Hw 2 LiChen Liang

1. a)

$$\begin{array}{ccc} X & Y & Z \end{array}$$
$$\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix}$$

$E(X) = (1+4)/2 = 2.5$

$E(Y) = (2+5)/2 = 3.5$

$E(Z) = (3+6)/2 = 4.5$

$Cov(X,Y) = Cov(Y,X) = E(XY) - E(X)E(Y)$

$= \left(\frac{1\cdot 2 + 4\cdot 5}{2}\right) - 2.5(3.5)$

$= 2.25$

$Cov(X,Z) = Cov(Z,X) = E(XZ) - E(X)E(Z) = 2.25$

$Cov(Y,Z) = Cov(Z,Y) = E(YZ) - E(Y)E(Z) = 2.25$

$Cov(X,X) = E(X^2) - E(X)^2$

$= \left(\frac{1^2+4^2}{2}\right) - (2.5)^2 = 2.25$

$Cov(Y,Y) = E(Y^2) - E(Y)^2 = 2.25$

$Cov(Z,Z) = E(Z^2) - E(Z)^2 = 2.25$

$$\begin{array}{c|ccc} & X & Y & Z \\ \hline X & & & \\ Y & & & \\ Z & & & \end{array} \quad \begin{bmatrix} 2.25 & 2.25 & 2.25 \\ 2.25 & 2.25 & 2.25 \\ 2.25 & 2.25 & 2.25 \end{bmatrix}$$

Covariance matrix

b) A zero eigenvalue means there is no variance between that eigenvector and the data.

c) If two eigenvalues are the same, then the variance between the corresponding eigenvectors to the data are the same. In this case we have to decide which eigenvalue to choose for the PCA.

2.

a) Yes, greedy algorithm will essentially reach the purest splitting. i.e. There are N students with N IDs, then there can be N branches achieving purest split.

b) Assume N values with equal possibility 1/N then the entropy is

$$\underbrace{-\tfrac{1}{N}\log_2\tfrac{1}{N} - \tfrac{1}{N}\log_2\tfrac{1}{N} \dots - \tfrac{1}{N}\log_2\tfrac{1}{N}}_{N} \Rightarrow N\cdot\left(-\tfrac{1}{N}\log_2\tfrac{1}{N}\right) = \log_2 N$$

c) Information gain is used to reduce the entropy by splitting into k partitions

$Gain_{split}$ = old entropy − new entropy $= Entropy(p) - \left(\sum_{i=1}^{k} \frac{n_i}{N} Entropy(i)\right)$

Assume there are N values and we split into $k=N$ partitions and each partition has only 1 value, then $Gain = Entropy(p) - 0$ will be the maximum Gain that can be achieved.

Since SplitInfo is the weighted total of $-\log {n_i}/{N}$. When we split into $k=N$ partitions with $n_i = 1$, then $-\log {n_i}/{N}$ goes to ∞ and Gain ratio $= \frac{Gain_{split}}{\infty} = 0$

d) Pre-Pruning: Using certain theshold or condition to stop in order to prevent the tree from fully grown

Post-Pruning: Let decision tree to fully grow then trim from bottom up.

3. a) $precision = \frac{5}{5+20} = 0.2$ $recall = \frac{5}{5+15} = 0.25$

$F\text{-measure} = \frac{2(5)}{2(5)+15+20} = 0.22$

b) ROC curve shows the performance tradeoff between True Positive and False Positive. so (0,0) means we make everything negative so it's not detecting any positives and (1,1) means we make everything positive and it's detecting all positives. The diagonal line on the ROC curve indicate random classifier.

4. a) Try each k and record their accuracy. Pick the k with highest accuracy.
If k is too small, it may contain too little of points which might be noise.
If k is too large, it may contain points that belong to other classes

b) Since we have to calculate the distance with each point in the training set, then it will take $O(n)$ time.

5.

a) $P(\text{Disease}) = 0.11$ $P(\text{no disease}) = 0.89$

$P(\text{positive} \mid \text{disease}) = 0.9$ $P(\text{Positive} \mid \text{no disease}) = 0.02$

$$P(\text{Disease} \mid \text{positive}) = \frac{P(\text{Positive} \mid \text{Disease})\, P(\text{Disease})}{P(\text{positive})}$$

$$P(\text{positive} \mid \text{Disease}) P(\text{Disease}) + P(\text{Positive} \mid \text{No disease}) P(\text{No disease})$$

$$= \frac{0.9 \times 0.11}{0.9 \times 0.11 + 0.02 \times 0.89} = 0.848$$

b) Discretize, Two way Split, Probability density estimation

c) Given the weather condition of a part of sea, such as wind, wave, rain, etc each day, decice whether should permit a ship to leave the dock.

d) Naive bayes is simple and efficient for most of the classification tasks. It's robust to noise and irrelvant points. It also handle missing value pretty well.

Duplicate data will make the trained model bias toward duplicates.

6. Accuracy $\approx 80\%$