Theodore Kim
CS-UY 4563 A – Introduction to Machine Learning
Homework #1

Question 1: *Provide N and d for each data set*

a) Collect a set of data (profit, number of employees, industry and CEO salary) on the top 500 firms to affect CEO salary. **N = 500**, **d = 3**
b) Collect data to determine if a new product is a *success* or *failure*. Collecting data on 20 similar products (success/failure, price, budget, competition price, and ten other variables). **N = 20, d = 13**

Question 2: *Describe a real life applications of the following ML strategies as well as the target and features (explaining inference / prediction)*

(a) Classification:
    1. A spam email classifier in which the target is whether or not an email is labelled as spam and the features are the sender's email address, the subject line, the time it was sent and the email message itself. This is an example of a prediction as one would use a dataset of previously identified spam/normal emails with the feature set described previously to build a model to classify future emails
    2. A program that determines whether an individual is qualified for a loan (*qualified* or *unqualified*, the target) based upon various features such as financial history, income, credit score, age and total savings / assets. This is an example of prediction as the algorithm would take a training set in which a human banker had determined whether or not an individual was eligible for a loan based on the descriptors.
    3. A program to associate the cardiovascular health with gender (the target). The features would be age, ethnicity, resting heartrate, blood pressure and cholesterol. This is an example of inference as the results of the classification program would be compared to actual values to determine the existence of an actual relationship between the data.
(b) Regression:
    1. Determining whether one's max salary is influenced by factors like level of education (i.e. years educated or highest degree), profession, work experience (years), and location (some quantified location, i.e. each state or city gets a number). This is an example of inference as the features are analyzed for whether the outcome fits a linear trend.
    2. A program to determine the ideal investment amount for a given stock / security based upon its average trading value, reported income, industry sector, and number of employees. This is an example of prediction as the function would predict the market value of a stock based on its metric data.
    3. Calculating lifespan (years, this is the target) for smokers based on age, gender, race, other genetic / chronic health conditions, number of years smoking. This is an example of inference as one is trying to determine if there exists a relationship between lifespan and whether or not someone is a smoker.

Question 3: *A university admissions office wants to predict success of students based on application material*

(a) A good target for this supervised learning problem is the alumni's salaries ten years after graduation.
(b) The target variable is continuous
(c) One possible predictor for the target variable is the prospective students' SAT scores
(d) A linear model for the relationship would fit, I expect the slope to be positive as higher SAT scores could reasonably mean a higher salary in the future

Question 4: *Data samples (x_i, y_i)*

x_i	0	1	2	3	4
y_i	0	2	3	8	17

(a) $\bar{x} = 2,\ \bar{y} = 6$

(b)

$$s_{xy} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

$$s_{xy} = ((0-2)(0-6) + (1-2)(2-6) + (2-2)(3-6) + (3-2)(8-6) + (4-2)(17-6))/(5-1)$$

$$s_{xy} = \frac{12 + 4 + 0 + 2 + 22}{4} = \mathbf{10}$$

$$s_y^2 = \frac{\sum_{i=1}^{N}(y_i - \bar{y})^2}{N-1}$$

$$s_y^2 = \frac{(0-6)^2 + (2-6)^2 + (3-6)^2 + (8-6)^2 + (17-6)^2}{4}$$

$$s_y^2 = \frac{36 + 16 + 9 + 4 + 121}{4}$$

$$\boldsymbol{s_y^2 = 46.5}$$

$$s_x^2 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N-1}$$

$$s_x^2 = \frac{(0-2)^2 + (1-2)^2 + (2-2)^2 + (3-2)^2 + (4-2)^2}{4}$$

$$s_x^2 = \frac{4 + 1 + 0 + 1 + 4}{4}$$

$$\boldsymbol{s_x^2 = 2.5}$$

(c) $y = w_0 + w_1 x + \epsilon$

$$w_1 = \frac{S_{xy}}{S_{xx}} = \frac{10}{2.5} = \mathbf{4}$$

$$w_0 = \bar{y} - w_1\bar{x} = 6 - 4(2) = \mathbf{-2}$$

(d) $\hat{y} = w_0 + w_1 x$

$$\hat{y} = (-2) + 4(2.5)$$

$$\hat{y} = 8$$

(e) $E_{in} = MSE = \frac{1}{N}RSS$

$$RSS = \sum_{i=1}^{N}(\hat{y}_i - y_i)^2$$

$$RSS = \big((-2) - 0\big)^2 + (2-2)^2 + (6-3)^2 + (10-8)^2 + (14-17)^2$$

$$RSS = 4 + 0 + 9 + 4 + 9$$

$$RSS = 26$$

$$E_{in} = \frac{42}{5} = \mathbf{8.4}$$

(f) $R^2 = 1 - \frac{RSS}{TSS}$

$TSS = \sum_{i=1}^{N} (y_i - \bar{y})^2$

$TSS = (0-6)^2 + (2-6)^2 + (3-6)^2 + (8-6)^2 + (17-6)^2$

$TSS = 36 + 16 + 9 + 4 + 121$

$TSS = 186$

$R^2 = 1 - \frac{26}{186} = 1 - 0.139784 = \mathbf{0.86021}$

(g) Adjusting the training set by increasing the 'y' value of x=4 would result in a small increase in w_1 as the regression slope would get slightly steeper because the right most point increased, while w_0 would probably not change as the entire dataset is unchanged. Reducing the 'y' value of x=4 would greatly decrease w_1 because the overall regression would level out because the most extreme point decreased in elevation. Again w_0 is unchanged as the overall dataset is unshifted, only a single point.

