Questions

You may also write "linear classifiers" model

$\{f \mid \exists \omega \in \mathbb{R}^n \ni f(x) = \text{sign}(\omega^T\varphi(x)) \; \forall x \in \mathcal{X}\}$ is also OK if you write 'linear classifiers model'

1. Consider the logistic regression setting taught in the lecture. Here, the learning task is a binary classification problem [[*Fill this blank with either "regression" or "multi-class classification" or "binary classification". 1/4 mark*]]. Let $\phi : \mathcal{X} \mapsto \mathbb{R}^n$ be a given feature map. The model employed in logistic regression is the linear model [[*Fill this blank with the appropriate proper noun. 1/4 mark*]]. The mathematical definition of this model is given by the expression:

$$\mathcal{L}_{n,\varphi} \equiv \{f \mid \exists \omega \in \mathbb{R}^n \ni f(x) = \omega^T\varphi(x) \; \forall x \in \mathcal{X}\}.$$

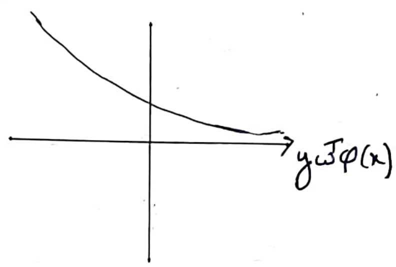
[[*In the above blank, use 'w' to denote the parameter of this model. 1/2 mark*]]

The mathematical expression for the loss function used here is:

$$l(w, x, y) \equiv \log\left(1 + e^{-y\omega^T\varphi(x)}\right).$$

[1 mark]

This loss function can be visualized using the plot below:



[[*Fill the plot appropriately to roughly depict the graph of logistic loss. Clearly label the axes, without which no marks will be given. 1/2 mark*]].

If p^* is the underlying (unknown) likelihood relating the inputs and labels, then, the Bayes optimal, restricted to the functions in the model, is given by the mathematical expression:

$$f^* = \underset{\omega \in \mathbb{R}^n}{\text{argmin}} \; \mathbb{E}_{X,Y \sim p^*}\left[\log\left(1 + e^{-Y\omega^T\varphi(X)}\right)\right]$$

[[*No marks will be given if you write general expressions for Bayes optimal. You need to write the specific expression for the logistic regression set-up. 1 mark*]]. Let the training data be $\mathcal{D} \equiv \{(x_1, y_1), \ldots, (x_m, y_m)\}$. The name of the important assumption that relates p^* to $\mathcal{D}$ is Supervised Learning. [[*Fill this blank with the appropriate proper noun. 1/2 mark*]]. This assumption formally means the following:

Any other equivalent sentence is fine.

$\mathcal{D}$ is set of iid samples from p^*

[1/2 mark]

The ERM problem in this case is the following mathematical optimization problem:

argmin is also fine.

$$\min_{w \in \mathbb{R}^n} \frac{1}{m} \sum_{i=1}^{m} \log\left(1 + e^{-y_i w^T \varphi(x_i)}\right)$$

[1/2 mark]

If the ERM solution is denoted by $\hat{w}_m$, then the label for any $x \in \mathcal{X}$ shall be computed using the formula: $\text{sign}(\hat{w}_m^T \varphi(x))$.

[1/2 mark]

Now, say, the training data actually is

$$\mathcal{D} = \left\{ \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, 1 \right), \left(\begin{bmatrix} 1 \\ 0 \end{bmatrix}, -1 \right), \left(\begin{bmatrix} 0 \\ 1 \end{bmatrix}, -1 \right), \left(\begin{bmatrix} 1 \\ 1 \end{bmatrix}, 1 \right) \right\}$$

and the feature map ϕ is defined by

$$\phi\left(\begin{bmatrix} z_1 \\ z_2 \end{bmatrix} \right) \equiv ((z_1 + z_2)\%2) - \frac{1}{2} \; \forall \; z_1, z_2 \in \mathbb{R}.$$

Here, $a\%b$ is the remainder when a is divided by b. For this case, in the box below, write down the optimal solution[1] of the ERM problem along with justification:

[1]You are welcome to solve this optimization problem in any way you prefer. For e.g., analytically, manually iterating through gradient descent etc.

