I, NEVIN LIANG WITH UID 705575353 have read & understood the policy on academic integrity.

1. (30 points) **True or False.**

Circling the correct answer is worth +3 points, circling the incorrect answer is worth −1 points. Not circling either is worth 0 points.

(a) The perceptron algorithm does not converge if the training samples are not linearly separable.

(TRUE) FALSE

(b) Logistic Regression is a linear classifier.

(TRUE) FALSE

(c) We learned that the soft margin SVM have the primal problem:

$$\begin{aligned} \min_{\xi,w,b} \quad & \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}\xi_i \\ s.t. \quad & y^{(i)}(w^T x^{(i)} + b) \geq 1 - \xi_i, \quad i = 1, \cdots, m \\ & \xi_i \geq 0, \quad i = 1, \cdots, m \end{aligned}$$

and the dual problem:

$$\begin{aligned} \max_{\alpha} \quad & W(\alpha) = \sum_{i=1}^{m}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{m} y^{(i)}y^{(j)}\alpha_i\alpha_j\langle x^{(i)}, x^{(j)}\rangle \\ s.t. \quad & 0 \leq \alpha_i \leq C, i = 1, \cdots, m \\ & \sum_{i=1}^{m}\alpha_i y^{(i)} = 0. \end{aligned}$$

Suppose we have solved the dual problem and have the optimal α. If we find that $0 < \alpha_i < C$, then $x^{(i)}$ is **inside** the margin, i.e., $y^{(i)}(w^T x^{(i)} + b) < 1$.

TRUE (FALSE)

(d) For $x_1, x_2 \in \mathbb{R}$, $K(x_1, x_2) = (1 + x_1 x_2)^2$ is a valid kernel.

(TRUE) FALSE

(e) The Naive Bayes assumption assumes that the features x_i are independent.

TRUE (FALSE)

(f) The decision boundary for the GDA model with equal class priors is linear.

TRUE (FALSE)

(g) For a given fixed set of data points and a fixed k, k-means always converges to the same clustering of the data.

TRUE (FALSE)

(h) In PCA, we want to **minimize** the variance of the projected data when we project the data onto a lower dimension.

TRUE (FALSE)

(i) To do bagging, we sample the original dataset without replacement and train weak classifiers using the sampled dataset. This procedure can be done in parallel.

TRUE (FALSE)

(j) In Adaboost, the weight α for each weak classifier is calculated based on the ratio of misclassified examples to the total number of examples.

(TRUE) FALSE

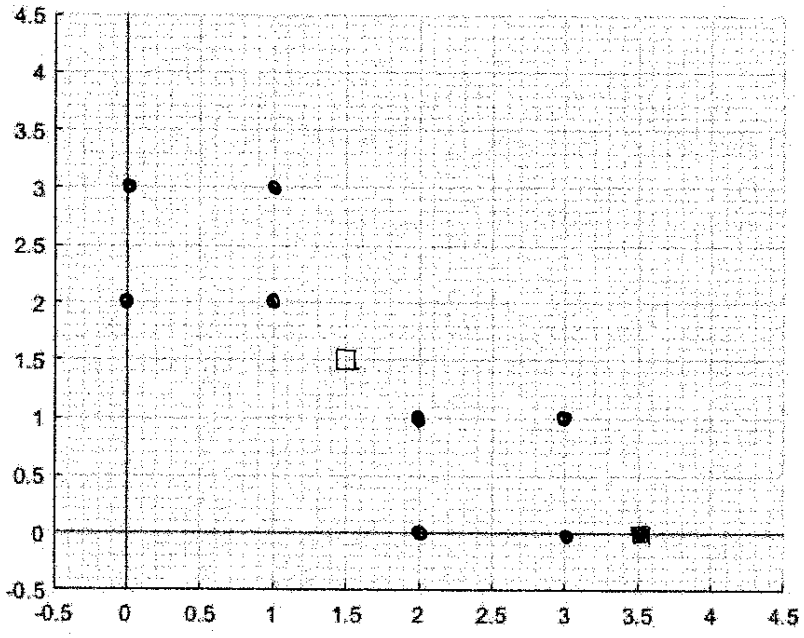
2. (15 pts) **k-Means Clustering**
The algorithm for k-means clustering are as follows: choose initial 'representatives' $z_1, ..., z_k$ for the k groups and repeat:

   (a) assign each vector x_i to the nearest group representative z_j.