Question 5: *Perform 2 steps of gradient descent,* $w_0 = 0,\ w_1 = 0,\ \alpha = 0.1$

$temp0 = w_0 - \frac{\alpha}{N}\sum_{i=1}^{N}(w_0 + w_1 x - y_i)$

$temp0 = 0 - \frac{0.1}{5}\sum_{i=1}^{N}(0 + 0 - y_i)$

$temp0 = -0.02\sum_{i=1}^{N}(-y_i)$

$temp0 = -0.02(0 - 2 - 3 - 8 - 17)$

$temp0 = 0.6$

$temp1 = w_1 - \frac{\alpha}{N}\sum_{i=1}^{N}(w_0 + w_1 x - y_i)x_i$

$temp1 = 0 - \frac{0.1}{5}\sum_{i=1}^{N}(0 + 0 - y_i)x_i$

$temp1 = -0.02\sum_{i=1}^{N}(y_i \cdot x_i)$

$temp1 = -0.02(0 - 2 - 6 - 24 - 68)$

$temp1 = -0.02 \cdot -100 = 2$

After 1 round of gradient descent the new values are: $\boldsymbol{w_0 = 0.6, w_1 = 2}$

$temp0 = w_0 - \frac{\alpha}{N}\sum_{i=1}^{N}(w_0 + w_1 x - y_i)$

$temp0 = -0.6 - \frac{0.1}{5}\sum_{i=1}^{N}(0.6 + 2x_i - y_i)$

$temp0 = -0.6 - 0.02((0.6 - 0) + (2.6 - 2) + (4.6 - 3) + (6.6 - 8) + (8.6 - 17))$

$temp0 = 0.6 - 0.02(0.6 + 0.6 + 1.6 - 1.4 - 8.4)$

$temp0 = 0.74$

$temp1 = w_1 - \frac{\alpha}{N}\sum_{i=1}^{N}(w_0 + w_1 x - y_i)x_i$

$temp1 = 2 - \frac{0.1}{5}\sum_{i=1}^{N}(-0.6 - 2x_i - y_i)x_i$

$$temp1 = 2 - 0.02\big((0.6)0 + (0.6)1 + (1.6)2 - (1.4)3 - (8.4)4\big)$$

$$temp1 = 2 - 0.02(0 + 0.6 + 3.2 - 2.8 - 16.8)$$

$$temp1 = 2 - 0.02 \cdot -15.8 = 2.316$$

Question 6: *For data in question 4, which is the most likely choice of parameters given noise is Gaussian, with mean of 0 and variance of 5.2*

(a) $w_0 = -1, w_1 = 4$
(b) $w_0 = -2, w_1 = 4$
(c) $w_0 = -2, w_1 = 3$

(b) is the most likely choice of parameters

Question 7: *Given the regression model:*

$$z(t) \approx z_0 e^{-\alpha t}$$

Model (1) is nonlinear, so the linear regression cannot be directly applied. Trying to find parameters z_0 and α.

(a) Rewrite the regression using logarithms to fit it in the linear regression format

$$z(t) \approx \ln\left(z_0 e^{-\alpha t}\right)$$

$$\boldsymbol{z(t) \approx \ln(z_0) - \alpha t}$$

(b) Least squares solution for the best estimate of parameters

$w_1 = \alpha, w_0 = \ln(z_0)$

$$\boldsymbol{\alpha = \frac{\sum_{i=1}^{N}(t_i - \bar{t})(z(t_i) - \overline{z(t)})}{\sum_{i=1}^{N}(t_i - \bar{t})^2}}$$

$$\ln(z_0) = \overline{z(t)} - \alpha\bar{t}$$

$$\boldsymbol{z_0 = e^{\overline{z(t)} - \alpha\bar{t}}}$$

(c) Python code

```
import numpy as np
from math import e
# ...
t = #input vector (x)
z = #training vector (y)

tbar = np.mean(t)
zbar = np.mean(z)
```

```
a = np.sum((t - tbar) * (z - zbar))/np.sum((t - tbar) ** 2)
z0 = e ** (zbar – a * tbar)
```

Question 8: A regression in which the intercept (w_0) is forced to 0:

$$y \approx wx$$

(a) Find the cost function in relation to the RSS of w

$$J(w) = \frac{1}{2N} RSS$$

$$RSS = \sum_{i=1}^{N} (wx_i - y_i)^2$$

$$J(w) = \frac{1}{2N} \sum_{i=1}^{N} (wx_i - y_i)^2$$

(b) Find the w that minimizes RSS

$$\frac{\delta J(w)}{\delta w} = \frac{1}{N} \sum_{i=1}^{N} (wx_i - y_i)x_i$$

$$temp = w - \frac{\alpha}{N} \sum_{i=1}^{N} (wx_i - y_i)x_i$$