Equivalent rephrasing is fine. ← $\boxed{\text{No Solution. Because } \mathcal{D} \text{ is linearly separable in feature space.}}$

[1 mark]

For this specific training data and feature map, suppose we wish to perform linear classification using the 0-1 loss. Then, run the perceptron algorithm in rough using manual calculations and write all parameter iterates until convergence including initialization in the box below. For each iterate write down the update equation too. No other details are required

$$w^{(0)} = 0$$
$$w^{(1)} = 0 + (-1)\left(\tfrac{1}{2}\right) = -\tfrac{1}{2}.$$

$\rightarrow w^{(1)} = 0 + (1)\left(-\tfrac{1}{2}\right) = -\tfrac{1}{2}$ is also correct.

[1 mark]

2. Consider the linear regression setting taught in lectures with training data as: $\mathcal{D} = \{(2,1),(4,5)\}$ (usual convention of set of input,label pairs). Consider the feature map $\phi(x) = x$. Analytically solve the ERM problem in rough work and write down the final ERM solution in this blank: $w_\phi^{ERM} =$ <u>1.1</u> [[*1/2 mark*]]. With this solution, the explained variance computed on the training set is <u>0.775</u> [[*Fill the blank with appropriate number. 1/2 mark*]]. Now, consider another feature map, $\psi(x) \equiv \begin{bmatrix} x \\ 1 \end{bmatrix}$. With this feature map, analytically solve the ERM problem in rough work and write down the final ERM solution in this blank: $w_\psi^{ERM} = \begin{bmatrix} 2 \\ -3 \end{bmatrix}$ [[*1 mark*]]. With this solution, the explained variance computed on the training set is <u>1</u> (one) [[*Fill the blank with appropriate number. 1/4 mark*]]. Now, consider another feature map, $\nu(x) \equiv \begin{bmatrix} x \\ 1 \\ -1 \end{bmatrix}$. With this feature map, analytically solve the ERM problem in rough work and write down the final ERM solution you obtained in this blank: $w_\nu^{ERM} = \begin{bmatrix} 2 \\ -3 \\ 0 \end{bmatrix}$ → Any other where $w_2 - w_3 = -3$ is also fine. For e.g. $\begin{bmatrix} 2 \\ 45 \\ 48 \end{bmatrix}$

[[*1 mark*]]. With this solution, the explained variance computed on the training set is 1 [[*Fill the blank with appropriate number. 1/4 mark*]].
(one)

3. In the lectures you were taught how to model the Bayes optimal in a binary classification task using linear functions (over input feature space). Now suppose you have a multi-class classification problem with 3 classes: 'Ξ', '‡', and '¶'. However, still you are only allowed to use the linear model taught in lectures. Think about how you can model the Bayes optimal in a 3-class classification task using these linear functions. With this way of modelling the Bayes optimal in mind, according to you, the loss function, l, appropriate for this task would be defined by $l(w, x, \Xi) \equiv$ $\mathbb{1}\{w^T\varphi(x) > -1\}$, $l(w, x, ‡) \equiv$ $\mathbb{1}\{w^T\varphi(x) \notin (-1,1)\}$, $l(w, x, ¶) \equiv$ $\mathbb{1}\{w^T\varphi(x) < 1\}$. Here, w denotes the parameter of the linear model.

Any other tie break is also fine

Any thresholds instead of 1, -1 are also OK.

Any cyclic permutation of these is also OK

[1.5 Marks]

"logistic" & "hinge" versions are also fine.

Observe that your way of modelling the Bayes optimal with linear functions has an inherent ('wrong'?) bias. More specifically, if the parameter changes a little then the label for a fixed x changes preferentially to one of the other two classes. In this sense, there is am implicit (unequal) nearness between different class pairs.