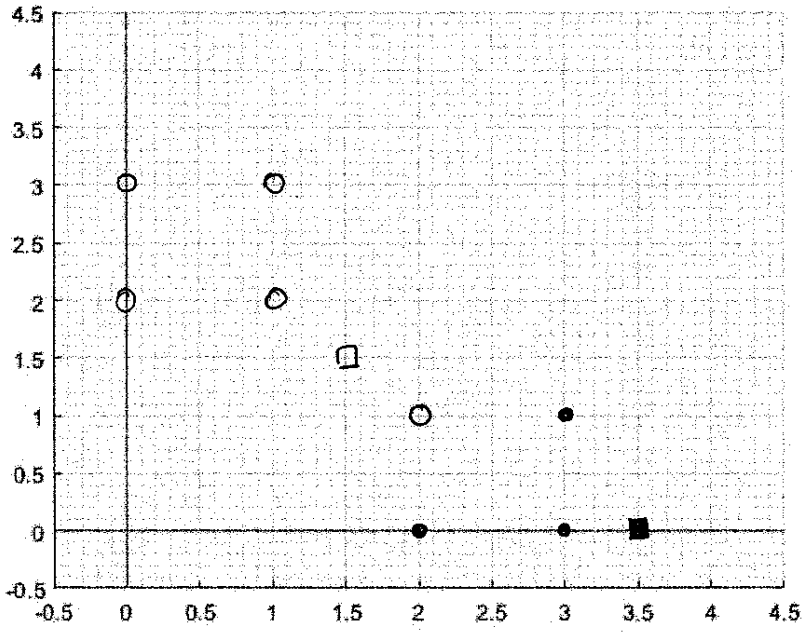
   (b) set the representative z_j to the mean of the vectors assigned to it.

   In this problem, you will perform k-means algorithm with $k = 2$ in the following data set: $x_1 = (0,3)$, $x_2 = (1,3)$, $x_3 = (0,2)$, $x_4 = (1,2)$, $x_5 = (2,1)$, $x_6 = (3,1)$, $x_7 = (2,0)$, $x_8 = (3,0)$, and initial 'representatives': $z_1 = (1.5, 1.5)$, $z_2 = (3.5, 0)$.

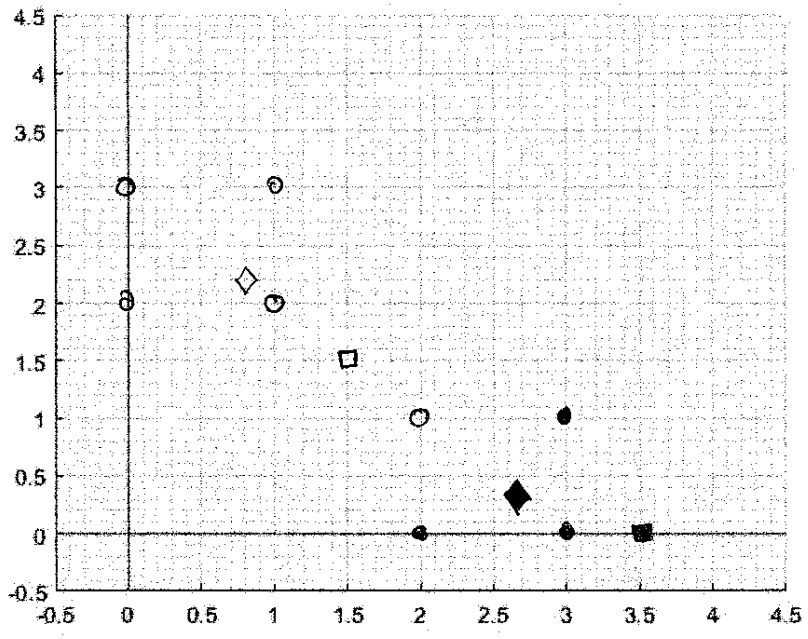
   (a) In the following figure, plot the data x_i with filled circle '●', one initial 'representatives' with empty square '□', and the other representatives with filled squares '■'.



