Data Mining: Classification 3

Jakramate Bootkrajang

School of Computer Science
Chiang Mai University

Adapted from materials by Ata Kabán

Outline

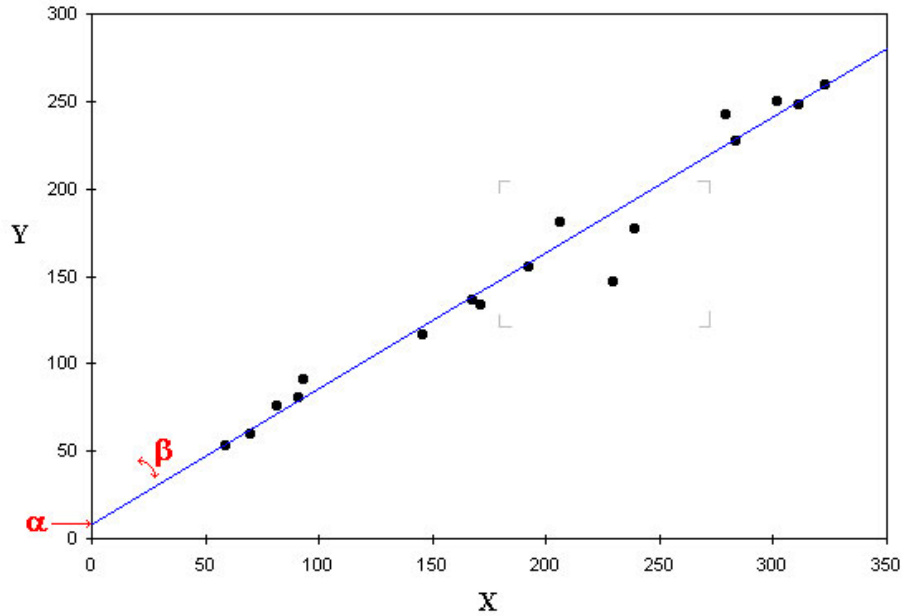
- Introduction
- Bayesian Learning
- Generative Classifier
- Discriminative Classifier
- Nearest Neighbour Classifier
- Classifier Evaluation [Maybe]

Introduction

- We have seen how we can construct a generative classifier.
- The generative classifier puts some assumption on the data (in our case Gaussian assumption)
- This lecture we will learn the counterpart of generative classifier called *discriminative* classifier
- Discriminative classifier aims to classify data directly without modelling data distribution.

Linear Regression

- To find linear relationship between $\mathbf{x}$: input and y: the real valued response.
- The function is $\hat{y} = \mathbf{w}^T\mathbf{x} + \alpha$. Want to find $\mathbf{w}$
- Such that the error $|y - \hat{y}|$ is minimised.



Logistic Regression (1/2)

- From real valued output, we want a classifier that gives discrete valued outputs.
- This can be done by thresholding the output of linear regression
- If $\hat{y} > 0$ predict 1, else predict 0.
- This is in fact the discriminant function $f_2(x) = \log \frac{P(h=1|x)}{P(h=0|x)}$.
- So then we have $f_2(\mathbf{x}) = \log \frac{P(h=1|\mathbf{x})}{P(h=0|\mathbf{x})} = \mathbf{w}^T\mathbf{x}$.

Logistic Regression (2/2)

- The logistic regression goes further by finding the probability supporting the prediction.
- This is done by inverting $\log \frac{P(h=1|\mathbf{x})}{P(h=0|\mathbf{x})} = \mathbf{w}^T\mathbf{x}$ to get $P(h=1|\mathbf{x})$
- Which turns out to be
  $P(h=1|\mathbf{x}) = \frac{1}{1+\exp(-\mathbf{w}^T\mathbf{x})}$.
  $P(h=0|\mathbf{x}) = ??$
- The functions is called the sigmoid function.

What's next?: Parameter estimation

- First construct the likelihood function as before, $\prod_i P(\mathbf{x}_i|\mathbf{w}, y_i)^{y_i}(1 - P(\mathbf{x}_i|\mathbf{w}, y_i)^{1-y_i})$
- It's easier to work with log-likelihood
- $\sum_i y_i \log P(\mathbf{x}_i|\mathbf{w}, y_i) + (1 - y_i) \log(1 - P(\mathbf{x}_i|\mathbf{w}, y_i))$
- The maximum of this is where the derivative of the function w.r.t $\mathbf{w}$ equals zero.
- The derivative is .....

Newton's method for numerical optimisation

- The method for finding successive better approx. to the roots of a real-valued function, $w : f(w) = 0$.
- The formula is given by $w_1 = w_0 - \frac{f(w_0)}{f'(w_0)}$
- Back to our problem we want to find $f'(w) = 0$.
- The Newton's method for our purpose is then
- The formula is given by $w_1 = w_0 - \frac{f'(w_0)}{f''(w_0)}$

Summary

- We learn another way to construct a classifier.
- The classifier is called *discriminative* classifier.
- Since it focuses on separating the data not modelling data generation
- One widely used classifier of this type is the Logistic Regression
- Optimising the parameter of the logistic regression can be done using numerical method, such as the Newton's method.