WQD7006
Machine Learning for Data Science

Experimenter and Linear Regression

Weka Experimenter

Run Logistic regression on Heart Rate using the default setup. Compare the performance between Naïve Bayes and Logistic Regression.

Linear Regression

Part 1

Last year, five randomly selected students took a math aptitude test before they began their statistics course. Their corresponding marks are as follows:

Student	Aptitude Test Score	Statistics Score
1	95	85
2	85	95
3	80	70
4	70	65
5	60	70

a. Assume that you want to use a linear regression hypothesis of the form, $h_\theta(x) = \theta_0 + \theta_1 x_1$ Whereby x_1 is the Aptitude Test Score" and $h_\theta(x)$ is the hypothesis for the final "Statistics Score". This is to predict the final "Statistics Score" given the "Aptitude Test Score".

Among the following parameter pairs for θ_0, θ_1

Which of the following pair, is the best parameter pair for the linear regression hypothesis in terms of **cost**? Please explain your answer.

- $\theta_0, \theta_1 = 26.77, 0.64$
- $\theta_0, \theta_1 = 52.77, 0.22$

Part 2

Use Pollution file to answer the rest of the questions.

% This is the pollution data
% PREC   Average annual precipitation in inches
% JANT   Average January temperature in degrees F
% JULT   Same for July
% OVR65 % of 1960 SMSA population aged 65 or older
% POPN   Average household size
% EDUC   Median school years completed by those over 22
% HOUS   % of housing units which are sound & with all facilities
% DENS   Population per sq. mile in urbanized areas, 1960
% NONW   % non-white population in urbanized areas, 1960
% WWDRK % employed in white collar occupations
% POOR   % of families with income < $3000
% HC   Relative hydrocarbon pollution potential
% NOX   Same for nitric oxides
% SO@   Same for sulphur dioxide
% HUMID Annual average % relative humidity at 1pm
% MORT   Total age-adjusted mortality rate per 100,000

```
@relation pollution
@attribute PREC real
@attribute JANT real
@attribute JULT real
@attribute OVR65 real
@attribute POPN real
@attribute EDUC real
@attribute HOUS real
@attribute DENS real
@attribute NONW real
@attribute WWDRK real
@attribute POOR real
@attribute HC real
@attribute NOX real
@attribute SO@ real
@attribute HUMID real
@attribute MORT real
@data
36.0,27.0,71.0,8.1,3.34,11.4,81.5,3243.0,8.8,42.6,11.7,21.0,15.0,59.0,59.0,921.870
35.0,23.0,72.0,11.1,3.14,11.0,78.8,4281.0,3.5,50.7,14.4,8.0,10.0,39.0,57.0,997.875
```

1. Describe the sample based on their Education, and Income

Education:
Income:
Relationship:

2. Provide the scatter plot between % of sulphur dioxide (x-axis) and nitric oxide (y-axis).

3. How many attributes are there in the file given?

16

4. Use all the attributes and run linear regression to predict mortality rate (default setting).
   - i. What is the RMSE value?
   - ii. What is the R-square value? (Square of the correlation coefficient)
   - iii. Write the hypothesis based on the regression.
   - iv. How many attributes do you observe? 9
   - v. Identify the missing attributes.

5. Based on your results in (4) – identify the two features with the highest coefficients. Remove these two features and run linear regression again.
   - i. Explain your results in terms of performance.
   - ii. Write the equation.
   - iii. What is the RMSE and R-square?
   - iv. Observe the attributes. What do you notice?
   - v. Explain (iii)