CS5691: Pattern recognition and machine learning
Final Exam

Name : **Roll No** :

Answer any 5 out of the 6 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 03 August, 2020, 09:00 AM. Submit on Moodle. If you cannot, then email the pdf to `hariguru@cse.iitm.ac.in` *. As a general hint, plots are a very useful tool, use them whenever you can.*

(1) **(PCA.)** Let P be the distribution $\mathcal{N}\left(\begin{bmatrix}0\\1\\2\end{bmatrix}, \begin{bmatrix}12 & 0 & -4\\0 & 1 & 0\\-4 & 0 & 12\end{bmatrix}\right)$. Let $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n \in \mathbb{R}^3$ be i.i.d. samples from P. Assume n is arbitrarily large.

i. Let $\mathbf{v}_1^\top = [1,0,0]$, $\mathbf{v}_2^\top = [0,1,0]$ and $\mathbf{v}_3^\top = [0,0,1]$. Evaluate the expression:

$$\min_{z_1,\ldots,z_n \in \mathbb{R}, b_2 \in \mathbb{R}, b_3 \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^{n} ||\mathbf{x}_i - z_i \mathbf{v}_1 - b_2 \mathbf{v}_2 - b_3 \mathbf{v}_3||^2$$

ii. Let $\mathbf{v}_1^\top = [1,0,0]$, $\mathbf{v}_2^\top = [0,1,0]$ and $\mathbf{v}_3^\top = [0,0,1]$. Evaluate the expression:

$$\min_{\mathbf{z}_1,\ldots,\mathbf{z}_n \in \mathbb{R}^2, b_3 \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^{n} ||\mathbf{x}_i - z_{i,1} \mathbf{v}_1 - z_{i,2} \mathbf{v}_2 - b_3 \mathbf{v}_3||^2$$

iii. Let $\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3$ be the principal components of the data $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$. What are $\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3$?

iv. Evaluate the expression:

$$\min_{z_1,\ldots,z_n \in \mathbb{R}, b_2 \in \mathbb{R}, b_3 \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^{n} ||\mathbf{x}_i - z_i \mathbf{u}_1 - b_2 \mathbf{u}_2 - b_3 \mathbf{u}_3||^2$$

v. Evaluate the expression:

$$\min_{\mathbf{z}_1,\ldots,\mathbf{z}_n \in \mathbb{R}^2, b_3 \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^{n} ||\mathbf{x}_i - z_{i,1} \mathbf{u}_1 - z_{i,2} \mathbf{u}_2 - b_3 \mathbf{u}_3||^2$$

vi. Let a_1, a_2, a_4, a_5 be the result of parts (i),(ii),(iv) and (v) above. Answer the following and give an argument for why that should be so.

A. Which is greater: a_1 or a_2?

B. Which is greater: a_4 or a_5?

C. Which is greater: a_1 or a_4?

D. Which is greater: a_2 or a_5?

(0.5+0.5+1+0.5+0.5+2 points)

(2) **(EM algorithm.)** Consider the following crowd-sourcing problem. Each data-point $1 \leq i \leq n$ has a true-label $Z_i \in \{0, 1\}$, and $P(Z_i = 1) = \alpha$. Each data point also has crowd-sourced labels $X_{i1}, \ldots, X_{id}$ given by d annotators, where $X_{ij} \in \{0, 1\}$. Each annotator j, is assumed to have an independent error μ_j, i.e.

$$P(X_{ij} = 1 | Z_i = 0) = \mu_j$$
$$P(X_{ij} = 0 | Z_i = 1) = \mu_j$$

The joint distribution of all the random variables involved is given by:

$$P(X_{11}, \ldots, X_{1d}, \ldots, X_{n1}, \ldots, X_{nd}, Z_1, \ldots, Z_n) = \prod_{i=1}^{n} \left(P(Z_i) \prod_{j=1}^{d} P(X_{ij} | Z_i) \right)$$

You have access to the dataset of n points annotated by d annotators, i.e. $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$, where $\mathbf{x}_i \in \{0, 1\}^d$. But the true labels $z_1, z_2, \ldots, z_n \in \{0, 1\}$ are hidden. Use the EM algorithm to estimate the error rates of the annotators μ_j and the prior probability parameter α.

i. For a given current value of the parameters α and $\mu_1, \ldots \mu_d$, give the E-step where you will compute $\gamma_i = P(Z_i = 1 | X_i = \mathbf{x}_i)$.

ii. Consider the estimation problem for annotator j. Given the complete data $(z_1, x_{1j}), \ldots, (z_n, x_{nj})$, with $z_i \in \{0, 1\}$ and $x_{ij} \in \{0, 1\}$, give the ML estimate of the probability of error μ_j.

iii. Use the γ_i from part (a), along with the solution to the maximum likelihood problem above to give an expression for the update of the parameters $\alpha, \mu_1, \ldots, \mu_d$ in the M-step.
(Hint: It might be helpful to use γ_i to hallucinate complete data.)

iv. Consider the annnotation data below with 20 points and 4 labellers. Apply the E-step and M-step derived above for 2 iterations. In particular, give the result of the E-step $\gamma_1, \ldots, \gamma_{20}$ at the end of each of the two iterations. Also give the result of the M-step $\alpha, \mu_1, \mu_2, \mu_3, \mu_4$ at the end of each of the two iterations. Initialize with $\alpha = 0.5, \mu_1 = \mu_2 = \mu_3 = \mu_4 = 0.1$. Interpret the likely true labels Z for the 20 points, and the quality of the 4 labellers based on the final γ and μ values. How does the final γ_i differ from a simple majority vote of the annotator labels?