   (b) In the following figure, assign each data point x_i to the nearest group representative z_j. Plot one group of data points with empty circle '○' and its 'representatives' with empty box '□', and the other group of data points with filled circle '●' and its 'representatives' with filled box '■'. (In the figure, your 'representatives' should be the same as what you have in part (a)).

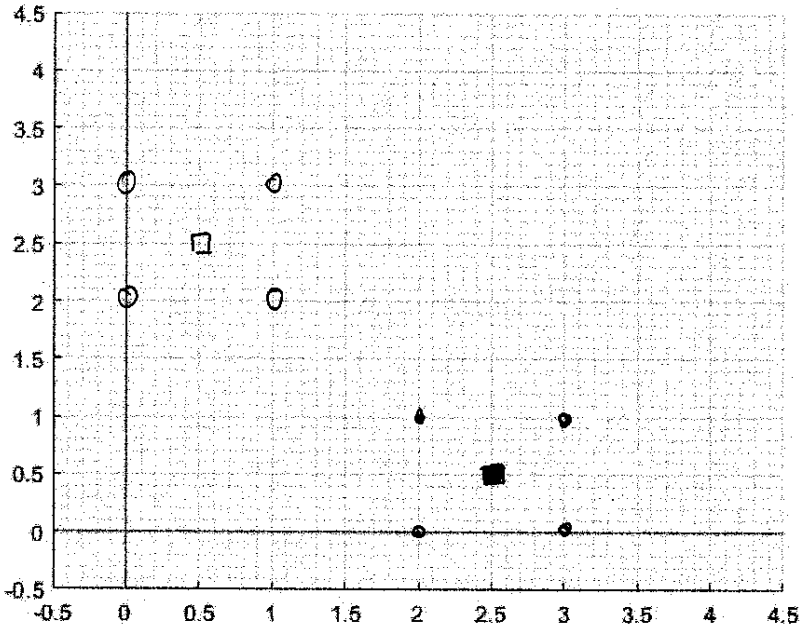


(c) Update the representative z_j to the mean of the vectors assigned to it. What are the values of updated 'representative' z_1 and z_2? Copy you data points from part b, and plot one of the updated 'representative' with empty diamond '◇' and the other with filled diamond '◆'.



(d) Perform the algorithm untill it converges, i.e., the 'representative' doesn't change anymore. Plot your final result in the following figure. Plot one group of data

points with empty circle '○' and its 'representatives' with empty box '□', and the other group of data points with filled circle '●' and its 'representatives' with filled box '■'.



3. (15 pts) **Optimization**
Solve the following optimization problem. Justify your answer.

$$\begin{aligned} \min_{x} \quad & x^T A x \\ \text{subject to} \quad & x^T x = 1, \end{aligned}$$

where $A = \begin{bmatrix} 5 & 3 \\ 3 & 5 \end{bmatrix}$ and x is a vector in $\mathbb{R}^2$. In you solution, you should justify the following:

(a) Is the problem a convex probelm, why?

(b) What is the Lagrangian?

(c) What is the optimal value of $x^T A x$?

(d) What is the optimal x?

a) Yes. We are tying to minimize a function over a convex set

$x^T x = 1$ = circle or sphere or hypersphere which is convex b/c $x_1 \rightarrow x_2$ does not go outside sphere.

b) $\mathcal{L} = x^T A x + \lambda(1 - x^T x)$

c) $\frac{\partial \mathcal{L}}{\partial x} = 2Ax - 2\lambda x = 0 \rightarrow Ax = \lambda x \rightarrow \begin{bmatrix} 5 & 3 \\ 3 & 5 \end{bmatrix} x = \lambda x \rightarrow$ find eigenvals.

$$\begin{vmatrix} 5-\lambda & 3 \\ 3 & 5-\lambda \end{vmatrix} = \lambda^2 - 10\lambda + 25 - 9 = (\lambda - 8)(\lambda - 2) = 0 \rightarrow \lambda = 2, 8$$

minimize $\rightarrow$ smallest $\lambda = \boxed{2}$

d) $\begin{bmatrix} 3 & 3 \\ 3 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 0 \qquad 3x_1 + 3x_2 = 0 \rightarrow x_1 = -x_2 \rightarrow$ normalize $\rightarrow -x_1 = x_1 = \frac{\sqrt{2}}{2}$

$$x = \boxed{\begin{bmatrix} \sqrt{2}/2 \\ -\sqrt{2}/2 \end{bmatrix}}$$

4. (15 pts) **Weighted Least Square**
We have a set of data $x_n \in \mathbb{R}^M, y_n \in \mathbb{R}, n = 1, \cdots, N$. Suppose we want to find $w \in \mathbb{R}^M$ that minimize the following objective weighted least square loss function:

$$J(w) = \sum_{n=1}^{N} \alpha_n (w^T x_n - y_n)^2,$$

where $\alpha_n > 0$ are the weights for each data point.

