ESE 542 Final Exam

Problem 1

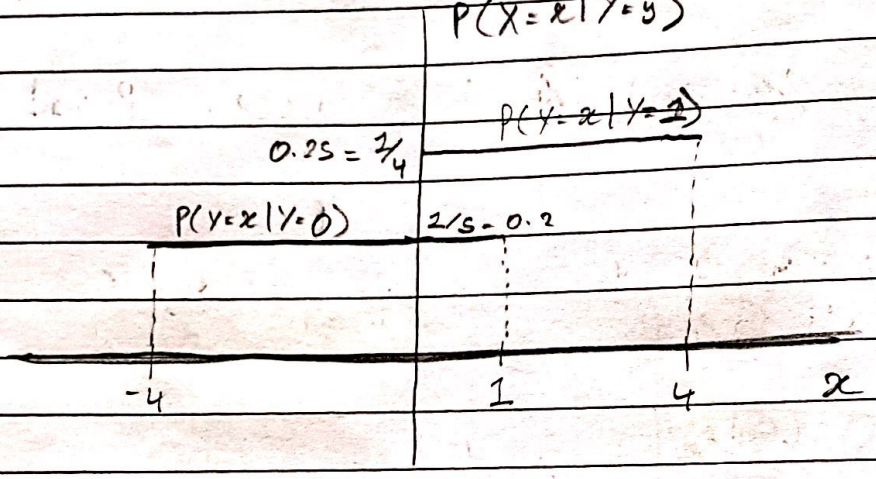
$P(Y=0) = P$

$P(Y=1) = 1-P$

$P(X=x \mid Y=0) = 1/5$ for $-4 \le x \le 1$

$P(X=x \mid X=1) = 1/4$ for $0 \le x \le 4$

a)



b) By Bayes' rule,

$$P(Y=y \mid X=x) = \frac{P(X=x \mid Y=y) \cdot P(Y=y)}{P(X=x)}$$

c) $h^*(x) = \underset{y \in \{0,1\}}{\operatorname{argmax}} P(Y=y \mid X=x)$

$= \underset{y \in \{0,1\}}{\operatorname{argmax}} \frac{P(X=x \mid Y=y) \cdot P(Y=y)}{P(X=x)}$ (bayes' Rule)

$= \underset{y \in \{0,1\}}{\operatorname{argmax}} P(X=x \mid y=y) \cdot P(Y=y)$ (Since $P(Y=x)$ doesn't depend on y)

$= \underset{y \in \{0,1\}}{\operatorname{argmax}} \{ P(X=x \mid Y=0) \cdot P(y=0), P(Y=x \mid Y=1) \cdot P(Y=1) \}$

$= \underset{y \in \{0,1\}}{\operatorname{argmax}} \left\{ \frac{1}{5} \cdot P, \frac{1}{4} \cdot (1-P) \right\}$

$= \begin{cases} 0 \text{ if } \frac{P}{5} > \frac{1-P}{4} \\ 1 \text{ otherwise} \end{cases}$

Let's look @ the condition for 0:

$\frac{P}{5} > \frac{1-P}{4}$

$\Rightarrow 4P > 5(1-P)$

$\Rightarrow 4P > 5 - 5P$

$\Rightarrow 4P + 5P > 5$

$\Rightarrow 9P > 5$

$\Rightarrow P > 5/9$

$\therefore h^*(x) = \begin{cases} 0 \text{ if } P > 5/9 \\ 1 \text{ if } P \leq 5/9 \end{cases}$

This makes intuitive sense because we simply return the argument for y that maximizes $P(X=x \mid Y=y)$.

Note, $5/9 > 1/2$, which means that if on average $Y=0$ is more likely, always output 0 irrespective of x.

a) $P(h^*(x) \neq y)$

$= P(h^*(x) = 0, y = 1) + P(h^*(x) = 1, y = 0)$

$= P(h^*(x) = 0 \mid y = 1) \cdot P(y = 1) + P(h^*(x) = 1 \mid y = 0) \cdot P(y = 0)$

$= P(p > s/a) \times 1 - p + P(p \leq s/a) \times p$

$= \mathbb{1}(p > s/a) \cdot (1 - p) + \mathbb{1}(p \leq s/a)\, p$

Therefore, the error rate is either p or 1-p, depending on whether p is less than or greater than s/a respectively.

