Instructions:
1. This question paper contains 2 pages (4 sides of paper). Please verify.
2. Write your name, roll number, department in **block letters** with **ink** on **each page**.
3. Write your final answers neatly **with a blue/black pen**. Pencil marks may get smudged.
4. Don't overwrite/scratch answers especially in MCQ – ambiguous cases may get 0 marks.

Q1. (Total Confusion) The *confusion matrix* is often used to evaluate classification models. For a C-class problem, this is a $C \times C$ matrix that tells us, for any two classes $c, c' \in [C]$, how many instances of class c were classified as c' by the model. In the example below, $C = 2$, there were $P + Q + R + S$ points in the test set where P, Q, R, S are strictly positive integers. The matrix tells us that Q points were in class $+1$ but were (incorrectly) classified as -1 by the model, S points were in class -1 and were (correctly) classified as -1 by the model, etc. **Give expressions for the specified quantities in terms of** P, Q, R, S. No derivations needed. Note that y denotes the true class of a test point and $\hat{y}$ is the predicted class for that point. **(5 x 1 = 5 marks)**

		Predicted class $\hat{y}$: $+1$	-1
True class y	$+1$	P	Q
	-1	R	S

Confusion Matrix

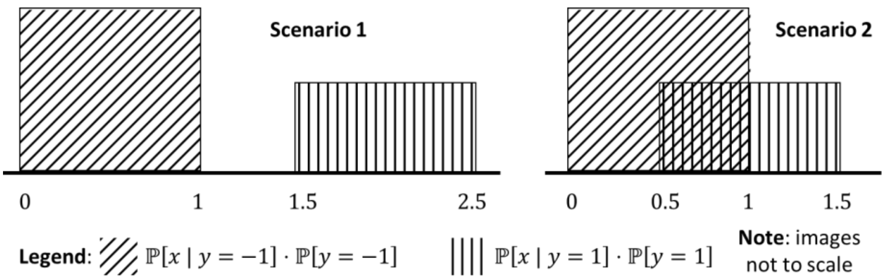
True positive rate (**TPR**) $\mathbb{P}[\hat{y} = 1 \mid y = 1]$	$\frac{P}{P+Q}$
False positive rate (**FPR**) $\mathbb{P}[\hat{y} = 1 \mid y = -1]$	$\frac{R}{R+S}$
True negative rate (**TNR**) $\mathbb{P}[\hat{y} = -1 \mid y = -1]$	$\frac{S}{R+S}$
False negative rate (**FNR**) $\mathbb{P}[\hat{y} = -1 \mid y = 1]$	$\frac{Q}{P+Q}$
Misclassification rate (**MIS**) $\mathbb{P}[\hat{y} \neq y]$	$\frac{Q+R}{P+Q+R+S}$

Q2. (Kernel Smash) Melbi has created two Mercer kernels $K_1, K_2: \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ with the feature map for K_i as $\phi_i: \mathbb{R} \to \mathbb{R}^3$ i.e., for any $x, y \in \mathbb{R}$, we have $K_i(x, y) = \langle \phi_i(x), \phi_i(y) \rangle$ for $i \in \{1,2\}$. Melbi tells us that $\phi_1(x) = (x, x^3, x^5)$ and $\phi_2(x) = (1, x^2, x^4)$. Melbo creates two new kernels K_3, K_4 so that for any $x, y \in \mathbb{R}$, $K_3(x, y) = K_1(x, y) + K_2(x, y)$ and $K_4(x, y) = K_1(x, y) \cdot K_2(x, y)$. Design feature maps $\phi_3, \phi_4: \mathbb{R} \to \mathbb{R}^6$ for the kernels K_3, K_4. **Note that ϕ_3, ϕ_4 must not have more than 6 dimensions each.** Write your answer in the space provided. No derivations required. If your feature map for either K_3 or K_4 requires fewer than 6 dimensions, you may fill-in the rest of the dimensions with 0 features. **(2 + 3 = 5 marks)**

