BMEG 3105 Fall 2022

Data analytics for personalized genomics and precision medicine

Course introduction

Lecturer: Yu LI (李煜) from CSE

Liyu95.com, liyu@cse.cuhk.edu.hk

Friday, 30 September2022

Outline of lecture

I. **Recap from last lecture**
   A. Classification
   B. K-nearest neighbour classification
   C. Clustering VS Classification
   D. Logistic regression

II. **Today's agenda**
   A. Logistic Regression Model Training
      a) How to train?
         1. Standard process
         2. Loss Function
         3. How to minimize
         4. To get the formula

   B. From logistic Regression to Neural Networks
      a) The simplest neural networks
      b) From logistic regression to neural network
      c) From network to deep learning

III. **Self-study material**

I. Recap from last lecture

A. Classification

a) What is classification?
Given that there is a collection of records (Training set)
Find a method to assign the class of previously unseen records based on their other attributes and the training set as accurately as possible.

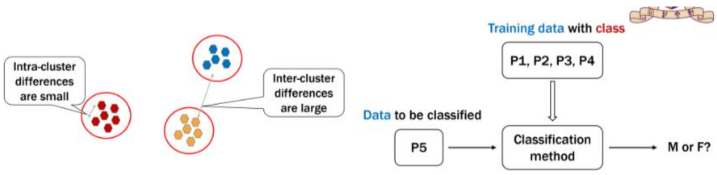
b) How to do classification
First, get the training data with class
Second, build a classification method
Third, insert the data to be classified
Then, get the result

B. K nearest neighbour classification

Standard process of KNN:

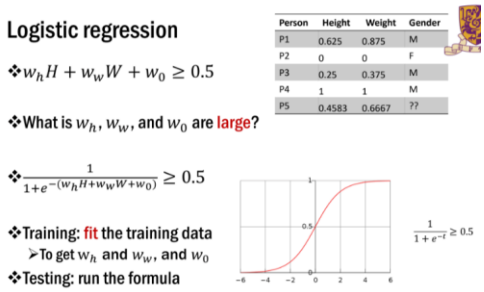
1. Normalization
2. Computer distances
3. Identify the K most similar data
4. Take their class out and find the mode class

C. Clustering VS Classification



	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)

D. Logistic regression



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	??

II. Today's agenda

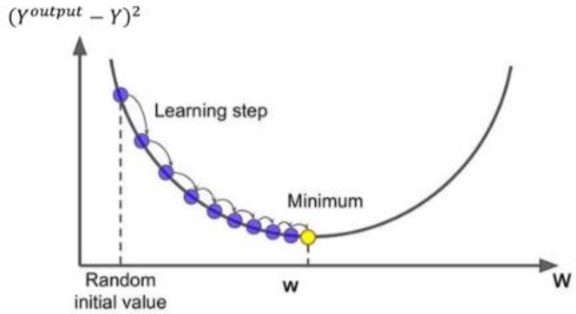
A. Logistic Regression Model Training

a) How to train?

1. Standard Process
   - Get W_h, W_w, and W_0, (Initialization)
   - Make the model fit the training data
   - **Make** $\frac{1}{1+e^{-(w_h H + w_w W + w_0)}} \geq 0.5$ **correct for the training data**
   - $Y^{output} = \frac{1}{1+e^{-(w_h H + w_w W + w_0)}}$ (1 for male and 0 for female).

2. Loss Function
   - As we want Y^{output} should be as close as true data possible.
   - $(Y^{output} - Y)^2$ should be as small as possible. (Y is the true data we have for training data)

3. How to minimize
   - Use differentiation to find the minimum, when $\frac{dy}{dx} = 0$, we reach the minimum.
   - We can use Gradient Descent Algorithm:
     $(Y^{output} - Y)^2$ is a function of w
     For each w, we want to find a value to make the function value smallest.



Step:

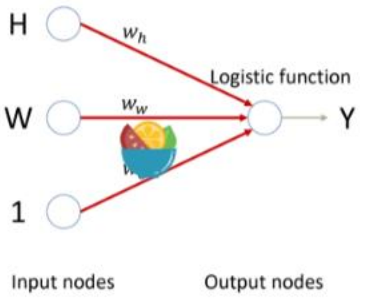
1. Find $\frac{dy}{dx}$ of $(Y^{output} - Y)^2$ at a point A.
2. If $\frac{dy}{dx}$ result is negative, we choose another point B to get a result.
3. Repeat this process to let the $\frac{dy}{dx}$ closer to 0
4. Until very close to 0 (maybe not 0)
5. Then it is the minimum we find

4. To get the formula
    - Initialize W_h, W_w, and W_0 (Random values)
    - Calculate the output Y^{output}
    - Update 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
    - Repeat the above steps until no more update

B. From Logistic Regression to Neural Networks

a) The simplest neural networks

One input node and one output nodes

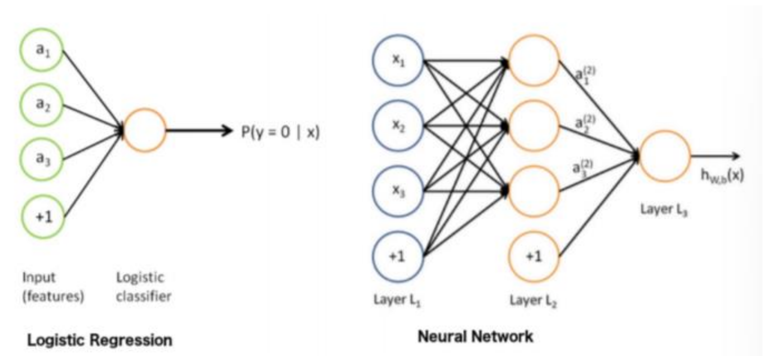


$$y^{output} = \frac{1}{1 + e^{-(w_h H + w_w W + w_0)}}$$

b) From logistic regression to neural networks

Several input nodes and several output nodes

We can create function of functions (such as layer L_3 below)

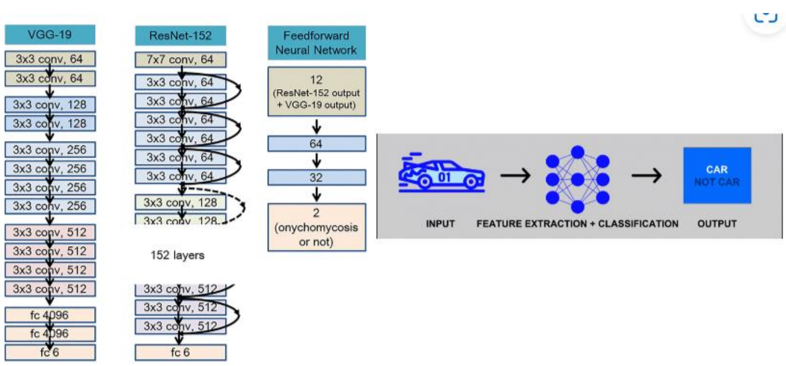


Advantage: Fast prediction; Successful in real-life problems; High tolerance to noisy data

Disadvantage: Long training time; Poor interpretability

c) From neural networks to deep learning

Successful deep learning application: AlphaFold



III. Self-learning Materials

Introduction to data mining: Chapter 5.4 & Appendix E

- Problems of neural network
- Decision tree/SVM/Bayesian...
- Model overfitting
- Cross-validation (next lecture)