This makes intuitive sense because if we always classify 0, then we make an error whenever the label is 1, and that happens with $P(Y = 1) = 1 - p$.

On the other hand, if we always classify as 1, then we make an error whenever the label is 0, & that happens w/ $P(Y = 0) = p$.

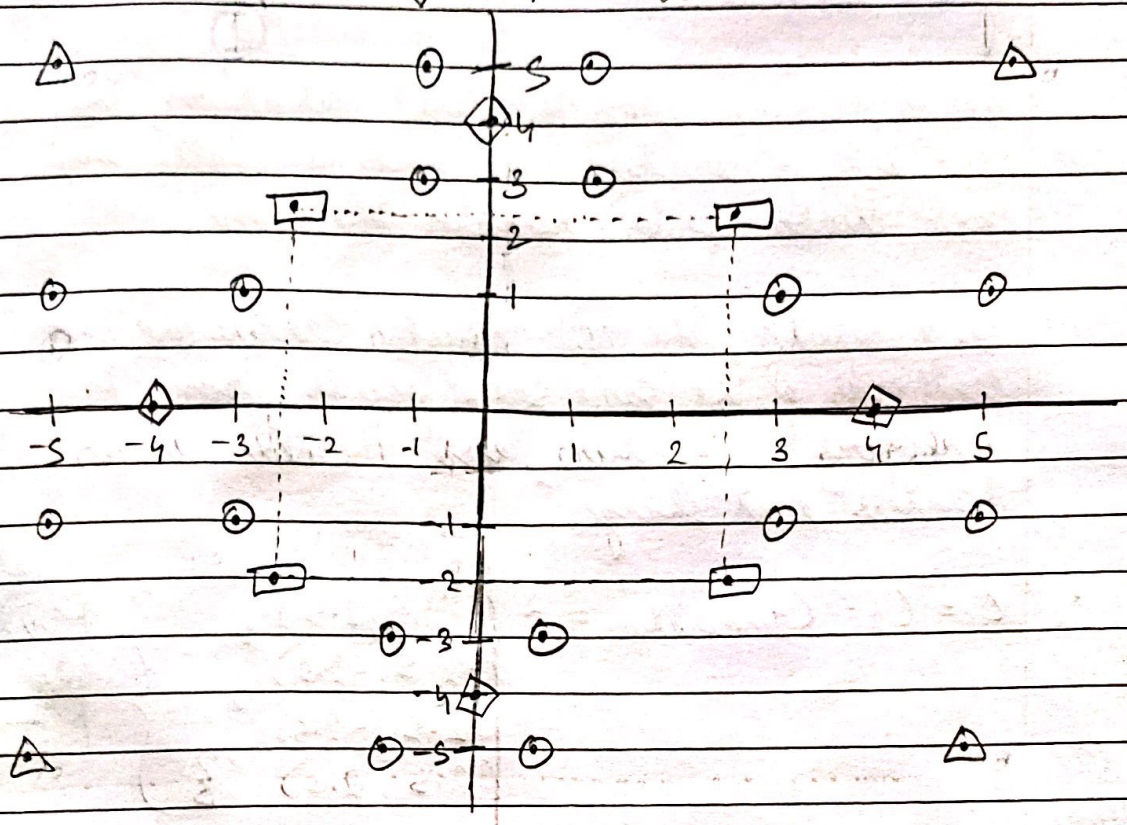
In order to activate only one of p or 1-p as errors, they are multiplied with the indicator fn of their corresponding classifier conditions.