$$\phi_3(x) = (1, x, x^2, x^3, x^4, x^5)$$

$$\phi_4(x) = (x, \sqrt{2}x^3, \sqrt{3}x^5, \sqrt{2}x^7, x^9, 0)$$

Q3. (The optimal classifier) Melbo is solving a binary classification problem with 1D features where features of datapoints labelled -1 are uniformly distributed in the interval $[a, a+1]$ and features of points labelled $+1$ are uniformly distributed in $[b, b+1]$ i.e., $\mathbb{P}[x \mid y=-1] = \mathcal{U}([a, a+1])$ and $\mathbb{P}[x \mid y=+1] = \mathcal{U}([b, b+1])$. Thrice as many points are labelled -1 as are labelled $+1$ i.e., $\mathbb{P}[y=-1] = 3 \cdot \mathbb{P}[y=+1]$. Melbo wants to learn a ***threshold classifier*** $f_\eta(\cdot)$ that classifies a data point with feature x as $+1$ if $x \geq \eta$ and as -1 if $x < \eta$. We have two scenarios for which the figure below depicts $\mathbb{P}[x \mid y] \cdot \mathbb{P}[y]$ (**images not to scale**). In scenario 1, we have $a = 0, b = 1.5$ while in scenario 2, we have $a = 0, b = 0.5$. $\mathbb{P}[y=+1]$ is the same for both scenarios. Your answer for parts 1, 2, 3, 7 and 11 should be a real number e.g., 0.5 or 1.414 etc. Your answer for parts 4, 5, 6, 8, 9 and 10 should be an expression in η e.g., $\frac{(1-\eta)}{2}$ or $\frac{\eta^2}{2} + \frac{1}{2}$, etc.



If a part has multiple correct answers, any correct answer will receive full marks. *Hint: pay attention to the form of* $f_\eta(x)$ *and how it uses the threshold.* **(1+2+2+1+1+1+2+1+1+1+2 = 15 marks)**

1	Find the value of $\mathbb{P}[y=+1]$. Recall that this value remains same in both scenarios.	$\frac{1}{4}$
	For parts 2 and 3, we consider scenario 1 i.e. $\boldsymbol{a = 0, b = 1.5}$	
2	Find a value η for which the misclassification rate of Melbo's classifier is the smallest i.e., find $\arg\min_\eta \{\mathbb{P}[f_\eta(x) \neq y]\}$.	Any value $\eta \in (1, 1.5]$ will give $\mathbb{P}[f_\eta(x) \neq y] = 0$
3	Find a value η for which the sum TPR and TNR is largest i.e., $\arg\max_\eta \{\mathbb{P}[f_\eta(x) = 1 \mid y = 1] + \mathbb{P}[f_\eta(x) = -1 \mid y = -1]\}$.	Any value $\eta \in (1, 1.5]$ will give $\text{TPR} + \text{TNR} = 2$
	For parts 4, 5, 6, 7, 8, 9, 10, 11 we consider scenario 2 i.e. $\boldsymbol{a = 0, b = 0.5}$	
4	Give an expression for the misclassification rate of Melbo's classifier $\mathbb{P}[f_\eta(x) \neq y]$ if Melbo chooses $\eta \in [0, 0.5)$	$\frac{3}{4} \cdot (1-\eta) + 0 = \frac{3}{4} - \frac{3\eta}{4}$
5	Give an expression for the misclassification rate of Melbo's classifier $\mathbb{P}[f_\eta(x) \neq y]$ if Melbo chooses $\eta \in [0.5, 1)$	$\frac{3}{4} \cdot (1-\eta) + \frac{1}{4} \cdot \left(\eta - \frac{1}{2}\right) = \frac{5}{8} - \frac{\eta}{2}$
6	Give an expression for the misclassification rate of Melbo's classifier $\mathbb{P}[f_\eta(x) \neq y]$ if Melbo chooses $\eta \in [1, 1.5]$	$0 + \frac{1}{4} \cdot \left(\eta - \frac{1}{2}\right) = \frac{\eta}{4} - \frac{1}{8}$
7	Using your solutions to parts 4, 5 and 6, find a value of η with smallest misclassification rate i.e., $\arg\min_\eta \{\mathbb{P}[f_\eta(x) \neq y]\}$.	1
8	Give an expression for the TPR + TNR of Melbo's classifier $\mathbb{P}[f_\eta(x) = 1 \mid y = 1] + \mathbb{P}[f_\eta(x) = -1 \mid y = -1]$ if Melbo chooses a $\eta \in [0, 0.5)$	$1 + \eta$
9	Give an expression for the TPR + TNR of Melbo's classifier $\mathbb{P}[f_\eta(x) = 1 \mid y = 1] + \mathbb{P}[f_\eta(x) = -1 \mid y = -1]$ if Melbo chooses a $\eta \in [0.5, 1)$	$(1.5 - \eta) + \eta = 1.5$

10 Give an expression for the TPR + TNR of Melbo's classifier $\mathbb{P}[f_\eta(x) = 1 \vert y = 1] + \mathbb{P}[f_\eta(x) = -1 \vert y = -1]$ if Melbo chooses a $\eta \in [1, 1.5)$	$(1.5 - \eta) + 1 = 2.5 - \eta$
11 Using your solutions to parts 8, 9, 10, find a value of η for which the TPR + TNR of the classifier is the largest i.e., $\arg\max_\eta \{\mathbb{P}[f_\eta(x) = 1 \vert y = 1] + \mathbb{P}[f_\eta(x) = -1 \vert y = -1]\}$	Any value $\eta \in (0.5,1)$ will give TPR + TNR $= 1.5$

