CS5691: Pattern recognition and machine learning
Final Exam

Name : **Roll No** :

Answer any 7 out of the 8 questions below. You may use any result proved in class (not tutorials) without proof. All figures must be neatly drawn using a ruler. No striking. All reference materials allowed. No seeking help from others. Write, scan convert to pdf, insert the question pages appropriately and submit. Deadline is 19 June, 2020, 11:59 PM. Email the pdf to me.

(1) **(EM algorithm.)** Consider random variables X, Z distributed as follows, with X taking values in $\{0,1\}^D$ and Z taking values in $\{1,2,\ldots,K\}$.

$$P(Z=z) = \pi_z$$
$$P(X=\mathbf{x}|Z=z) = \prod_{i=1}^{D} (\mu_{z,i})^{x_i}(1-\mu_{z,i})^{1-x_i}$$

The π and μ values are the parameters of the distribution.

Answer the following:

i. Give the total number of parameters in terms of D and K.

ii. Given i.i.d. samples $\mathbf{x}_1,\ldots,\mathbf{x}_N$ of the variable X (the variable Z is hidden), apply the EM algorithm to the mixture of Bernoulli model specified above. In particular, show the E and M steps in detail.

iii. Apply your algorithm for the below data with $N=8, D=3$ and $K=2$. Run the E and M steps for two times each, with the initialisation given by $\pi_1 = \pi_2 = 0.5$, and $\mu_1 = [0.1, 0.1, 0.1]$ and $\mu_2 = [0.5, 0.9, 0.5]$.

$$X = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}$$

(1+2+2 points)

(2) **(PCA.)**

i. Let n be a large number. Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be i.i.d samples from $\mathcal{N}\left(\begin{bmatrix}1\\2\end{bmatrix}, \begin{bmatrix}4 & -3\\-3 & 9\end{bmatrix}\right)$.
Let $S_{a,b,c} = \{\mathbf{x} \in \mathbb{R}^2 : ax_1 + bx_2 + c = 0\}$ be a line in $\mathbb{R}^2$. Let the approximation error of this line be:

$$R(a,b,c) = \min_{\mathbf{y}_1,\ldots,\mathbf{y}_n \in S_{a,b,c}} \frac{1}{n}\sum_{i=1}^{n} ||\mathbf{x}_i - \mathbf{y}_i||^2$$

Give a, b, c which minimises the above error.

ii. Let n be a large number. and $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be i.i.d. samples from D. Let distribution D be a mixture of two Gaussians over $\mathbb{R}^d$, i.e.

$$D = \frac{1}{2}\mathcal{N}(\mathbf{0}, A\Sigma A^\top) + \frac{1}{2}\mathcal{N}(\mathbf{0}, B\Lambda B^\top)$$

where A, B are orthonormal matrices. Σ, Λ are diagonal matrices, given as follows:

$$\Sigma = \text{diag}(8, 4, 2, 1, \frac{1}{2}, \frac{1}{2}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}, 0, 0, 0, 0, 0, 0)$$
$$\Lambda = \text{diag}(9, 5, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)$$

Let L_k be the approximation error of the datapoints $\mathbf{x}_i$ with the best k-dimensional hyperplane, i.e.

$$L_k = \min_{\mathbf{u}_1,\ldots,\mathbf{u}_d} \min_{\substack{\mathbf{z}_1,\ldots,\mathbf{z}_n \\ b_{k+1}\ldots,b_d}} \frac{1}{n}\sum_{i=1}^{n} ||\mathbf{x}_i - \sum_{j=1}^{k} z_{i,j}\mathbf{u}_j - \sum_{j=k+1}^{d} b_j \mathbf{u}_j||^2$$

where $\mathbf{u}_j \in \mathbb{R}^d$ and $\mathbf{z}_i \in \mathbb{R}^k$ and b values are scalars.
Show the following:

(a) $L_2 \leq 7.5$

(b) $L_5 \leq 2$.

(c) Also show that the above two bounds are tight. i.e. there exists orthonormal matrices A and B such that $L_2 = 7.5$ and $L_5 = 2$.