(a) Write the loss function in matrix-vector form, i.e., as a squared 2-norm of some vector. Hint: you may use the following matrix: $A^{\frac{1}{2}}$ with $A^{\frac{1}{2}}_{nn} = \sqrt{\alpha_n}, n = 1, \cdots, N$ and all other elements being 0.

(b) The normal equation for the w that minimizes $J(w)$ is of the form $Bw = c$. Find B and c in terms of the data and the weights. Hint: you may use the following matrix: A with $A_{nn} = \alpha_n, n = 1, \cdots, N$ and all other elements being 0.

(c) Show that $J(w)$ has a global minimum. Hint: show that the Hessian matrix is positive-definite.

a) $J(w) = \sum_{n=1}^{N} a_n (w^T x_n - y_n)^2 = \left\| \begin{bmatrix} \sqrt{a_1}(w^T x_1 - y_1) \\ \sqrt{a_2}(w^T x_2 - y_2) \\ \vdots \end{bmatrix} \right\|^2 = \| A^{1/2} \cdot (Xw - y) \|^2$

where $X = \begin{bmatrix} — x_1^T — \\ — x_2^T — \\ \vdots \end{bmatrix}$ — some vector.

b) Normal equation = $A^TAx = A^Tb$ for $\|Ax-b\|^2$:

$$(A^{1/2}X)^T (A^{1/2}X) w = (A^{1/2}X)^T A^{1/2} y$$

$$\underbrace{X^T A X}_{B} w = \underbrace{X^T A y}_{c}$$

← diff c's and b's oops.

c) $J(w) = \|Cw - b\|^2$ where $C = A^{1/2}X$ and $b = A^{1/2}y$

$= (Cw-b)^T(Cw-b) = (w^TC^T - b^T)(Cw - b) = b^Tb - 2b^TCw + w^TC^TCw$

$\nabla_w J = 2C^TCw - 2C^Tb$

$H_w J = 2C^TC$

proof of P.D. $y^T H y = y^T 2C^T C y = 2(Cy)^T Cy = 2\|Cy\|^2 \geq 0$

It is only 0 when $y = 0$. P.D. states $y^T H y > 0$ for $y \neq 0$.

5. (15 pts) **Maximum Likelihood Estimation**
Let $x_1, x_2, \cdots, x_n$ be independent samples from the following distribution:

$$P(x|\theta) = \theta x^{-\theta-1} \quad \text{where} \quad \theta > 1, x \geq 1.$$

(a) Find the maximum likelihood estimate of θ that maximizes $P(x_1, x_2, \cdots, x_n|\theta)$.

(b) Show that the estimator you get in (a) indeed maximizes $P(x_1, x_2, \cdots, x_n|\theta)$ instead of minimizing it. I.e, show that the second derivative of $P(x_1, x_2, \cdots, x_n|\theta)$ with respect to θ is non-negative.

(c) Estimate θ if you have data $x_1 = 3, x_2 = 5, x_3 = 2$, and $x_4 = 10$. You may leave you answer with natural log(s).