Q4. (Opt. to Prob.) Melbo has a regression problem with 1D features, n datapoints $(x_i, y_i), i \in [n]$ with $x_i, y_i \in \mathbb{R}$ and tried learning a 1D linear model w by solving the optimization problem: $\min_{w \in [-1,1]} \left\{|w| + \frac{1}{2}\sum_{i\in[n]}(y_i - w \cdot x_i)^2\right\}$. **Note**: w is a 1D scalar, is constrained in the interval $[-1,1]$ and is also L_1 regularized. Melbo's friend Melba claims that this is just a MAP solution. To convince Melbo, create a likelihood distribution over labels $\mathbb{P}[y \mid x, w]$ and a prior distribution over models $\mathbb{P}[w]$ s.t. $\underset{w\in\mathbb{R}}{\arg\max} \left\{\mathbb{P}[w] \cdot \prod_{i\in[n]} \mathbb{P}[y_i \mid x_i, w]\right\} = \underset{w\in[-1,1]}{\arg\min} \left\{|w| + \frac{1}{2}\sum_{i\in[n]}(y_i - w \cdot x_i)^2\right\}$. **Give brief derivation**. *Hint: the prior (and not the likelihood) will introduce the constraint. The PDF for a Gaussian with mean μ and variance $\sigma^2 > 0$ is $\mathcal{N}(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$. The PDF for a Laplacian with mean m and scale $s > 0$ is $\mathcal{L}(x; m, s) = \frac{1}{2s}\exp\left(-\frac{|x-m|}{s}\right)$.* **(2 + 3 = 5 marks)**

$$\mathbb{P}[y \mid x, w] = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(y - w\cdot x)^2}{2}\right)$$

$$\mathbb{P}[w] = \begin{cases} \frac{e}{2(e-1)} \exp(-|w|) & |w| \le 1 \\ 0 & |w| > 1 \end{cases}$$

The loss term for a single point looks like $\frac{(y_i - w\cdot x_i)^2}{2}$ which corresponds to the negative log likelihood. Negating and exponentiating gives us a likelihood that looks like $\mathbb{P}[y_i \mid x_i, w] \propto \exp\left(-\frac{(y_i - w\cdot x_i)^2}{2}\right)$. Normalizing so that the likelihood integrates to unity directs us to choose a Gaussian with $\sigma = 1$.

Consider the following barrier regularization function $r(w) = \begin{cases} |w| & |w| \le 1 \\ \infty & |w| > 1 \end{cases}$. Note that we have $\underset{w\in[-1,1]}{\arg\min} \left\{|w| + \frac{1}{2}\sum_{i\in[n]}(y_i - w \cdot x_i)^2\right\} = \underset{w\in\mathbb{R}}{\arg\min} \left\{r(w) + \frac{1}{2}\sum_{i\in[n]}(y_i - w \cdot x_i)^2\right\}$. Negating and exponentiating as before tells us that the prior should look like $\mathbb{P}[w] \propto \begin{cases} \exp(-|w|) & |w| \le 1 \\ 0 & |w| > 1 \end{cases}$.

Normalizing so that the prior integrates to unity solves the problem.

Q5. (Unity in Diversity) We have 2 datasets with d-dim features. Dataset 1 has m points, features $X = [\mathbf{x}_1, \dots, \mathbf{x}_m] \in \mathbb{R}^{m \times d}$ and labels $\mathbf{u} = [u_1, \dots, u_m] \in \mathbb{R}^m$. Dataset 2 has n points, features $Z = [\mathbf{z}_1, \dots, \mathbf{z}_n] \in \mathbb{R}^{n \times d}$ and labels $\mathbf{v} = [v_1, \dots, v_n] \in \mathbb{R}^n$. Melbo wants to learn a linear model using both datasets but has been told by Melbi that the while the optimal model vector $\mathbf{w}^* \in \mathbb{R}^d$ is the same for both datasets, the datasets require different bias terms b_1^* and b_2^* such that $b_2^* = 2 \cdot b_1^*$. Thus, Melbo wants to solve the following problem: ($\mathbf{1}_m = [1, \dots, 1] \in \mathbb{R}^m$ and $\mathbf{1}_n = [1, \dots, 1] \in \mathbb{R}^n$)

$$\min_{\mathbf{w} \in \mathbb{R}^d, b_1, b_2 \in \mathbb{R}} \{\|X\mathbf{w} + b_1 \cdot \mathbf{1}_m - \mathbf{u}\|_2^2 + \|Z\mathbf{w} + b_2 \cdot \mathbf{1}_n - \mathbf{v}\|_2^2 + \|\mathbf{w}\|_2^2\} \quad \text{such that} \quad b_2 = 2 \cdot b_1$$

