ECE241 PROJECT 3: First Steps Towards Machine Learning

Due: Dec 9, 2020, 11 PM on Gradescope

Submission Guide:

- Solve the problems below and report your findings in a well formatted PDF file.
- Upload your report to Gradescope.
- Since problem 2 is completely analytical actual code (plus the additional information we ask for) has only to be submitted for problem 1.

Problem 1

Suppose you are the chief data analyst for Walmart. The CEO has tasked you to select a new town for expansion. You are provided with data from different towns which already have a walmart location. This data contains the population and profits (loss if value is negative) of each town.

Example:

7.11	17.59	(population of 7.11 x 10000 and an annual profit of $17.59 x 10000.)
6.88	-2.33	(population of 6.88 x 10000 and an annual loss of $2.33 x 10000.)

Task Overview

a) Plot the given data using matplotlib, profit on y-axis and population on x-axis. Include your graph in your report.

b) Use linear regression to fit a line to the data which will be used for prediction. You can use available machine learning libraries (sklearn, numpy etc.)
sklearn numpy matplotlib

The objective of linear regression is to minimize the cost function :

$$J(\theta) = \frac{1}{2m}\sum_{i=1}^{m}(h_\theta(x^i) - y^{(i)})^2$$

c) Plot the generated line using matplotlib and include the plot in your report.

d) What can you say about the nature of the line?

e) To prepare for a board meeting at Walmart, the CEO has given you a list of cities and their population. Use your machine learning algorithm to rank the cities in descending order of profit potential. In your report, included the selected cities and their estimated revenue in the correct order.

City	Population (x10,000)
A	7.8
B	4.4
C	4.7
D	6.12
E	8.55
F	6.7
G	9.8
H	7.01

f) What is the accuracy of your model?

Problem 2

Include your solution to the problem below in your report file.
Consider regression in one dimension, with a data set $\{ (\mathrm{x}^{\mathrm{i}}, \mathrm{y}^{\mathrm{i}}) \}_{\mathrm{i} = 1,2,\ldots\mathrm{m}}$

a) Find a linear model that minimizes the training error, i.e ω and β to minimize f(x) given that:

$$\mathrm{f(x)} = \sum_{i=1}^{m} (\omega x_i + \beta - y_i)^2$$

b) Variance describes how much a random variable differs from its expected value. It is defined as follows:

$$S^2 = \frac{\sum(x_i - \mu)}{N - 1}$$

Where,

$$S^2 = Sample\ variance$$
$$x_i = Value\ of\ one\ observation$$
$$\mu = \text{Mean of all observations}$$
$$\mathrm{N} = \text{Number of observations}$$

Covariance is a measure of how much two random vaiables vary together. It is similar to variance, but where variance tells you how a single variable varies, covariance tells you how two variables vary together. Covariance is defined as follows:

$$COV_{x,y} = \frac{\sum(x_i - \mu_x)(y_i - \mu_y)}{N - 1}$$

Simplify your answer from (a) in terms of variance and covariance.