Instructions: Feel free to write and use a computer code to get these answers. If you are using a computer code, you can run 10 steps of EM algorithm, instead of 2, for a much easier interpretation of the final answer. Give answers only upto 3 significant digits. Give the numerical results for the E-step and M-step in two tables of size 20*2 and 5*2 respectively.

$x_{:,1}$	1	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	0	1	1	1
$x_{:,2}$	0	0	1	0	0	0	0	0	0	0	1	1	1	1	0	1	1	1	1	1
$x_{:,3}$	1	0	0	1	1	1	0	1	0	0	1	0	1	0	1	1	0	0	1	0
$x_{:,4}$	1	1	1	1	1	0	1	1	1	1	0	0	0	1	0	0	0	0	0	0

(1+1+2+(1+2) points, including 2 bonus points.)

(3) **(Collaborative filtering.)** Consider the three movie, three user rating dataset below, represented by the ratings matrix R.

	M1	M2	M3
U1	4	2	1
U2	2	4	1
U3	5	5	3

Consider the baseline model where $R_{u,i} = \mu + a_u + b_i + \epsilon_{u,i}$ for some parameters $\mathbf{a} \in \mathbb{R}^3$ and $\mathbf{b} \in \mathbb{R}^3$ and $\mu \in \mathbb{R}$ and $\epsilon_{u,i}$ is noise.

i. Set μ to an appropriate value. What would that be?

ii. Solve for $\mathbf{a}$ and $\mathbf{b}$ by solving the following optimisation problem:

$$\min_{\mathbf{a}\in\mathbb{R}^3, \mathbf{b}\in\mathbb{R}^3} \frac{1}{2} \sum_{u=1}^{3} \sum_{i=1}^{3} (R_{u,i} - \mu - a_u - b_i)^2$$

using full gradient descent till convergence. (Hint: Initialise $\mathbf{a}$ and $\mathbf{b}$ to the zero vector, and use step size $\eta = \frac{1}{3}$)

iii. Interpret the values $\mathbf{a}$ and $\mathbf{b}$ got above.

iv. Let $\widehat{R}_{u,i} = \mu + a_u + b_i$ be the estimate of the ratings from the baseline model. Give $\widehat{R}_{u,i}$ for all u, i in the form of a matrix.

v. What property does the difference matrix $R - \widehat{R}$ satisfy?

vi. What aspect of the data R does the baseline prediction $\widehat{R}$ capture?

vii. What aspect of the data R does the baseline prediction $\widehat{R}$ fail to capture?

viii. Give scalars p_1, p_2, p_3 for the users, and scalars q_1, q_2, q_3 for the movies such that $R_{u,i} = \mu + a_u + b_i + p_u q_i$. Interpret the p values and q values.

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

(4) **(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	1	2	3	4	3	4	7	7
x_2	1	0	4	3	6	6	2	3
y	1	1	2	2	3	3	4	4

(1+1+1+2 points)

(5) **(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 constant $\epsilon > 0$. 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	2	3	5	6
y	1	1	2	4

Give both the primal and dual solution. Argue how you got the solution, and show that the solution you have got is optimal.
(Hint: 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)

(6) **(Kernels.)** Let K_1 and K_2 be valid kernel functions, with feature mappings $\phi_1 : \mathbb{R}^d \to \mathbb{R}^{d_1}$ and $\phi_2 : \mathbb{R}^d \to \mathbb{R}^{d_2}$.

i. Show that $K_3 = K_1 + K_2$ is also a valid kernel function. Give the feature mapping ϕ_3 corresponding to K_3 in terms of ϕ_1 and ϕ_2.

ii. Show that $K_4 = K_1 \cdot K_2$ is also a valid kernel function. Give the feature mapping ϕ_4 corresponding to K_4 in terms of ϕ_1 and ϕ_2.

iii. Show that kernel defined as $K_5(\mathbf{u}, \mathbf{v}) = f(\mathbf{u})K_1(\mathbf{u}, \mathbf{v})f(\mathbf{v})$ is also a valid kernel for any function $f : \mathbb{R}^d \to \mathbb{R}$. Give the feature mapping ϕ_5 corresponding to K_5 in terms of ϕ_1, f

iv. Show that the kernel given by $K(\mathbf{u}, \mathbf{v}) = \exp(2\mathbf{u}^\top \mathbf{v})$ is a valid kernel.

v. Show that the kernel given by $K(\mathbf{u}, \mathbf{v}) = \exp(-||\mathbf{u} - \mathbf{v}||^2)$ is a valid kernel.

(1+1+1+1+1 points)

(Hint: Express $\exp(t)$ as a polynomial expansion and use previous results.