Part 1: Melbo wants to use alternating optimization to solve the problem. Given fixed values of b_1 and b_2 s.t. $b_2 = 2 \cdot b_1$, solve $\arg\min_{\mathbf{w} \in \mathbb{R}^d} \{\|X\mathbf{w} + b_1 \cdot \mathbf{1}_m - \mathbf{u}\|_2^2 + \|Z\mathbf{w} + b_2 \cdot \mathbf{1}_n - \mathbf{v}\|_2^2 + \|\mathbf{w}\|_2^2\}$ to get a value for $\mathbf{w}$. Give brief derivation. *Note: optimization is only over $\mathbf{w}$ in this part.* **(4 marks)**

Stationarity tells us that at the optimum, we must have

$$X^\top(X\mathbf{w} + b_1 \cdot \mathbf{1}_m - \mathbf{u}) + Z^\top(Z\mathbf{w} + b_2 \cdot \mathbf{1}_n - \mathbf{v}) + \mathbf{w} = 0$$

Another way of writing this is the following (I_d is the d-dim identity matrix)

$$(X^\top X + Z^\top Z + I_d)\mathbf{w} = X^\top(\mathbf{u} - b_1 \cdot \mathbf{1}_m) + Z^\top(\mathbf{v} - b_2 \cdot \mathbf{1}_n)$$

This gives us the following value for the model

$$\mathbf{w} = (X^\top X + Z^\top Z + I_d)^{-1}\left(X^\top(\mathbf{u} - b_1 \cdot \mathbf{1}_m) + Z^\top(\mathbf{v} - b_2 \cdot \mathbf{1}_n)\right)$$

Note that the matrix $X^\top X + Z^\top Z + I_d$ is always invertible due to the identity matrix being added.

Part 2: Now instead, if we are given a fixed value of $\mathbf{w} \in \mathbb{R}^d$, find out values for b_1, b_2 by solving $\arg\min_{b_1, b_2 \in \mathbb{R}} \{\|X\mathbf{w} + b_1 \cdot \mathbf{1}_m - \mathbf{u}\|_2^2 + \|Z\mathbf{w} + b_2 \cdot \mathbf{1}_n - \mathbf{v}\|_2^2\}$ subject to the constraint $b_2 = 2 \cdot b_1$. Give brief derivation. *Note: optimization is only over b_1 and b_2 in this part.* **(6 marks)**

Let us introduce a new variable t and replace $b_1 = t, b_2 = 2t$ to get

$$\arg\min_{t \in \mathbb{R}} \{\|X\mathbf{w} + t \cdot \mathbf{1}_m - \mathbf{u}\|_2^2 + \|Z\mathbf{w} + 2t \cdot \mathbf{1}_n - \mathbf{v}\|_2^2\}$$

Stationarity tells us that the optimal value of t satisfies

$$(X\mathbf{w} + t \cdot \mathbf{1}_m - \mathbf{u})^\top \mathbf{1}_m + (Z\mathbf{w} + 2t \cdot \mathbf{1}_n - \mathbf{v})^\top (2 \cdot \mathbf{1}_n) = 0$$

This means that $(m + 4n) \cdot t = (\mathbf{u} - X\mathbf{w})^\top \mathbf{1}_m + (\mathbf{v} - Z\mathbf{w})^\top (2 \cdot \mathbf{1}_n)$. This gives us

$$t = \frac{1}{m + 4n}\left((\mathbf{u} - X\mathbf{w})^\top \mathbf{1}_m + (\mathbf{v} - Z\mathbf{w})^\top (2 \cdot \mathbf{1}_n)\right)$$
$$= \frac{1}{m + 4n}\left(\sum_{i \in [m]} (u_i - \mathbf{w}^\top \mathbf{x}_i) + 2\sum_{i \in [n]} (v_i - \mathbf{w}^\top \mathbf{z}_i)\right)$$

This gives us $b_1 = \frac{1}{m+4n}\left(\sum_{i \in [m]} (u_i - \mathbf{w}^\top \mathbf{x}_i) + 2\sum_{i \in [n]} (v_i - \mathbf{w}^\top \mathbf{z}_i)\right)$

Similarly, we get $b_2 = \frac{2}{m+4n}\left(\sum_{i \in [m]} (u_i - \mathbf{w}^\top \mathbf{x}_i) + 2\sum_{i \in [n]} (v_i - \mathbf{w}^\top \mathbf{z}_i)\right)$