204456: Machine Learning

Ch03 - Regression

Jakramate Bootkrajang

Department of Computer Science
Chiang Mai University

Objective

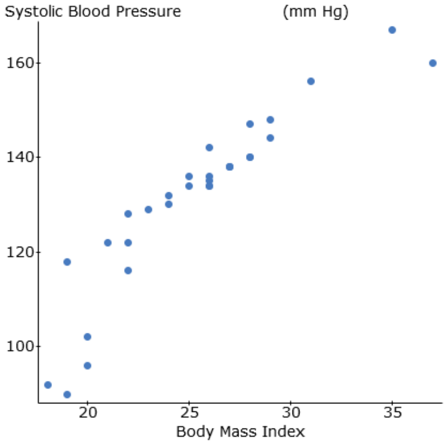
- Understand the basic concepts of regression problem

Outlines

- Motivation
- Regression task and some examples
- Group activity

Motivation

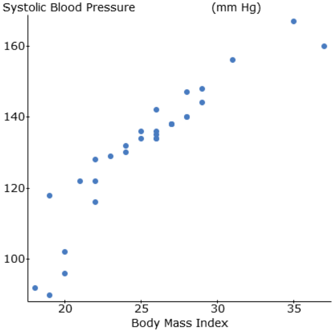
Suppose we have records of patients BMI and blood pressures



Motivation

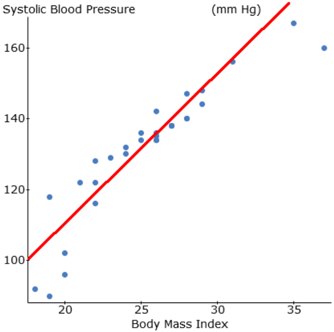
And we would like to estimate relationship between BMI and blood pressure, so that we could

- predict blood pressure if we know BMI



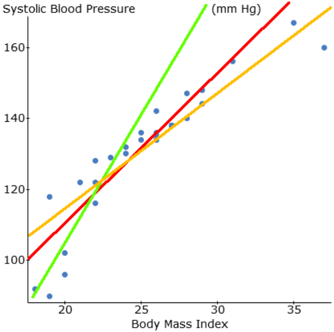
Motivation

Assuming that the relationship between feature and target is linear, we could fit a linear function (a straight line) to the data



Motivation

There are many possible lines, which one is the best fit ?



How to measure goodness of fit ?

Regression task

- Regression analysis is a study of methods for estimating the relationship between independent variable(s)/input feature(s) and dependent variable/target
- A method may assume linear relationship between features and target (this class)
- Or it can assume non-linear relationship (not in this class)

Formally

- Given a dataset $\{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots (\mathbf{x}_n, y_n)\}$
- We want to find a real-valued function $f(\mathbf{x})$ such that the Sum of Squared Residuals (SSR)

$$\sum_{i=1}^{n} (y_i - f(\mathbf{x}_i))^2$$

is as small as possible

Example: BMI and BP data

- x_i is a BMI of one patient (since we have one feature (BMI) our data is in 1-D)
- y_i is his blood pressure
- Data is of the form $\{(20, 102), (23, 119), \ldots, (35, 170)\}$
- Assuming linear relationship we want to find a linear function

$$f(x) = ax + b$$

Example: BMI and BP data

- $f(x)$ is our model of data

$$f(x) = ax + b$$

- a and b are model's parameters that need to be tuned
- so that SSR $= \sum_{i=1}^{n}(y_i - f(x_i))^2$ is minimised
- This is a supervised-learning because we use targets (answers) to help finding the best model's parameters.

Example: Secondhand car pricing

- In many problems, we may have several features. $\mathbf{x} = [x_1, x_2]$
  - x_1 = car's CC, x_2 = mileage
  - y is its second hand price
- The dataset is in the form $\{([1600, 150000], 200000), ([2000, 50000], 450000), \ldots, \}$
- Our linear function is then

$$f(\mathbf{x}) = ax_1 + bx_2 + c$$

where a, b, c are model's parameters that need to be tuned

Linearity check

There are 4+ key assumptions on data that can be used to check if linear regression is appropriate for the data

- Linear Relationship between the features and target
- Little or no Multicollinearity between the features
- Normal distribution of error terms (residual)
- No auto-correlation
- Homoscedasticity
- etc.

Linearity check: assignment

As a group of two teams (5-6 groups expected)

- Do research on how to check for key assumptions for linear relationship and if the assumption is not met, can it be rectified ?
- Present the outcome of your research on the 7th of January (approx 5 mins for each group)
- This assignment worths 3 marks.

Lastly

Q&A ?