Name: Alisher Karim ID: 20192014

Problem 1

① supervised learning is the process of learning where the label input and target is given while in unsupervised it is not.
Unsupervised Learning is descriptive analysis of data

② Classification is mainly about predicting the class from the input data. Regression is mainly from continuous response variable, classification is from categorical response variable.

③ Pearson correlation gives the measure of linear relationship. Euclidean distance gives the length of how far away the points are. Pearson correlation is unit independent.

④ In gradient descent, the function is maximized from getting the derivative.
In Newton-Raphson the function is maximized from second derivation and used in Logistic Regression.

⑤ Hard Clustering predicts whether the data in a cluster or not. Insoft clustering the probability or likelyhood of that is given.

⑥ Single linkage is mainly about minimum distances and nearest neighbors. Complet linkage is about maximum distance and furthest neighbors.

⑦ Support vectors are used when classifying data where margin is maximized

Problem 2

$y_1, y_2, \dots y_n$ - dependent variables

$x_{i1}, x_{i2}, \dots x_{im}$ - independent variables where $i = 1, 2, \dots n$

ϵ_i - errors

$\beta_0, \beta_1, \dots \beta_m$ - coefficients.

$$y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_m x_{im} + \epsilon_i \quad (i = 1, 2, \dots n)$$

$\epsilon_i \sim N(0, \sigma^2)$

Question 2-1

$$Q = \sum_{i=1}^{n} [y_i - (\beta_0 + \dots + \beta_m x_{im})]^2 = SSE$$

Statistical Inference: (on β)

$$E(\hat{\beta}) = \begin{pmatrix} E(\hat{\beta}_0) \\ \vdots \\ E(\hat{\beta}_m) \end{pmatrix} = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_m \end{pmatrix}$$ where $\hat{\beta}_i$ is the root of $\frac{\partial Q}{\partial \beta_i} = 0$

$$Var(Y) = \begin{pmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \ddots & & \vdots \\ \vdots & & \ddots & \\ 0 & \dots & & \sigma^2 \end{pmatrix} = \sigma^2 I_n$$

$$\hat{\beta} = (X^T X)^{-1} X^T Y = CY \Rightarrow Var(\hat{\beta}) = \sigma^2 (X^T X)^{-1}$$

Problem 3.

(1) PCA is a linear dimension-reduction techinque that maximizes the variance in the data. by using linear combination.

(2) There should be a linear relationship between variables.
There shouldn't be no unique variance.
Variables must be strongly correlated.

Problem 4.

$$\text{Log}\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = X\beta = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p \quad (1)$$

↑
Logit

$$P(Y=1) = \frac{e^{\beta_0+\beta_1 X_1+\ldots+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\ldots+\beta_p X_p}} \quad (2)$$

Here, by taking power to e, in equation (1), we get

$$\frac{P(Y=1)}{1-P(Y=1)} = e^{\beta_0+\ldots+\beta_p X_p} \Leftrightarrow \text{equation (2)}$$

Problem 5.

for Regression, it uses variance reduction, discretization, binary decision

for Classification, it uses GINI index, GaiN, and [chi-square].

MSE is also used for regression.

Entropy is used for regression.

Problem 6.

$$AUC = \frac{(\text{Sum of all positive}) - |\text{all positive}| \cdot (|\text{all positive}| + 1)/2}{|\text{all positive}| + |\text{all negative}|}$$

Problem 8.8

Collaborative filtering:

$$\begin{bmatrix} 5 & 3 & 4 & 4 \\ 3 & 1 & 2 & 3 \\ 3 & 3 & 4 & 3 \end{bmatrix} \begin{bmatrix} 2 \\ 3 \\ 4 \end{bmatrix}$$

given users data (each row) and some features (each column). We have to find which feature will be in the new user.

Content-Based:

User1	5,25	3,11	0
User2	1,21	6,1	1
User3	8,59	2,54	1,54

This is almost same as collaborative, Encodes text documents into multi-dimesional Euclidean space.

Problem 9.

① (1-1)

②

③ True. Both of them use derivatives, but slight difference may occur as Newton-Raphson uses second derivative.

④. True. The initial values are used to get improved and more accurate version of them.

⑤

⑥

⑦

⑧ True. If the samples are finite, there may be not enough information

⑨ True. They don't use no assumption on underlying data.

⑩ True. creating many of decision trees, averaging the variance can reduce.