BMEG 3105

Data analytics for personalized genomics and precision medicine

Lecture 9: Clustering and classification performance evaluation

Lecturer: Yu LI

Scriber: Ng Tsoi Yin

Wednesday, 3 October 2022

Outline of the Lecture:

- Performance Evaluation
- Cross-Validation
- Multi-class Classification
- Clustering Evaluation

Content:

- **Training**
  - What is training ?
    - To get w_w, w_w, w_0
    - To make model fit the training data
    - To make $\frac{1}{1+e^{-(w_h H+w_w W+w_0)}} \geq 0.5$ correct
  - $Y^{output} = \frac{1}{1+e^{-(w_h H+w_w W+w_0)}}$
  - Loss function L
    - Loss function = $(Y^{output} - Y)^2$
    - Example: Data from P1-P4

Person	Height	Weight	Gender
P1	0.625	0.875	M

    - For P1,
    - $L = (Y^{output} - Y)^2 = (1 - \frac{1}{1+e^{-(w_h H+w_w W+w_0)}})^2$
    - In a whole, $L = \sum_{P_1}^{P_4}(Y^{output} - Y)^2$
    - L = the smaller, the better

- **Gradient descent algorithm**
  - Aim: fine smallest value for L using different values of w
  - Step 1: initialise w_w, w_w, w_0

    Step 2: For each data (P1,P2,P3,P4)

    Calculate Y^{output}

    Update new weights

    New $w_i = w_i + \Delta w_i$

    $\Delta w_i = 2\alpha(Y - Y^{output})\frac{\partial Y^{output}}{\partial w_i}$, α is a constant

    Repeat until no update

- **Performance evaluation -binary classification evaluation**
  - Purpose: pinpoint strong points and weak points of one method→model selection
  - Method: confusion matrix

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

  - Accuracy $= \frac{a+d}{a+b+c+d} = \frac{TP+TN}{TP+TN+FP+FN}$
  - Higher accuracy, better the classifier is
  - Exception: when there is imbalanced classes
  - $Precision = \frac{a}{a+c}$ $Recall = \frac{a}{a+b}$ $F1\ score = \frac{2(precision)(recall)}{precision+recall}$

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

- **Cross-validation**
  - KNN
    - Standard procedure
    - After chosen distance metric and K,
      1. Normalization
      2. Compute distances
      3. Identify the K most similar data
      4. Take their class out and find the mode class
    - Good K = good prediction accuracy
    - Problem: no label for testing data

Solution: use part of training data as testing data

(use each part one by one and calculate the average)

- Can use L_∞ for convenience

- Cross-fold validation
  - Procedure to measure the performance of models
  - N-fold cross-validation:

    Step 1: randomly partition data into n disjoint subsets

    Step 2: for i=1 to n,

    Validaton data =i-th subset

    H← classifier trained except validation data

    Accuracy (i) = accuracy of h

    Step3: final accuracy = mean of n recorded accuracies

- **Multi-class classification**
  - Consider each class as binary classification problem
  - Aggregate multiple values into one value:
  - Macro-average = $\frac{sum\ of\ accuracy\ of\ each\ class}{number\ of\ classes}$
  - Micro-average = $\frac{sum\ of\ (accuracy*number\ of\ data\ of\ the\ class)}{number\ of\ data}$

- **Clustering evaluation** (different from classification !!)
  - Messy classification can be a good clustering
  - Should be evaluate a pair of cells
  - Rand index R=$\frac{a+d}{a+b+c+d} = \frac{a+d}{Number\ of\ all\ pair\ combinations}$
  - Number of pairs =$\binom{n}{2} = \frac{n(n-1)}{2}$