Now, suppose you are allowed to model functions of the form $f(x) = W^\top\phi(x)$, where W is $n \times 3$, where ϕ is a feature map. You may use the notation $W = [w_1\ w_2\ w_3]$, where $w_i \in \mathbb{R}^n$. Think about how you can model the Bayes optimal in a 3-class classification task using these '3-dimensional linear functions'. With this way of modelling the Bayes optimal in mind, according to you, the loss function, l, appropriate for this task would be defined by $l(W, x, \Xi) \equiv$ $\mathbb{1}\{w_1^T\varphi(x) < w_2^T\varphi(x)\} + \mathbb{1}\{w_1^T\varphi(x) < w_3^T\varphi(x)\}$ $l(W, x, ‡) \equiv$ $\mathbb{1}\{w_2^T\varphi(x) < w_1^T\varphi(x)\} + \mathbb{1}\{w_2^T\varphi(x) < w_3^T\varphi(x)\}$ $l(W, x, ¶) \equiv$ $\mathbb{1}\{w_3^T\varphi(x) < w_1^T\varphi(x)\} + \mathbb{1}\{w_3^T\varphi(x) < w_2^T\varphi(x)\}$.

[2.5 Marks]

Many alternatives (see above part).

ROUGH

① Perceptron question.

At iteration k

$w^{(k-1)}$ if ($y_i w^T \varphi(x_i) \leq 0$ for some i)

then, $w^{(k)} = w^{(k-1)} + y_i \varphi(x_i)$

$w^{(1)}$

$w^{(0)} = 0$

$\varphi(x_1) = \varphi(x_4) = -\frac{1}{2}$

$\varphi(x_3) = \varphi(x_2) = \frac{1}{2}$.

$y_1 w^{(0)T} \varphi(x_1) = y_4 w^{(0)T} \varphi(x_4) = 0 \leq 0$

$y_3 w^{(0)T} \varphi(x_3) = y_2 w^{(0)T} \varphi(x_2) = 0 \leq 0$

We can choose any i = 1 to 4.

Let's choose i = 3. $w^{(1)} = 0 + (-1)\frac{1}{2} = -\frac{1}{2}$

---

② $w_\varphi^{ERM} = \underset{w \in \mathbb{R}}{\operatorname{argmin}} \frac{1}{2}\left[(w(2)-1)^2 + (w(4)-5)^2\right]$

$= \underset{w \in \mathbb{R}}{\operatorname{argmin}} \quad 10w^2 - 22w = \frac{22}{20} = 1.1$

$MSE(w_\varphi^{ERM}) = \frac{1}{2}\left[(1.1\times2-1)^2 + (1.1\times4-5)^2\right]$

$= \frac{1.44}{2} + \frac{0.36}{2} = 0.9$

$MSE(\bar{w}) = \frac{1}{2}\left[(3-1)^2 + (3-5)^2\right] = 4$

$\bar{w}^T\varphi(x) = \frac{5+1}{2} = 3$

$\forall x$

$\therefore$ exp. val $= 1 - \frac{0.9}{4} = 1 - 0.225 = 0.775$

(2) $$w_\psi^{ERM} \equiv \underset{w_1\in\mathbb{R}, w_2\in\mathbb{R}}{\operatorname{argmin}} \underbrace{\frac{1}{2}\left[(2w_1+w_2-1)^2+(4w_1+w_2-5)^2\right]}_{g(w)}$$

$$\frac{\partial g(w)}{\partial w_1} = 2(2w_1+w_2-1)+4(4w_1+w_2-5)=0 \iff 20w_1+6w_2-22=0 \iff w_2 = \frac{22}{6} - \frac{20w_1}{6}$$

$$\frac{\partial g(w)}{\partial w_2} = (2w_1+w_2-1)+(4w_1+w_2-5)=0 \iff 6w_1+2w_2-6=0$$

$$\Updownarrow$$

$$18w_1+(22-20w_1)-18=0$$

$$\iff w_1 = 2$$

$$\Downarrow$$

$$w_2 = \frac{-18}{6} = -3$$

$$MSE(w_\psi^{ERM}) = \frac{1}{2}\left[(2\times2-3-1)^2+(4\times2-3-5)^2\right]$$

$$=0 \iff$$ expr. variance = 1

(2) $$w_{\gamma f}^{ERM} \equiv \underset{w_1\in\mathbb{R}, w_2\in\mathbb{R}, w_3\in\mathbb{R}}{\operatorname{argmin}} \frac{1}{2}\left[(2w_1+w_2-w_3-1)^2+(4w_1+w_2-w_3-5)^2\right]$$

equations will be same as above except $w_2 \rightarrow w_2-w_3$.