a) $P(x_1, x_2, x_3 \ldots x_n|\theta) = \prod_{i=1}^{n} P(x_i|\theta) = \prod_{i=1}^{n} \theta x_i^{-\theta-1}$

maximize $\prod_{i=1}^{n} \theta x_i^{-\theta-1}$ = maximize $\ln\left(\prod_{i=1}^{n} \theta x_i^{-\theta-1}\right)$

$$\frac{\partial}{\partial\theta}\left(\sum_{i=1}^{n} \ln\left(\theta x_i^{-\theta-1}\right)\right) = \sum_{i=1}^{n} \frac{1}{\theta x_i^{-\theta-1}} \cdot \left(x_i^{-\theta-1} + \theta \cdot (-x_i^{-\theta-1} \ln x_i)\right)$$

$$= \sum_{i=1}^{n} \frac{1}{\theta} + -\ln x_i = \frac{n}{\theta} - \sum_{i=1}^{n} \ln x_i = 0$$

$$\boxed{\theta = \frac{n}{\sum_{i=1}^{n} \ln x_i}}$$

b) $\frac{\partial^2}{\partial\theta^2}\left(\sum_{i=1}^{n} \ln\left(\theta x_i^{-\theta-1}\right)\right) = \frac{\partial}{\partial\theta}\left(\frac{n}{\theta} - \sum_{i=1}^{n} \ln x_i\right) = n \cdot (-1)\theta^{-2}$

$= -n \cdot \theta^{-2} < 0$

so maximum

c) $\theta = \frac{4}{\sum_{i=1}^{n} \ln x_i} = \frac{4}{\ln 3 + \ln 5 + \ln 2 + \ln 10} = \boxed{\frac{4}{\ln 300}}$

6. (10 pts) **Expectation maximization**

You learned that the log likelihood function for the Gaussian mixture model is of this form:

$$J = \ln P(X; \pi, \mu, \Sigma) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(x_n; \mu_k, \Sigma_k) \right\}.$$

Here, π_k is the prior probability of the latent variable; μ_k and Σ_k are the mean and covariance matrix for the k-th Gaussian component. Suppose the data points x_n are D-dimensional.

Suppose we want to maximize J with respect to μ_l. Show that the μ_l that maximize J is of the form:

$$\hat{\mu}_l = \frac{\sum_{n=1}^{N} \gamma_{nl} x_n}{\sum_{n=1}^{N} \gamma_{nl}},$$

where

$$\gamma_{nk} = \frac{\pi_k \mathcal{N}(x_n; \mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x_n; \mu_j, \Sigma_j)}.$$

You may assume that all γ_{nk} are known for this step. You **must show all necessary steps (e.g., application of chain rules) to get full credit** for this question. Matrix calculus results can be used without proof.

$$\frac{\partial J}{\partial \mu_l} = \sum_{n=1}^{N} \frac{\pi_l \frac{\partial}{\partial \mu_l}\left(\mathcal{N}(x_n; \mu_l, \Sigma_l)\right)}{\sum_{k=1}^{K} \pi_k \mathcal{N}(x_n; \mu_k, \Sigma_k)}$$

$$\frac{\partial}{\partial \mu_l}\left(\mathcal{N}(x_n; \mu_l, \Sigma_l)\right) = \frac{\partial}{\partial \mu_l}\left(\frac{1}{(2\pi)^{D/2} \det(\Sigma_l)^{1/2}} \exp\left(\frac{-1}{2}(x_n - \mu_l)^T \Sigma_l^{-1} (x_n - \mu_l)\right)\right)$$

$$= \frac{1}{(2\pi)^{D/2} \det(\Sigma_l)^{1/2}} \exp\left(\frac{-1}{2}(x_n - \mu_l)^T \Sigma_l^{-1} (x_n - \mu_l)\right) \cdot \left(\frac{-1}{2}\right)\left(2\Sigma_l^{-1}(x_n - \mu_l)\right)(-1)$$

$$= \mathcal{N}(x_n; \mu_l, \Sigma_l) \cdot \Sigma_l^{-1}(x_n - \mu_l)$$

$$= \sum_{n=1}^{N} \frac{\pi_l \mathcal{N}(x_n; \mu_l, \Sigma_l)}{\sum_{k=1}^{K} \pi_k \mathcal{N}(x_n; \mu_k, \Sigma_k)} \cdot \Sigma_l^{-1}(x_n - \mu_l)$$

$$= \sum_{n=1}^{N} \gamma_{nl} \cdot \Sigma_l^{-1}(x_n - \mu_l) = 0 \quad \rightarrow \sum_{n=1}^{N} \gamma_{nl}(x_n - \mu_l) = 0$$

$$\mu_l = \frac{\sum_{n=1}^{N} \gamma_{nl} x_n}{\sum_{n=1}^{N} \gamma_{nl}} \quad \blacksquare$$