Question 1

1. Briefly explain the difference between supervised and unsupervised learning methods.

(1) Input Data:

The input data used in supervised learning is well known and is labeled. However, the data used in unsupervised learning is not known nor labeled.

(2) Computational Complexity:

One has to understand and label the inputs in supervised learning, while in unsupervised learning, one is not required to understand and label the inputs.

(3) Accuracy of Results:

The supervised method is more accurate than unsupervised method, because the input data is well known and labeled which means that the machine will only analyze the hidden patterns.

(4) Number of Classes:

All the classes used in supervised learning are known which means that the answers in the analysis are likely to be known. But there is no prior knowledge in unsupervised method of machine learning and the number of classes are not known which clearly means that no information is known.

(5) Real Time Learning

The supervised learning takes place off-line while unsupervised learning takes place in real time.

2. Briefly explain the terms “overfitting” and “generalization”

Overfitting is "the production of an analysis that corresponds too closely or exactly to a particular set of data, and may therefore fail to fit additional data or predict future observations reliably".

Generalization is the formulation of general concepts from specific instances by abstracting common properties.

3. Categorize the following methods as “parametric” or “non-parametric”: (i) Maximum likelihood, (ii) Linear discriminant analysis, (iii) k-nearest neighbour method.

Maximum likelihood: parametric

Linear discriminant analysis: parametric

k-nearest neighbour method: non-parametric

4. Explain the difference between the maximum likelihood and Bayesian estimation methods.

Maximum likelihood estimation assumed that the parameter of the distribution, $\boldsymbol{\theta}$, is fixed while Bayesian estimation assumes that $\boldsymbol{\theta}$ is a random variable.

5. Explain the essence of Bayesian classification

The essence of Bayesian inference is in the rule, known as Bayes' theorem, that tells us how to update A priori probability , in order to find out Posteriori probability.

6. Explain the Naive Bayes classification

Naive Bayes is a simple technique for constructing classifiers: models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set.

7. Explain the idea of a linearly separable data when considering support vector machines.

The linear separability is a property of a pair of sets of points. This is most easily visualized in two dimensions by thinking of one set of points as being colored blue and the other set of points as being colored red. These two sets are linearly separable if there exists at least one line in the plane with all of the blue points on one side of the line and all the red points on the other side. This idea immediately generalizes to higher-dimensional Euclidean spaces if line is replaced by hyperplane.

8. Explain three methods that can be used to extend two-class classification methods to multi-class versions.

(1) One-against-all:

One-against-all strategy involves training a single classifier per class, with the samples of that class as positive samples and all other samples as negatives. This strategy requires the base classifiers to produce a real-valued confidence score for its decision, rather than just a class label.

(2) One-against-one:

In the One-against-one, one trains K (K − 1) / 2 binary classifiers for a K-way multiclass problem; each receives the samples of a pair of classes from the original training set, and must learn to distinguish these two classes. At prediction time, a voting scheme is applied: all K (K − 1) / 2 classifiers are applied to an unseen sample and the class that got the highest number of "+1" predictions gets predicted by the combined classifier.

(3) Discriminant functions:

Define C linear discriminant functions $g_i(x), i = 1,\ldots\ldots,C$, Assign pattern vector to class w_i if $g_i(x) = max_j g_j(x)$.

9. What is the significance of support vectors in a support vector classification method?

A Support Vector Machine (SVM) performs classification by finding the hyperplane that maximizes the distance margin between the two classes. The extreme points in the data sets that define the hyperplane are the support vectors.

10. Describe two ways of achieving dimensionality reduction.

(1) Principal Component Analysis

The main linear technique for dimensionality reduction, principal component analysis, performs a linear mapping of the data to a lower-dimensional space in such a way that the variance of the data in the low-dimensional representation is maximized.

(2) Kernel PCA

Principal component analysis can be employed in a nonlinear way by means of the kernel trick. The resulting technique is capable of constructing nonlinear mappings that maximize the variance in the data.

Question 2-1

1. $Tr(A) = 2 + 1 = 3$

2. $\det(A) = 2 * 1 - 2 * 3 =- 4$

3. $\lambda_1 =- 1, \lambda_2 = 4$

4. $p_1 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}, p_2 = \begin{bmatrix} 1.5 \\ 1 \end{bmatrix}$

Question 2-2

1. Suppose that this customer's account is overdrawn in the first month. How does this alter the bank's opinion of this customer's creditworthiness?

$P(good) = P(A_1) = 0.7, P(bad) = P(A_2) = 0.3, P(over) = P(B_1), P(notover) = P(B_2)$

$P(over|good) = P(B_1|A_1) = 0.01, P(over|bad) = P(B_1|A_2) = 0.1$

$P(notover|good) = P(B_2|A_1) = 0.99, P(notover|bad) = P(B_2|A_2) = 0.9$

$$P(good|over) = \frac{P(B_1|A_1)P(A_1)}{P(B_1|A_1)P(A_1) + P(B_1|A_2)P(A_2)} = \frac{7}{37} \approx 0.189$$

$$P(bad|over) = \frac{P(B_1|A_2)P(A_2)}{P(B_1|A_1)P(A_1) + P(B_1|A_2)P(A_2)} = \frac{30}{37} \approx 0.811$$

2. Given the result in part (1), what would be the bank's opinion of the customer's creditworthiness at the end of the second month if there was not an overdraft in the second month?

$P(A_1) = \frac{7}{37} \approx 0.189, P(A_2) = \frac{30}{37} \approx 0.811$

$$P(good|notover) = \frac{P(B_2|A_1)P(A_1)}{P(B_2|A_1)P(A_1) + P(B_2|A_2)P(A_2)} = \frac{77}{377} \approx 0.204$$

$$P(bad|notover) = \frac{P(B_2|A_2)P(A_2)}{P(B_2|A_1)P(A_1) + P(B_2|A_2)P(A_2)} = \frac{300}{377} \approx 0.796$$

Question 3

1. Explain the significance of area under the receiver operating characteristics (ROC) curve as it pertains to classifier performance measure.

AUC (area under the curve) represents the probability that a randomly chosen negative example will have a smaller estimated probability of belonging to the positive class than a randomly chosen positive example.

2. $\widetilde{\mathrm{A}}_X = \frac{1}{N_1 N_2}\left\{S_0 - \frac{1}{2}N_1(N_1+1)\right\} = \frac{1}{5*5}\{39 - 15\} = \frac{24}{25}$, $e_x = \frac{1}{5}$

$\widetilde{\mathrm{A}}_Y = \frac{1}{5*5}\{31 - 15\} = \frac{16}{25}$, $e_y = \frac{1}{5}$

3. $Z = \frac{|n_{01} - n_{10}| - 1}{\sqrt{n_{01} + n_{10}}} = \frac{|2 - 2| - 1}{\sqrt{2+2}} = -\frac{1}{2}$