$\therefore$ equations will be same as above except the optimality conditions.

$\therefore$ $w_1=2$, $w_2-w_3=-3$ are optimality conditions.

Any choice of w that satisfies is fine.

My choice is $\begin{bmatrix} 2 \\ -3 \\ 0 \end{bmatrix}$.

expr. variance will remain 1.

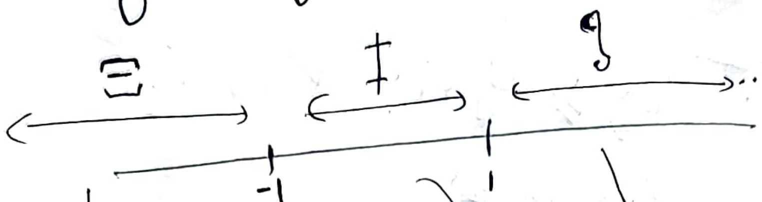
③ I part

I'll ~~map~~ bin reals into three bins instead of two ~~fo~~

Two bins → require one threshold

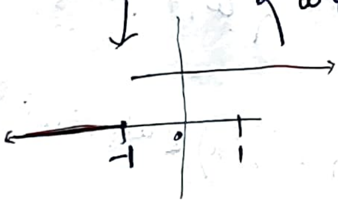
Three bins → require two thresholds.

My choice for thresholds is $-1, 1$

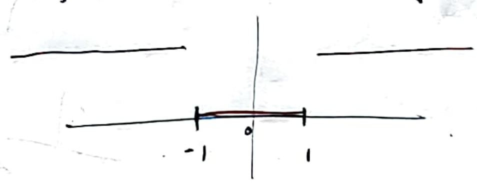


$$l(\omega, x, \varsigma) \equiv \mathbb{1}_{\{\omega^T\varphi(x) < 1\}}$$

$$l(\omega, x, \Xi) \equiv \mathbb{1}_{\{\omega^T\varphi(x) > -1\}}$$



$$l(\omega, x, \mp) = \mathbb{1}_{\{\omega^T\varphi(x) \geq 1\}} + \mathbb{1}_{\{\omega^T\varphi(x) \leq -1\}}$$



$$= \mathbb{1}_{\{\omega^T\varphi(x) \geq 1 \,\&\, \omega^T\varphi(x) \leq -1\}}.$$

$$= \mathbb{1}_{\{\omega^T\varphi(x) \notin (-1, 1)\}}$$

③ II part

We may say high $w_i^T\varphi(x) \Longleftrightarrow i^{th}$ class.

~~But how to~~ Let threshold be "h".

then $w_i^T\varphi(x) \geq h$, $w_j^T\varphi(x) < h \Rightarrow i^{th}$ class
$(j \neq i)$

This is same as $w_i^T\varphi(x) \geq w_j^T\varphi(x) \quad \forall j \neq i$.

$\downarrow$ loss for 1^{st} class

$$\mathbb{1}\{w_1^T\varphi(x) < w_2^T\varphi(x)\} + \mathbb{1}\{w_1^T\varphi(x) < w_3^T\varphi(x)\}$$

$$= \mathbb{1}\{w_1^T\varphi(x) < w_2^T\varphi(x) \text{ or } w_1^T\varphi(x) < w_3^T\varphi(x)\}$$

$$= \mathbb{1}\{w_1^T\varphi(x) < \min_{j=2,3} w_j^T\varphi(x)\}.$$