ECE 449. Assignment 1. Conceptual Question

Q1

a). supervised learning.

use the book this costomer bought before as training data.
with features like book type, language, number of words, price and so on.

b). reinforcement learning.

use the decision make every gone as training data, like a sequence of location where to move, also some feature to judge the decision efficiency like whether win or lose, number of step used to win and so on.

c) both supervised or unsupervised.

it is a clustering problem essentially. with unsupervised learning. we can effectly get some different group of movies with traning data as some basic features of movies like length, nation, language, director, title and so on, and actually we don't know whether any one movie is what type. as we have no label.

with supervised learning, we need training data above as well as all the type labels each movie belong to.

d) unsupervised learning.

use basic features like style, rhythm, language and so on as training data

e) both supervised learning

use data of previous costomers and their maxium debt.

Q2

the key difference is : linear regression is a regression problem whose output is continous while logistic regression, although named in regression, is actually a classification problem, which has discrete output.

liear regression example : Stock trend forecast

logistic regression example: Spam mail detection.

Q3 :

(1,2), (2,1), (3,2). $\hat{y} = w_0 + w_1 x = H\hat{w}$

with $\hat{w} = \begin{bmatrix} w_0 \\ w_1 \end{bmatrix}$, $y = \begin{bmatrix} 2 \\ 1 \\ 2 \end{bmatrix}$, $H = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix}$

so. $H^T y = H^T H w$, $H^T H$ must be invertible.

$\Rightarrow (H^T H)^{-1} H^T y = (H^T H)^{-1}(H^T H) w = w$

$w = \left(\begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix}^T \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix} \right)^{-1} \cdot \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix}^T \cdot \begin{bmatrix} 2 \\ 1 \\ 2 \end{bmatrix} = \begin{pmatrix} \frac{5}{3} \\ 0 \end{pmatrix}$

so $w_0 = \frac{5}{3}$, $w_1 = 0$.

Q4 : (-2,1), (-1,0) (1,0), (2,2) , $y = w_0 + w_1 x_1 + w_2 x_2^2$

$y = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 2 \end{bmatrix}$ $w = \begin{bmatrix} w_0 \\ w_1 \\ w_2 \end{bmatrix}$ $H = \begin{bmatrix} 1 & -2 & 4 \\ 1 & -1 & 1 \\ 1 & 1 & 1 \\ 1 & 2 & 4 \end{bmatrix}$

Q5 : For Gradient descent, assume we fit our model with $h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \cdots$, $\theta = [\theta_0, \theta_1, \cdots]^T$, then we analyze the loss by $J(\theta) = \min_\theta \frac{1}{2N} \sum_{i=1}^{N} (h_\theta(x^{(i)}) - y^{(i)})^2$

start by some random guess of θ, we update θ_j $(j=0,1,2\cdots)$ with $\theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1)$, with α some learning rate defined until θ_j convergence or through some iteration, one important thing is all θ_j should update simultaneously.

The different is gradient ascent algorithm are used to get max value, with the same idea of use gradient, out descent algorithm update θ by substract but ascent use addition, which cause their different use.

Q6:

large learning rate:

Advantage: we may quickly converge that can shorter operation time

Disadvantage: as it jump large step each time, we may fail to converge or even diverge

Small learning rate:

Advantage: we can alway find converge point

Disadvantage: we may get into some local min that are not good enough and take long operation time

Q7:

$$H(X) = -\sum_{i=1}^{4} P_i \log_2 P_i$$

$$= -4 \times \frac{1}{4} \log_2 \frac{1}{4}$$

$$= 2.$$

Q8:

define X as different records of scores.

Y as different postgraduate admission

X margin $(\frac{7}{16}, \frac{1}{4}, \frac{5}{32}, \frac{5}{32})$

Y margin $(\frac{1}{4}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8})$

$H(Y|X) = \sum_{x \in X, y \in Y} p(x,y) \log \frac{p(x)}{p(x,y)} = \frac{3\times1}{8}\log\frac{\frac{7}{16}}{\frac{1}{8}} + \frac{1}{16}\log\frac{\frac{7}{16}}{\frac{1}{16}} + \frac{2\times1}{16}\log\frac{\frac{1}{4}}{\frac{1}{16}} + \frac{1}{8}\log\frac{\frac{1}{4}}{\frac{1}{8}} + \frac{2\times1}{32}\log\frac{\frac{5}{32}}{\frac{1}{32}} + \frac{4\times1}{16}\log\frac{\frac{5}{32}}{\frac{1}{11}}$

$= \frac{3}{8}\log\frac{7}{2} + \frac{1}{16}\log 7 + \frac{1}{4} + \frac{1}{4} + \frac{1}{16}\log 5 + \frac{1}{4}\log\frac{5}{2}$

