COMP47460/COMP47490 Tutorial

Regression and Gradient Descent

Q1.

In 1929 Edwin Hubble published a paper which revolutionised astronomy. It is the basis of Hubble's law which was the first observational evidence for the expansion of the universe. Hubble was able to measure the recession velocity (km/sec) (how fast it is moving away from us) and the distance from the earth in megaparsecs of various galaxies (1 megaparsec is about 3×10^{22} metres, a long way!)

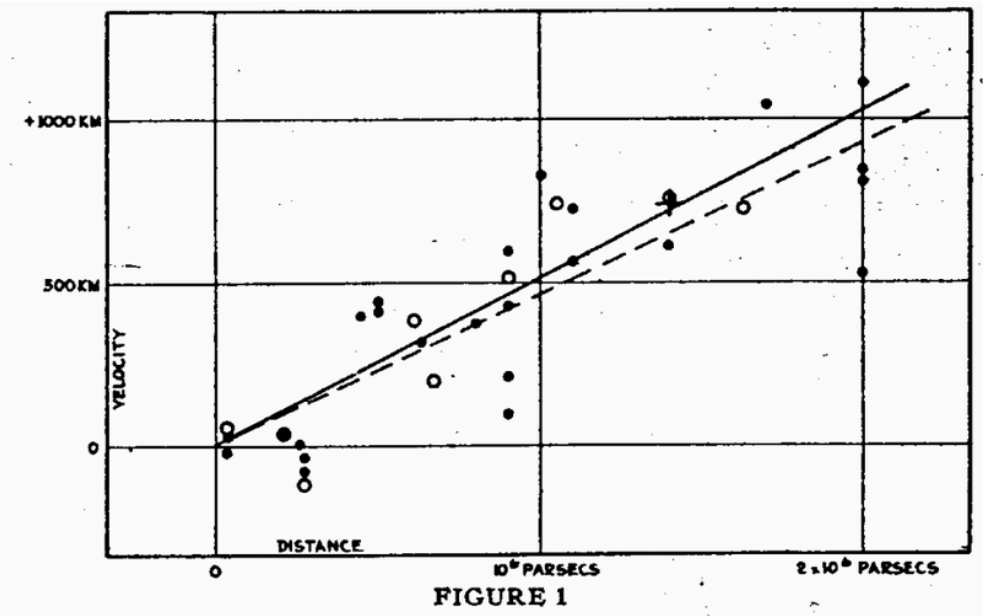

FIGURE 1

a) Just looking at the graph from Hubble's original paper, would you have confidence in his conclusion that there is a linear relationship between the speed of galaxy and its distance from earth?
b) Using OLS find the best fit linear model of the data (β_0, β_1). (do this by hand)
c) Compute the correlation coefficient. (by hand)
d) Using a two-tailed t-test, determine if the relationship between distance and velocity is significant for a p-value of 0.05. We can reject the null hypothesis if $-0.630 < t < 0.630$.

e) The Andromeda Galaxy is the closest spiral galaxy to us at 0.613 megaparsecs. What is it's recession velocity?
f) The slope β_1 is known as the Hubble constant H_0. The latest measurements of the Hubble Space Telescope determine it to be 73.00±1.75 (km/s/megaparsec). How close was Hubble with his original data?
g) Using Weka, run the LinearRegression model and confirm your results.

You can find the "hubble_constant.csv" on moodle.

Q2.

The sinking of the RMS Titanic is one of the most infamous shipwrecks in history. On April 15, 1912, during her maiden voyage, the Titanic sank after colliding with an iceberg, killing 1502 out of 2224 passengers and crew. One of the reasons that the shipwreck led to such loss of life was that there were not enough lifeboats for the passengers and crew. Although there was some element of luck involved in surviving the sinking, some groups of people were more likely to survive than others, such as women, children, and the upper-class.

The data can be found in "titanic.arff" on moodle.

1. Using Weka, build a logistic regression model for (Age, Survived) and (Fare, Survived).
2. Based on the odd-ratio what is the relationship between the variables (Age or Fare and Survived).
3. How do the odds of survival change for each additional year in the age of a passenger?

Q3.

Using Weka, we will build a Gradient Descent regression model for the Hubble dataset. Remember to set the batch size to the size of the dataset, and use the Squared Loss as the lossFunction.

1. Set the learningRate (alpha) to 0.01 and the number of epochs to 1. What values of beta0 and beta0 does the Gradient Descent model give?
2. Increase the number of iterations (epochs)- how many epochs are required for the Gradient Descent algorithm to find the parameters to within 0.01 of the OLS values you found in Q1? (Hint: try

increasing epochs by factor of 10 each time, no need to be too precise).

3. Adjust the alpha parameter (learningRate) to 0.001. How many iterations does it take to find parameters within 0.01 of the OLS parameters?
4. For extra work: apply the formulas on slide 48 by hand and check that your computed values for 1 step of the gradient descent algorithm match with Weka's values, for a learning rate of 0.01.

Q4.

For this question, make sure you have the 'Haberman' Dataset. The arff file should be accessible on moodle. The dataset contains cases from a study that was conducted between 1958 and 1970 at the University of Chicago's Billings Hospital on the survival of patients who had undergone surgery for breast cancer.

Attribute Information:

1. Age of patient at time of operation (numerical)

2. Patient's year of operation (year - 1900, nominal)

3. Number of positive axillary nodes detected (numerical)

4. Survival status (class attribute)
   1 = the patient survived 5 years or longer
   2 = the patient died within 5 year

http://mlr.cs.umass.edu/ml/machine-learning-databases/haberman/haberman.data

Load this dataset in Weka, and perform a logistic regression on the 'Survival Status'

a) What is the best fit logistic regression model to the data?
b) What age is a patient most likely to survive after receiving surgery?
c) What age is a patient least likely to survive after receiving surgery?
d) Comment on how you interpret the odds of surviving if the surgery is performed at these ages?
e) Compare the accuracy of the logistic regression with Naive Bayes and another classifier of your choice?