for a simple 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	??

Figure 1. Normalized data matrix

To predict the Gender of P5, we can propose a formula first, which is a function derived from observation of training data.

❖ $\frac{1}{1+e^{-(w_h H + w_W W + w_0)}} \geq 0.5$

Then using training data to train this model, and obtaining the values of parameters that make the model fit the training data.

B. Gradient descent algorithm

It's an algorithm used to find the parameters of the assumed model. The working procedure is:

1. initialize parameters (wh, ww & w0) using random values
2. calculate the function output for each training data
3. update the parameters using the following formula
   - $w_i = w_i + \Delta w_i$
   - $\Delta w_i = 2 * \alpha(Y - Y^{output}) \frac{\partial Y^{output}}{\partial w_i}$
4. repeat the above step until no parameters need to be updated

C. Loss function

$$L = \sum_{P_1}^{P_4} (Y^{output} - Y)^2$$

This function quantitatively calculates the differences between the current model and the actual training data. For each w, we want to find a value to make the function value smallest.

Part 2. Performance evaluation

A. Classification performance evaluation

- We can get various models using different method, but which classification method should we trust?
- We need some quantitative values to summarize the performance of different methods

B. The purpose of model evaluation:

1. Characterize the performance of a model
2. Pinpoint the strong points and weak points of a method
3. Method selection/Model selection

C. Confusion matrix

The confusion matrix is a table used to evaluate the performance of a classification model.

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

Figure 2. Confusion matrix

D. Most widely used metric: Accuracy

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

Usually: good classifier will have higher accuracy.

But there is a serious limit:

- For imbalanced classes, using Accuracy may be misleading:

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

Imbalanced classes

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

Maybe misleading for imbalanced data

E. Other metrics:

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

Among the predicted positive samples, how many of them are correct?

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

How many actual positive samples are predicted to be positive?

$$F1\ score = \frac{2*precision*recall}{presicion+recall}$$

The weighted average of precision and recall

F. Special case:

Considering the following matrix: is it a terrible prediction?

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

Figure 3. Special case

- If we only calculate the accuracy, then it will become a terrible prediction
- But when considering the actual situation, the result can be quite different: if this is a model used for predicting cancer, then this model can avoid missing out on potential patients.

Part 3. Cross-validation

A. What is cross-validation]

Consider a situation:

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	??

Figure 4. Sample matrix

For these data matrix, we want to apply KNN method. But how can we choose a gook K without the class of P5?

Solution: Applying <u>cross-validation</u>

- Use part of the training data as testing data
- Use each part one by one
- Calculate the average of all the parts

Cross-validation/rotation estimation is a technique for assessing how the results of a machine learning analysis will generalize to an independent dataset.

- 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 (called the training set), and validating the analysis on the other subset (called the testing set).

B. N-fold cross-validation

- Idea: train multiple times, leaving out a disjoint subset of data each time for validation. Average the validation set accuracies
- Process:
  1. Randomly partition data into n disjoint subsets
  2. 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
  3. Final Accuracy =mean of the n recorded accuracies

C. Leave-one-out cross-validation

- Idea: 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 subset
     - ✧ 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

Part 4. Multi-class classification

A. Considering data with more than 2 classes

For this situation, we have different approaches to deal with:

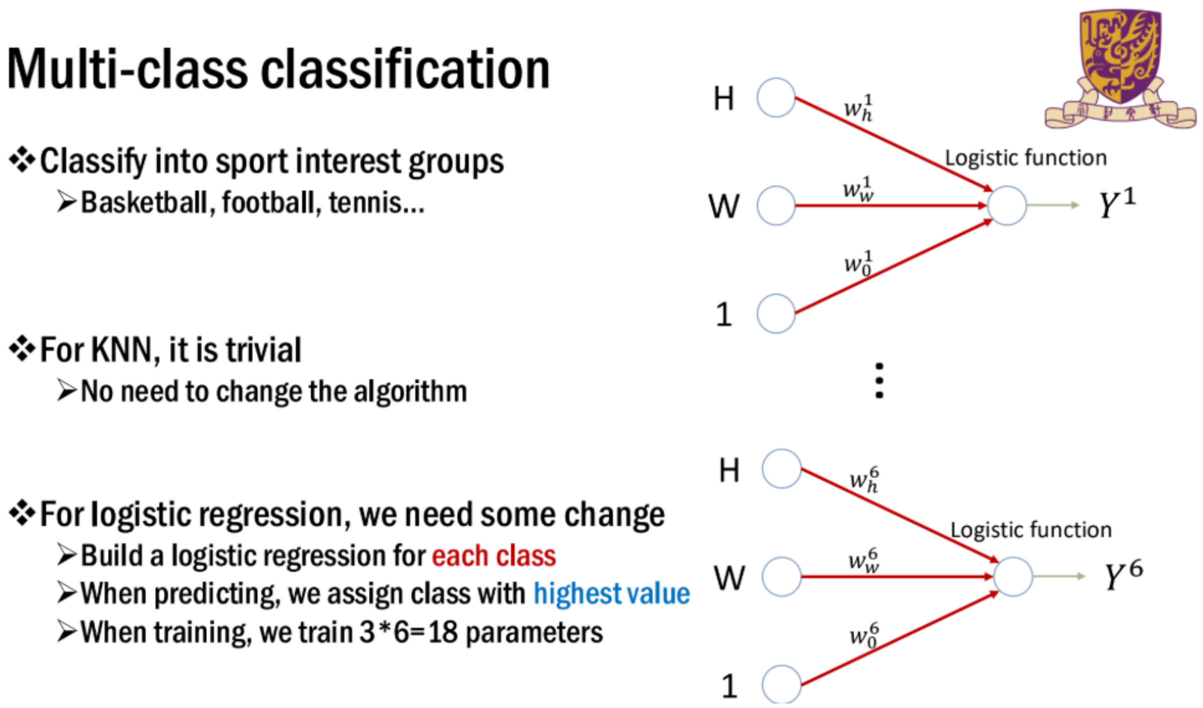


Figure 5. Different approaches

B. Multi-class evaluation

Still using accuracy, precision, recall, F1 score and so on.

✧ Considering each class as a binary classification problem.

But how to aggregate multiple values into one value?

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

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

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

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

Part 5. Clustering evaluation

A. Main difference between clustering and classification

✧ In classification, we are correct for a cancer cell only if we predict it as cancer cell;

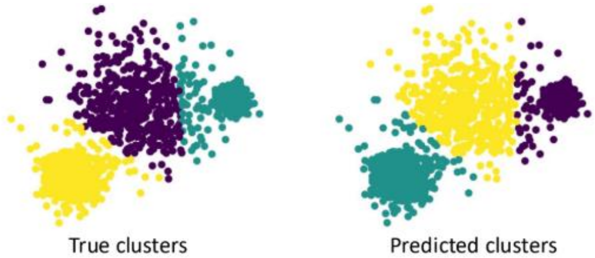
✧ In clustering, we are correct as long as similar cells are in the same cluster.

That is: Classification needs the accurate class of different data, but Clustering only needs the correct grouping (The specific group labels are not important).

B. How to evaluate clustering

We should evaluate a pair of cells (using confusion matrix):

		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

raluation Yu Li Lecture 10-43

Figure 6. Clustering evaluation

Using 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

References

Li, Yu (2025). “Clustering and classification performance evaluation”.