$= 1.704$

$$H(X|Y) = \sum_{x\in X, y\in Y} p(x,y)\log\frac{p(x)}{p(x,y)} = \frac{1}{8}\log\frac{\frac{1}{4}}{\frac{1}{8}} + \frac{1}{16}\log\frac{\frac{1}{4}}{\frac{1}{16}} + \frac{3}{16}\log\frac{\frac{1}{4}}{\frac{1}{16}} + \frac{1}{16}\log\frac{\frac{1}{4}}{\frac{1}{16}} + \frac{1}{8}\log\frac{\frac{3}{8}}{\frac{1}{16}} + \frac{1}{8}\log\frac{\frac{1}{4}}{\frac{3}{8}}$$

$$= \frac{1}{8} + \frac{1}{8} + \frac{3}{16} + \frac{1}{2} + \frac{1}{4}\log 3 + \frac{1}{8}\log 6 + \frac{1}{8}$$

$$= 1.657$$

$$I(X,Y) = H(X) - H(X|Y)$$

$$= -\frac{7}{16}\log\frac{7}{16} - \frac{1}{4}\log\frac{1}{4} - \frac{5}{16}\log\frac{5}{16} - \frac{5}{32}\log\frac{5}{32} - H(X|Y)$$

$$= 0.202$$

Q9.

Follow this hint. first prove $h_i = x_i (X^T X)^{-1} x_i^T$

$$x_i (X^T X)^{-1} x_i^T = (1, x_i)\left[\begin{pmatrix} 1 & \cdots & 1 \\ x_0 & \cdots & x_n \end{pmatrix}\begin{pmatrix} 1 & x_0 \\ \vdots & \vdots \\ 1 & x_0 \end{pmatrix}\right]^{-1}\begin{pmatrix} 1 \\ x_i \end{pmatrix}$$

$$= (1, x_i)\begin{bmatrix} N & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix}^{-1}\begin{pmatrix} 1 \\ x_i \end{pmatrix}$$

$$= (1, x_i)\begin{pmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & N \end{pmatrix}\begin{pmatrix} 1 \\ x_i \end{pmatrix} \Big/ (N\cdot\sum x_i^2 - (\sum x_i)^2)$$

$$= \frac{\sum x_i^2 - 2x_i\sum x_i + N x_i^2}{N\sum x_i^2 - (\sum x_i)^2}$$

$$= \frac{1}{N} + \frac{N\bar{x}^2 - 2N\bar{x}x_i + N x_i^2}{N\sum x_i^2 - N^2\bar{x}^2}$$

$$= \frac{1}{N} + \frac{(\bar{x} - x_i)^2}{\sum x_i^2 - N\bar{x}^2} = \frac{1}{N} + \frac{(\bar{x} - x_i)^2}{\sum (x_i^2 - \bar{x}^2)}$$

$$= \frac{1}{N} + \frac{(\bar{x} - x_i)^2}{\sum (x_i^2 + \bar{x}^2) - \sum 2x_i\bar{x}} = \frac{1}{N} + \frac{(\bar{x} - x_i)^2}{\sum (x_i - \bar{x})^2}$$

Then for $CV = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{y_i - \hat{y}_i}{1-h_i}\right)^2$, say x_i, y_i be X, Y without i^{th} row

$$X_i^T X_i = (X^T X - x_i x_i^T), \text{ and with}$$

$$(X_i^T X_i)^{-1} = (X^T X)^{-1} + \frac{(X^T X)^{-1} x_i x_i^T (X^T X)^{-1}}{1 - h_i}$$

$$\hat{\beta}_{(i)} = (X_{(i)}^T X_{(i)})^{-1} X_{(i)}^T Y_{(i)}$$

$$= \left((X^T X)^{-1} + \frac{(X^T X)^{-1} x_i x_i^T (X_{(i)}^T X_{(i)})^{-1}}{1 - h_i} \right) (X^T Y - x_i y_i)$$

$$= \hat{\beta} - \left(\frac{(X^T X)^{-1} x_i}{1 - h_i} \right) \left(y_i (1 - h_i) - x_i^T \hat{\beta} + h_i y_i \right)$$

$$= \hat{\beta} - \frac{(X^T X)^{-1} x_i \, (y_i - \hat{y}_i)}{1 - h_i}$$

given $(y_i - \hat{y}_{(i)})^2 = (y_i - x_i^T \hat{\beta}_{(i)})^2$

$$= \left(y_i - x_i^T \left(\hat{\beta} - \frac{(X^T X)^{-1} x_i (y_i - \hat{y}_i)}{1 - h_i} \right) \right)^2$$

$$= \left(y_i - \hat{y}_i + \frac{h_i (y_i - \hat{y}_i)}{1 - h_i} \right)^2$$

$$= \left(\frac{y_i - \hat{y}_i}{1 - h_i} \right)^2 = \text{PRESS}$$