Camlin Page
Date / /

Problem 2

Legend
□ = Updated cluster center ○ = Data point
◇ = Optimal cluster center △₀ = Initial cluster center

a)



b) Points closest to (-5, 5) → {(-5, 1), (-3, 1), (-1, 3), (-1, 5)}
" " " (-5, -5) → {(-5, -1), (-3, -1), (-1, -3), (-1, -5)}
" " " (5, -5) → {(5, -1), (3, -1), (1, -3), (1, -5)}
" " " (5, 5) → {(5, 1), (3, 1), (1, 3), (1, 5)}

Computing new center for first cluster = $\left(\frac{-5-3-1-1}{4}, \frac{1+1+3+5}{4}\right)$

$= (-10/4, 10/4) = (-2.5, 2.5)$

Similarly New center for second cluster = (-2.5, -2.5)
" " " " third " = (2.5, -2.5)
" " " " fourth " = (2.5, 2.5)

c) Final output of clusters remains the same as those after termination

This is because after the first iteration, the new sets of clusters (ie points associated with each cluster center) remain the same.

As a result, the new cluster centers are the centroids of the same set of points from the previous iteration, and the new cluster centers don't change.

$\therefore C_1 = C_2 = C_{Termination} = \{(-2.5, 2.5), (-2.5, -2.5), (2.5, -2.5), (2.5, 2.5)\}$ } Marked on plot with □

d) The optimal cluster centers should be

$C_{OPT} = \{(-4, 0), (0, -4), (4, 0), (0, 4)\}$ } Marked on plot with ◇

Therefore in this instance, k-means does not find the optimal clusters

Camlin | Page
Date / /

Problem 3

$S = \{(x_i, y_i)\}_{i=1}^{n}$

$x_i \in \mathbb{R}^d$ $\quad y_i = \beta^T x_i + \epsilon_i$

$y_i \in \mathbb{R}$ $\quad \beta = (\beta_1, \ldots, \beta_d)$

This notation is used later

a) ϵ_i iid Laplace $(0,1) = \frac{1}{2} e^{-|z|} = l(z)$

Show: $f(y_1 \ldots y_n | \beta) = \left(\frac{1}{2}\right)^n \prod_{i=1}^{n} e^{-|y_i - \beta^T x_i|}$

Now, we know that

$\therefore f(y_1 \ldots y_n | \beta) = \prod_{i=1}^{n} f(y_i | \beta)$ (Since $\epsilon_i \sim$ iid & features also iid)

~~In class (& in Piazza @ 804), we saw that~~

Now, $f(y_i | \beta) = P(Y_i = y_i | \beta)$

$= P(y_i = \beta^T x_i + \epsilon_i | \beta)$

$= P(\epsilon_i = y_i - \beta^T x_i | \beta)$

$= l(y_i - \beta^T x_i)$ (Since ϵ_i is Laplace $(0,1)$ distributed).

$\therefore f(y_1 \ldots y_n | \beta) = \prod_{i=1}^{n} f(y_i | \beta)$

$= \prod_{i=1}^{n} l(y_i - \beta^T x_i)$

$= \prod_{i=1}^{n} \left(\frac{1}{2}\right) e^{-|y_i - \beta^T x_i|}$

$= \left(\frac{1}{2}\right)^n \prod_{i=1}^{n} e^{-|y_i - \beta^T x_i|}$

Since $\frac{1}{2}$ does not depend on index i

b) We have now established that

$$f(y_1, \ldots, y_n \mid \beta) = L(\beta) = \left(\frac{1}{2}\right)^n \prod_{i=1}^{n} e^{-|y_i - \beta^T x_i|}$$

$$\therefore \hat{\beta}_{MLE} = \underset{\beta}{\operatorname{argmax}}\, f(y_1, \ldots, y_n \mid \beta)$$

$$= \underset{\beta}{\operatorname{argmax}}\, L(\beta)$$

$$= \underset{\beta}{\operatorname{argmax}}\, \log L(\beta)$$ (due to monotonicity of log transform)

$$= \underset{\beta}{\operatorname{argmax}}\, \log \left\{ \left(\frac{1}{2}\right)^n \cdot \prod_{i=1}^{n} e^{-|y_i - \beta^T x_i|} \right\}$$

$$= \underset{\beta}{\operatorname{argmax}} \left\{ \underbrace{\log\left[\left(\frac{1}{2}\right)^n\right]}_{\text{doesn't depend on } \beta} + \log\left[\prod_{i=1}^{n} e^{-|y_i - \beta^T x_i|}\right] \right\}$$

