$$K(x, x^{(i)}) = e^{-\frac{\|x - x^{(i)}\|^2}{2}}$$

**Midsem Solution 1**

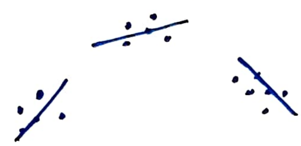
(a) for $d = 1$ $\quad x \in \mathbb{R}$

$$K(x, x^{(i)}) = e^{-\frac{(x - x^{(i)})^2}{2}}$$

$$J_{new}(w|x) = \frac{1}{2n} \sum e^{-\frac{(x - x^{(i)})^2}{2}} (y^{(i)} - x^{(i)T} w)^2$$

term large for $x^{(i)}$ near ~~x~~ $x$

↳ so essentially the model fits local linear hyperplanes because the loss is mostly contributed by points $x^{(i)}$ ~~x for~~ that are near point $x$ for any given $x$.



(b)
(c) $\quad J_{new}(w|x)$

$$(y^{(i)} - x^{(i)T} w)^T$$

$$[y^{(1)} - x^{(1)T} w, \quad y^{(2)} - x^{(2)T} w, \quad y^{(3)} - x^{(3)T} w]$$

$$\underbrace{[y^{(1)} - x^{(1)T} w \quad y^{(2)} - x^{(2)T} w \quad y^{(3)} - x^{(3)T} w]}_{1 \times 3} \underbrace{\begin{bmatrix} K(x, x^{(1)}) & 0 & 0 \\ 0 & K(x, x^{(2)}) & 0 \\ 0 & 0 & K(x, x^{(3)}) \end{bmatrix}}_{3 \times 3} \underbrace{\begin{bmatrix} y^{(1)} - x^{(1)T} w \\ y^{(2)} - x^{(2)T} w \\ y^{(3)} - x^{(3)T} w \end{bmatrix}}_{3 \times 1}$$

↑ Example of 3 data points.

$K$

$\boxed{[K]_{n \times n}}$ ← (b)

(c) →
$$\boxed{K_{ij} = \begin{cases} K(x, x^{(i)}) & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}}$$

$$J = \frac{1}{2n} \sum_{i=1}^{n} K(x, x^{(i)}) \, (y^{(i)} - x^{(i)T} w)^2$$

(d)

$$= \frac{1}{2n} \sum_{i=1}^{n} e^{\frac{-\|x - x^{(i)}\|^2}{2}} (y^{(i)} - x^{(i)T} w)^2$$

this can be written as

$$\frac{1}{2n} \underbrace{(Y - Xw)^T}_{1\times n} \underbrace{K}_{n\times n} \underbrace{(Y - XW)}_{n\times 1}$$

$$\frac{1}{2n} (Y^T - W^T X^T)(KY - KXW)$$

For loss we want

$$\nabla_W (Y^T K Y - W^T X^T K Y - Y^T K X W + W^T X^T K X W) = 0$$

$X_{n\times d}$ $Y_{n\times 1}$ $W_{d\times 1}$ $K_{n\times n}$

$$\nabla(Y^T K Y) = 0$$

$$\nabla(W^T X^T K Y) = X^T K Y$$

$$\nabla(Y^T K X W) = X^T K^T Y = X^T K Y \quad \text{as} \quad K = K^T$$

$$\nabla(W^T X^T K X W) = 2X^T K X W$$

So derivative become

$$-X^T K Y - X^T K^T Y + 2X^T K X W$$

$\nabla A w = A^T$

$\nabla w^T B = B$

$$-2X^T K Y + 2X^T K X W$$

Put to 0

$$X^T K (Y - XW) = 0$$

$$X^T K X W = X^T K Y$$

$$\boxed{W^* = (X^T K X)^{-1} X^T K Y}$$

## Q2

**P1. (I)**
Multiple possible answers for this question. One possible answer is as follows:

1. Assume there is a training split $\mathcal{U}$ and a validation split $\mathcal{V}$ .

2. Learn the label distribution from $\mathcal{U}$ . This means that we know the quantiles and threshold of all labels. For example:

- label: $0 \to q_0 : 0.15 \quad t_0 : -0.2$
- label: $1 \to q_1 : 0.41 \quad t_1 : 1.28$
- label: $2 \to q_2 : 0.88 \quad t_2 : 3.41$
- label: $3 \to q_3 : 1.00 \quad t_3 : 4.48$

We can use either of quantiles or thresholds to find the predictions. We will use thresholds for $\mathcal{U}$ and quantiles for $\mathcal{V}$ as follows:
For $\mathcal{U}$ :

$$\hat{y} = t_1(m_\theta(x)) = \begin{cases} 0 & \text{if} \quad m_\theta(x) \leq t_0, \\ 1 & \text{if} \quad t_0 \geq m_\theta(x) \leq t_1, \\ 2 & \text{if} \quad t_1 \geq m_\theta(x) \leq t_2, \\ 3 & \text{if} \quad t_2 \geq m_\theta(x) \end{cases}$$

Next, assume $q(x)$ gives the quantile of x from a batch $X$. For $\mathcal{V}$ :

$$\hat{y} = t_2(q(m_\theta(x))) = \begin{cases} 0 & \text{if} \quad q(m_\theta(x)) \leq q_0, \\ 1 & \text{if} \quad q_0 \geq q(m_\theta(x)) \leq q_1, \\ 2 & \text{if} \quad q_1 \geq q(m_\theta(x)) \leq q_2, \\ 3 & \text{if} \quad q_2 \geq q(m_\theta(x)) \end{cases}$$

The reason for using quantiles for $\mathcal{V}$ is that an example for $\mathcal{V}$ might come from out of distribution.

3. Accuracy can be calculated on the thresholded predictions and MSE loss on $m_\theta(x)$ can be used for training.

**P2. (II)**
Again, multiple possible formulations. The pointwise formulation is as follows:

$$\min_\theta \left[ \sum_{i \in S} \mathcal{L}(h(x_i), y_i) + \sum_{j \in \mathcal{D} \setminus S} \mathcal{L}(m_\theta(x_j), y_j) \right]$$

**P3. (III)**
Use the formula and the given data to evaulate the 6 terms of the ranking loss. Some mistakes that students made were the following:

1. Made a mistake in understanding the sign operator.
2. Applied the sign operator to the product $(r_{ii'}(\hat{y}_i - \hat{y}_{i'}))$
3. Forgot applying relu at the end $[\ldots]_+$

partial marks were awarded regardless.

**P3. (IV)**

1. Ranking loss evaluates the relative ranks of the predictions. If $y_i > y_{i'}$ and if $\hat{y}_i < \hat{y}_{i'}$, then the ranking loss will penalize such a pair of predictions.
2. Margin $\Delta$ ensures the margin of separation. If $y_i > y_{i'}$, the penalty is zero only if $\hat{y}_i - \hat{y}_{i'} \geq \Delta$.

# Q3

**Error 1:** model.fc2 should have the structure (50, 30).
**Error 2:** model outputs 10 logits, while the dataset had 4 classes.

Both are structural errors.