Lec 11 Scribing

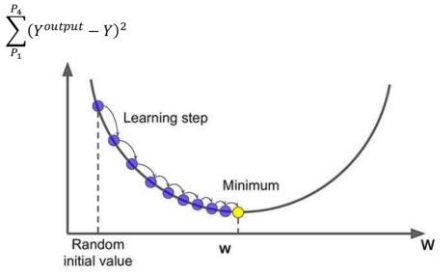
- **Recap:**

  Logistic Regression:

  1. Logistic function & Loss function
  2. Training Procedure:

$$w_i = w_i + \Delta w_i$$
$$\Delta w_i = 2 * \alpha(Y - Y^{output}) \frac{\partial Y^{output}}{\partial w_i}$$
$$\alpha \text{ is a small constant}$$

  3. Gradient descent algorithm



- **New Content:**

  Model evaluation:

1. Purpose of model evaluation:

   Characterize the performance of a model

   ➢Pinpoint the strong points and weak points of a method

   ➢Method selection/Model selection

2. Classification Performance Evaluation

   1) Confusion Metrix eg. Gender Classification

   a.

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

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

   b. Accuracy: (Most widely-used metric)

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

   *Imbalanced classes maybe misleading for imbalanced data.

c. Precision, recall, and F1 score:

$$Precision = \frac{a}{a+c} = \frac{4949}{4949+51} = 0.99$$

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

$$Recall = \frac{a}{a+b} = 1$$

Recall: How many actual positive samples are predicted to be positive?

$$F1\ score = \frac{2 * precision * recall}{presicion + recall} = 0.995$$

F1 score: The weighted average of precision and recall

*However, still may be misled by imbalanced data

d. Balanced accuracy:

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

e. *Personal Method Dealing with Imbalanced Data :if knowing it's an imbalanced dataset, suggest to look at the confusion matrix directly

2) How to find a good K for KNN with the below data?

a. What is a good K?

The K can give us good prediction accuracy

b. Method: (Get the distance metric first)

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

use part of the training data as the testing data

➢Use each part one by one

➢Calculate the average over all the parts

Eg. Gender Classification

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

⇨ Select K=3

2.1 Cross-fold Validation/Rotation estimation:

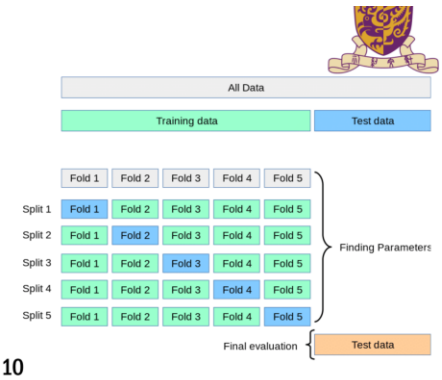
1) Def: a technique for assessing how the results of a machine learning analysis will generalize to an independent dataset.
2) Procedure: Partitioning a set of data into complementary subsets;
   Performing the analysis on one subset (the training set);
   Validating the analysis on the other subset (the testing set)
3) n-fold cross-validation:
   a. Train multiple times, leaving out a disjoint subset of data each time for validation. Average the validation set accuracies.
   b. 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
   Eg.5-fold cross-validation

5-fold cross-validation

❖10 data points:
➢P1-P10

❖5-fold
➢P1-2, P3-4, P5-6, P7-8, P9-10
➢The grouping can be random

❖Procedure
➢P1-2's results based on the model from P3-10
➢...
➢P9-10's results based on the model from P1-8
➢Averaging



4) Leave-one-out cross-validation
   a. a special case of n-fold cross-validation, where n = N

2.2 Multi-class Classification

1) For KNN, it is trivial ➢No need to change the algorithm
2) For logistic regression, we need some change:
   ➢Build a logistic regression for each class
   ➢When predicting, we assign class with highest value
   ➢When training, we train 3*6=18 parameters

3) Still using accuracy, precision, recall, F1 score and so on

➢Considering each class as a binary classification problem

❖**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

3. Clustering evaluation

1)How to evaluate Cluster: In clustering, we are correct as long as two similar cells are in the same cluster

➢We should evaluate a pair of cells

➢We also have a confusion matrix

		Predicted clusters	
		The same	Not the same
Actual clusters	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

2)Rand Index:

Rand index, R:

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

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

Eg.

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?

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

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	✓

3)More clustering performance evaluation:

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

*Model evaluation in Python:

https://scikit-learn.org/stable/modules/generated/sklearn.metrics.classification_report

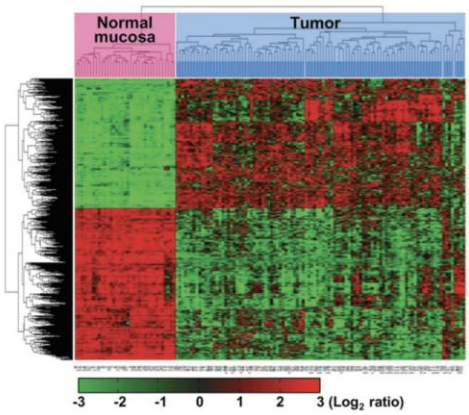
https://scikit-learn.org/stable/modules/clustering.html#clustering-performance-evaluation

https://scikit-learn.org/stable/modules/cross_validation.html

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**



Resource 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