(2+3 points)

(3) **(Multiclass Logistic Regression.)**

i. Let $X|Y = i$ be distributed as the multivariate normal given by $\mathcal{N}(\boldsymbol{\mu}_i, \sigma^2 I)$ for all $i \in [K]$. Let π_i be equal to $P(Y = i)$. What is the posterior probability $P(Y = i|X = \mathbf{x})$?

ii. Consider the three class, 1-dimensional dataset, with 6 data points. With feature given by x and class label given by y.

x	3	2	5	5	7	8
y	1	1	2	2	3	3

The multinomial logistic loss is given as :

$$L = \sum_{i=1}^{6} -\log\left(\left[\mathrm{SM}\left(w_1 x_i + b_1, w_2 x_i + b_2, w_3 x_i + b_3\right)\right]_{y_i}\right)$$

where SM is the softmax function from $\mathbb{R}^3 \to \mathbb{R}^3_+$ and the parameters are w_j, b_j for $j \in \{1, 2, 3\}$. Give a setting for the parameters so that $L < 0.1$. Argue that the loss can be made arbitrarily close to zero for some setting of the parameters.

iii. Consider the same dataset as above. The loss minimised in one-vs-all logistic regression is:

$$L = \sum_{i=1}^{6} \sum_{j=1}^{3} -\log\left(\sigma(y_{ij}(w_j x_i + b_j))\right)$$

where σ is the sigmoid function. $y_{ij} = +1$ if $y_i = j$ and -1 otherwise. Show that for any setting of $w_1, w_2, w_3, b_1, b_2, b_3$ the loss L is greater than $2\log(2)$.

iv. Repeat the two sub-problems above, with the 2-dimensional 4-class dataset with 8 points given below as well. Note that the parameters are $\mathbf{w}_1, \mathbf{w}_2, \mathbf{w}_3, \mathbf{w}_4$ and b_1, b_2, b_3 and b_4, with $\mathbf{w}_j \in \mathbb{R}^2$ and $b_j \in \mathbb{R}$. The multinomial logistic and one-vs-all loss expressions also change appropriately.

x_1	7	8	4	5	1	0	4	3
x_2	7	8	4	4	7	6	0	0
y	1	1	2	2	3	3	4	4

(1+1+1+2 points)

(4) **(Support vector regression.)**

i. Consider a regression problem with training data $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ with $\mathbf{x}_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$. The Support Vector Regression algorithm, essentially solves the below problem:

$$\begin{aligned} & \min_{\mathbf{w},b} ||\mathbf{w}||^2 \\ \text{s.t.} \quad & \mathbf{w}^\top \mathbf{x}_i + b \leq y_i + \epsilon \\ & \mathbf{w}^\top \mathbf{x}_i + b \geq y_i - \epsilon \end{aligned}$$

for some $\epsilon > 0$. Derive the Lagrangian dual optimisation problem to the above problem.

ii. Solve the SVR problem for the 4 point 1-dimensional dataset below. Use $\epsilon = 1$.

x	1	2	3	4
y	1	1	3	4

Give both the primal and dual solution. Put some thought into how will the Lagrange multipliers will look like, and which constraints in the primal will be active for this problem for getting a jump-start.

(**3+2 points**)

(5) **(Vapnik-Chervonenkis dimension.)**

i. Let $\mathcal{H}_1, \mathcal{H}_2$ be hypothesis classes of functions from $\mathcal{X}$ to $\{0,1\}$, i.e. $\mathcal{H}_i \subseteq \{0,1\}^{\mathcal{X}}$. Let $\mathcal{H}_3$ be the hypothesis class obtained from the point-wise multiplication of $\mathcal{H}_1$ and $\mathcal{H}_2$, i.e.

$$\mathcal{H}_3 = \{h : \mathcal{X} \to \{0,1\} : h(x) = h_1(x) * h_2(x), \text{ for some } h_1 \in \mathcal{H}_1, h_2 \in \mathcal{H}_2\}.$$

In other words, every function in $\mathcal{H}_3$ is a product of a function in $\mathcal{H}_1$ and a function in $\mathcal{H}_2$. Let the VC-dimensions of $\mathcal{H}_1$ and $\mathcal{H}_2$ be V. Let V' be the VC-dimension of $\mathcal{H}_3$. Show that

$$\frac{V'}{\log_2(V')} \leq 2V$$

(Hint: Use Sauer's Lemma)

ii. Based on the above argument, and using the fact that the VC dimension of the set of half-spaces in $\mathbb{R}^d$ is $d+1$, show that the following hypothesis class $\mathcal{H}_3$ from $\mathbb{R}^2 \to \{0,1\}$ has VC-dimension at most 30.