$$= \underset{\beta}{\operatorname{argmax}} \left\{ \log \prod_{i=1}^{n} e^{-|y_i - \beta^T x_i|} \right\}$$

$$= \underset{\beta}{\operatorname{argmax}} \left\{ \sum_{i=1}^{n} \log e^{-|y_i - \beta^T x_i|} \right\}$$ (Since $\log e^{-x} = -x$)

$$= \underset{\beta}{\operatorname{argmax}} \left\{ \sum_{i=1}^{n} -|y_i - \beta^T x_i| \right\}$$

$$= \underset{\beta}{\operatorname{argmax}} \left\{ -\sum_{i=1}^{n} |y_i - \beta^T x_i| \right\}$$ (Since $\underset{\beta}{\operatorname{argmax}}\{-f(\beta)\} = \underset{\beta}{\operatorname{argmin}}\{f(\beta)\}$)

$$= \underset{\beta}{\operatorname{argmin}} \left\{ |y_i - \beta^T x_i| \right\}$$ Hence Proved

Camlin | Page
Date / /

Problem 4

a) We have here the union of 3 disjoint intervals

We must show that $\{1, 2, 3, 4, 5, 6\}$ is shattered

Suppose the labellings are $(c_1, c_2, c_3, c_4, c_5, c_6)$ where c_i is the labelling of i

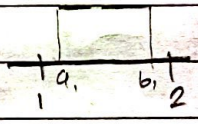
We must show that all 2^6 labellings are possible

Let us associate $[a_1, b_1] \rightarrow c_1, c_2$ for 1, 2
$[a_2, b_2] \rightarrow c_3, c_4$ for 3, 4
$[a_3, b_3] \rightarrow c_5, c_6$ for 5, 6

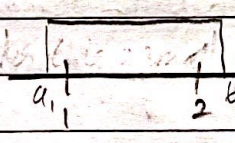
WLOG, if one interval can produce all possible 2^2 labellings for 2 points, then the union of 3 disjoint intervals can generate $(2^2)^3 = 2^6$ labellings

WLOG, let's take $[a_1, b_1]$ & try to shatter $\{1, 2\}$
Let ϵ be a small +ve value that is $0 < \epsilon \ll 1$

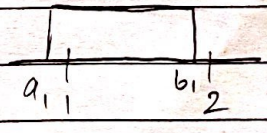
Labelling: 00 → if $a_1 = 1 + \epsilon$
$b_1 = 2 - \epsilon + 1$



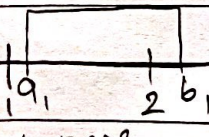
Labelling: 11 → if $a_1 = 1 - \epsilon$
$b_1 = 2 + \epsilon$



Labelling: 10 → if $a_1 = 1 - \epsilon$
$b_1 = 2 - \epsilon$



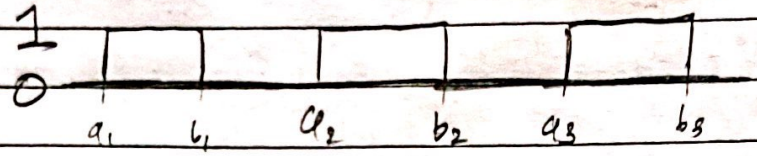
Labelling: 01 → if $a_1 = 1 + \epsilon$
$b_1 = 2 + \epsilon$



∴ Since 2^2 labellings are possible for 1 interval $(2^2)^3$ is possible for 3
∴ $\{1, 2, 3, 4, 5, 6\}$ is shattered

b) Suppose we have some sample of size 7, say $\{1,2,3,4,5,6,7\}$

We know that the graph of the hypothesis is as follows:



& There are 3 rising edges & 3 dipping edges.

For a sample of 7 points, consider the labelling

1 0 1 0 1 0 1

This particular labelling has 4 separate rises, & we would need 4 disjoint intervals to generate this.

However, this function only gives us 3 jumps but we need 4 to generate one of the labellings, specifically 1010101

∴ A set of size 7 may not be shattered

This is because the function class of union of 3 disjoint intervals is not sufficiently complex.