Lec 10 Scribing

- ## Recap

Classification:

1. What:

   Given a collection of records (training set) Each record contains a set of attributes, one of the attributes is the class; Find a method to assign the class of previously unseen records based on their other attributes and the training set as accurately as possible

2. How:

   a. use training data to create a classification method

   b. classify the data with the method

3. Procedure of KNN:

   a. Normalization

   b. Compute distances

   c. find K instances

   d. return the most frequent class label among K instances

4. Classification vs Clustering:

	Clustering	Classification
Goal	Find similarity (clusters) in the data	Assign class to the new data
Data	Data without class	Training data with class and testing data without class
Classes	Unknown number of classes	Known number of classes
Output	The cluster index for each point	The class assignment of the testing data
Algorithm	One phase	Two phases (training and application)

5. Shortcomings of KNN and Adjustments:

   a. Shortcomings: Space costly--need to store all data;

   Computility --Need to calculate dis metrix

   Predicting is slow

   b. Adjustments: generate a formula

- ## New Content

Logistic Regression:

1. Logistic function:

eg. Classify gender

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	??

a. Generate a formula:

$H + W \geq 0.5$ -> Male

b. Add weights w(h),w(w) and bias w0 if wh,ww,w0 are large, we use:

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

c. Training: fit the training data to get wh,ww,w0

$Y^{output} = \frac{1}{1+e^{-(w_h H + w_W W + w_0)}}$

➢ **1 for male, 0 for female**

(logistic function)

2. Loss function:

$(Y^{output} - Y)^2$ **is a function of** ws

Y**: the true label we have for training data**

For P1

➢ $(Y^{output} - Y)^2 = \left(1 - \frac{1}{1+e^{-(0.625*w_h + 0.875*w_W + w_0)}}\right)^2$

$L = \sum_{P_1}^{P_4}(Y^{output} - Y)^2$ **is a function of** ws

➢ **Goal: find** ws **to make** L **the smallest**

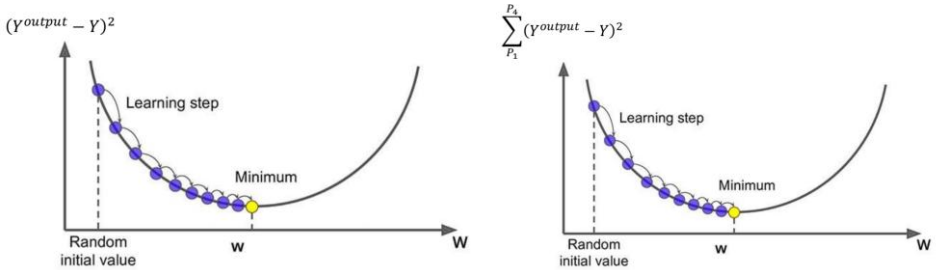
(L: loss function)

*How to find the minimum value for **L**?-- Gradient descent algorithm

3. Logistic Regression Model Training—Gradient descent algorithm

1)Def:

an algorithm to update parameters in a model (can identify the local min and may never exactly reach the min point *Ureply)



2)Method: Go the direction where the gradient of the function decreases until we reach the point that gradient=0

(*Ureply: the initialization of function WS can affect the training time)

3)Procedure: (included in the “fit” function of python)

a. Initialize wh and, ww and w0 with Random values

b. calculate the Y(output) of P1,P2,P3,P4

c. update the weights

$$w_i = w_i + \Delta w_i$$

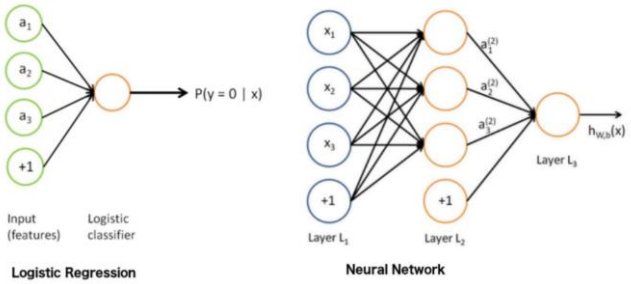
$$\Delta w_i = 2 * \alpha(Y - Y^{output}) \frac{\partial Y^{output}}{\partial w_i}$$

α **is a small constant**

d. Repeat the above step, until no more to update

4. From logistic regression to neural networks

1) From **LR** to **NN:**



Advantage: Fast prediction; Successful in real-life problems;High tolerance to noisy data

Disadvantage: Long training time; Poor interpretability

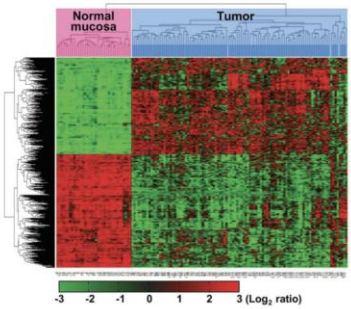
2)From **NN** to **DL**:

Eg.AlphaFold

Potential Project 3:

❖**Data preprocessing for the gene expression matrix**

- ➢**Data collecting and merging (if needed)**
- ➢**Exploration**
- ➢**Visualization**
- ➢**Data cleaning**
- ➢**Get distance matrix**
- ➢**Perform classification**



- **Next Lecture:**

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	
Actual class		Class=Yes	Class=No
	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