$$\mathcal{H}_3 = \{h : \mathbb{R}^2 \to \{0,1\} : h(\mathbf{x}) = \mathbf{1}(\mathbf{w}_1^\top \mathbf{x} + b_1 \geq 0) * \mathbf{1}(\mathbf{w}_2^\top \mathbf{x} + b_2 \geq 0) \text{ for some } \mathbf{w}_1, \mathbf{w}_2, b_1, b_2\}$$

iii. For the above hypothesis class, show that the VC-dimension is at least 5. Give 5 points in $\mathbb{R}^2$ and 2^5 classifiers in $\mathcal{H}_3$ that achieve all the 2^5 bit patterns on these 5 points.

(2+1+2 points)

(6) **(AdaBoost.)** Consider the following binary classification dataset. Run AdaBoost for 3 iterations on the dataset, with the weak learner returning a best decision stump (equivalently a decision tree with one node) (equivalently a horizontal or vertical separator). Ties can be broken arbitrarily. Give the objects asked for below. Highlight your answer by boxing it.

x_1	x_2	y
-1	-2	$+1$
1	-2	$+1$
-1	0	$+1$
1	0	-1
-1	2	-1
1	2	-1

i. Give the weak learners h_t for $t = 1, 2, 3$.

ii. Give the "edge over random" γ_t, and the multiplicative factor β_t for $t = 1, 2, 3$.

iii. Give the predictions of the final weighted classifier h on the training points.

(**2+2+1 points**)

(7) **(Precision and Recall.)** Consider the following distribution of random variables (X, Y) over $\mathbb{R} \times \{-1, +1\}$ defined as follows.

$$P(Y = -1) = 0.7$$
$$P(Y = +1) = 0.3$$
$$f_{X|Y}(x|y = -1) = \begin{cases} \frac{x}{9} & \text{if } x \in [0, 3] \\ \frac{6-x}{9} & \text{if } x \in [3, 6] \\ 0 & \text{otherwise} \end{cases}$$
$$f_{X|Y}(x|y = +1) = \begin{cases} \frac{x-2}{9} & \text{if } x \in [2, 5] \\ \frac{8-x}{9} & \text{if } x \in [5, 8] \\ 0 & \text{otherwise} \end{cases}$$

Consider the 9 classifiers $h_0, h_2, \ldots, h_8$ given by thresholding at $0, 1, \ldots, 8$, i.e.

$$h_j(x) = \begin{cases} +1 & \text{if } x \geq j \\ -1 & \text{otherwise} \end{cases}.$$

i. Hand-plot the class conditional distributions approximately.

ii. Derive the true positive, true negative, false positive and false negative fractions of these 9 classifiers. (Note that they must sum to 1, for each classifier)

iii. Compute the precision and recall for all the 9 classifiers.

iv. Show the precision and recall of all the 9 classifiers in a hand drawn scatter plot.

(**1+2+1+1 points**)

(8) **(Kernels.)** Let the data instance X be a d-dimensional vector. A function $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is a valid kernel function if there exists $\phi : \mathbb{R}^d \to \mathbb{R}^{d'}$ such that $K(\mathbf{u}, \mathbf{v}) = \phi(\mathbf{u})^\top \phi(\mathbf{v})$. Such a ϕ is called a feature map for the kernel K.

i. Let $d = 2, k = 2$. Prove that $K(\mathbf{u}, \mathbf{v}) = (\mathbf{u}^\top \mathbf{v})^k$ is a valid kernel. Give the feature map corresponding to this kernel.

ii. Repeat the above for $d = 3, k = 2$.

iii. Repeat the above for $d = 2, k = 3$.

iv. Infer the general form of the feature map ϕ of the kernel $K : (\mathbf{u}, \mathbf{v}) \mapsto (\mathbf{u}^\top \mathbf{v})^k$ for any d, k.

v. Repeat the four items above for the kernel $K(\mathbf{u}, \mathbf{v}) = (1 + \mathbf{u}^\top \mathbf{v})^k$.

(**0.5+0.5+0.5+1+2.5 points**)