Foundations of Machine Learning

Duration: 2 hours **Full Marks: 75**

Q1. K Nearest Neighbours [10+5 = 15 Marks]

1. The table below lists a dataset that was used to create a nearest neighbour model that predicts whether it will be a good day to go surfing.

ID	WAVE SIZE (FT)	WAVE PERIOD (SECS)	WIND SPEED (MPH)	GOOD SURF
1	6	15	5	yes
2	1	6	9	no
3	7	10	4	yes
4	7	12	3	yes
5	2	2	10	no
6	10	2	20	no

Assuming that the model uses Euclidean distance to find the nearest neighbour, what prediction will the model return for each of the following query instances.

ID	WAVE SIZE (FT)	WAVE PERIOD (SECS)	WIND SPEED (MPH)	GOOD SURF
Q1	8	15	2	?
Q2	8	2	18	?
Q3	6	11	4	?

Repeat the same for K Nearest neighbor model with K=3.

Q2. Naive Bayes Classifier [3 + 7 + 5 = 15 Marks]

Consider a NB model for text classification with vocabulary V = "secret", "offer", "low", "price", "valued", "customer", "today", "dollar", "million", "sports", "is", "for", "play", "healthy", "pizza". Consider a binary valued bag of words representation of the text, where each dimension represents if a word is present or not.

We have the following spam messages:
"million dollar offer", "secret offer today", "secret is secret".
Normal/non-spam/ham messages:
"low price for valued customer", "play secret sports today", "sports is healthy", "low price pizza".

(i) Represent the data in the described bag of words format. Provide input and output for each data point.

(ii) Which NB classifier (in terms of the distribution of likelihood) will you use ? Explain the parameters of the NB classifier and obtain the maximum likelihood (ML) estimates of the parameters of the NB classifier considering the training data as defined above.

(iii) We have a new message "million dollar price". Assuming ML estimate of the parameters, what is the probability of classifying this message to the spam/ham classes ?

Q3. Decision Trees [3 + 5 + 7 = 15 Marks]

Consider the train data in the table below. One would like to determine if somebody finds a particular type of food appealing based on the food's temperature, taste, and size. You are asked to build a decision tree to learn from this data and predict if a food is appealing.

Appealing	Temperature	Taste	Size
No	Hot	Salty	Small
No	Cold	Sweet	Large
No	Cold	Sweet	Large
Yes	Cold	Sour	Small
Yes	H	Sour	Small
No	H	Salty	Large
Yes	H	Sour	Large
Yes	Cold	Sweet	Small
Yes	Cold	Sweet	Small
No	H	Salty	Large

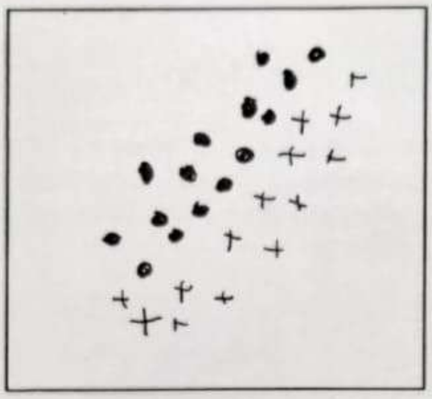
(i) What is the initial entropy of Appealing?

(ii) Assume that Taste is chosen for the root of the decision tree. What is the information gain associated with this attribute?

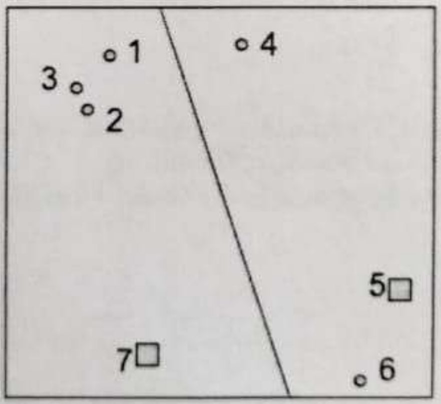
(iii) Draw the full decision tree learned for this data (without any pruning).

Q4. Unsupervised Learning [4+4+7= 15 Marks]

(i) Consider the data points from two classes distributed as in the figure below. Consider the unsupervised dimensionality reduction technique - principal component analysis (PCA) and supervised linear discriminant analysis (LDA). Plot separately the directions of the first PCA and LDA components for the following data



(ii) Perform K-means on the dataset given below. Circles are data points and there are two initial cluster centers, at data points 5 and 7. Draw the cluster centers (as squares) and the decision boundaries that define each cluster. If no points belong to a particular cluster, assume its center does not change. Use as many of the pictures as you need for convergence.



(iii) Consider clustering 1D data with a mixture of 2 Gaussians using the EM algorithm. You are given the 1-D data points x=[1 10 20]. Suppose the output of the E step is the following matrix:

$$\mathbf{R} = \begin{bmatrix} 1 & 0 \\ 0.4 & 0.6 \\ 0 & 1 \end{bmatrix}$$

where entry ri,c is the probability of observation xi belonging to cluster c (the responsibility of cluster c for data point i). Consider the maximization step in the EM algorithm for this particular problem setup

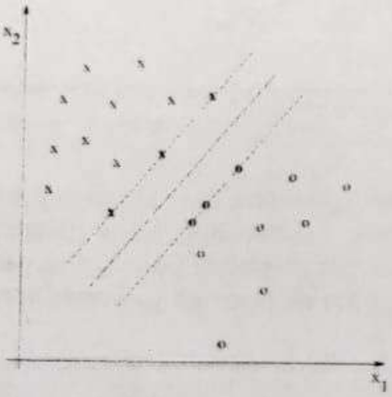
a. Write down the likelihood function you are trying to optimize.

b. After performing the M step for the mixing weights π1, π2, what are the new values?

c. After performing the M step for the means μ1 and μ2, what are the new values?

Q5. Support Vector Machines [3 + 8 + 4 = 15 Marks]

(a) Consider a linear SVM learnt on the full data. Assume you perform a leave one out cross validation on the data given in the following picture. In each iteration you fit a linear SVM on the training data and predict on the validation data point. What do you expect the overall validation error is going to be ? Explain your answer.



(b) Consider the soft margin SVM formulation as shown below. Derive the dual formulation for the soft margin SVM. How is the dual formulation of soft margin SVM different from the hard margin SVM. From the derivation, reason how SVM dual formulation leads to a sparse solution.

$$\begin{aligned} \min_{w,b,\xi} \quad & \frac{1}{2}||w||^2 + C\sum_n \xi_n \\ \text{subj. to} \quad & y_n\left(w \cdot x_n + b\right) \geq 1 - \xi_n \\ & \xi_n \geq 0 \end{aligned}$$

(c) Demonstrate through an example, how kernels can help to improve the dot product computation in a higher dimensional space. Provide the dual optimization problem associated with the nonlinear SVMs and explain how kernels would improve the computational complexity of the dual optimization problem.

ALL THE BEST