CS973: Machine Learning for Cyber Security

Endsem (30 Nov 2024)

Q3. The Essence of Exfiltration **(8 marks)**

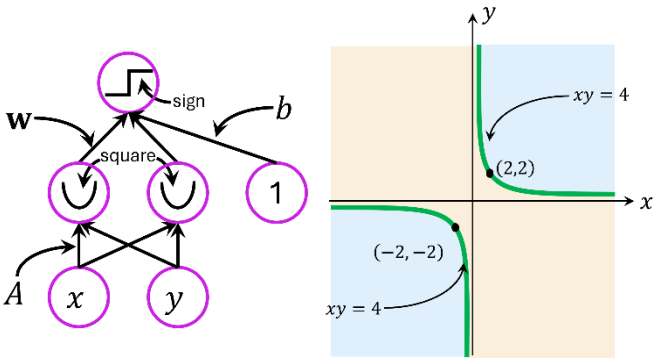
Melba is doing a side-channel exfiltration experiment. It is known that if the CPU workload is p and the distance of the EM probe from the CPU is q, then the strength of the EM signal from the CPU as measured by the probe is $\frac{p}{q}$. Melba conducted experiments with

	l	m	n	o
d	12			
e	30	40.5		
f			1.25	
g				4.5

4 CPU loads l, m, n, o and for each experiment, Melba placed the probe at 4 distances d, e, f, g from the CPU, thus obtaining 16 EM readings. Melba meticulously recorded the experiment readings in the following matrix. However, as you can see, some of the entries are missing since Melbu again spilled coffee on the spreadsheet where Melba recorded the EM readings. Melbu is angry because not only is a lot of the recorded data gone, but Melba also forgot to write down the values of l, m, n, o, d, e, f, g used in the experiment. Help Melbu find out these values so that the experiment can be repeated. It is known that the values l, m, n, o, d, e, f, g are all positive integers. It is known that $d + e + f + g = 25$. It is also known that if we take the 4 readings in the first column of the matrix (i.e. those corresponding to $p = l$ of which the first two are $12, 30$) and arrange those as a 4D vector $\mathbf{v}$, then $\|\mathbf{v}\|_2^2 = 1169, \|\mathbf{v}\|_1 = 57$.

We have $144 + 900 + a^2 + b^2 = 1169$ and $12 + 30 + a + b = 57$ i.e., $a + b = 15, a^2 + b^2 = 125$ i.e. $a - b = \sqrt{a^2 + b^2 - ((a+b)^2 - a^2 - b^2)} = \pm 5$ which gives $a = 5, b = 10$ or else $a = 10, b = 5$. Thus, the first column of the matrix is $[12,30,5,10]$ or $[12,30,10,5]$. Let us take the first solution – since the matrix is rank-one, it tells us that the second column is $[16.2,40.5,6.75,13.5]$, the third column is $[3,7.5,1.25,2.5]$, the fourth column is $[5.4,13.5,2.25,4.5]$. Now, the value of l must be a multiple of all entries in the first column i.e. it must be a multiple of $\text{LCM}(12,30,5,10) = 60$ i.e. $l = 60x$ giving us $d = 5x, e = 2x, f = 12x, g = 6x$. However, since $d + e + f + g = 25$, we get $x = 1$. This gives us $d = 5, e = 2, f = 12, g = 6$ which gives us $l = 60, m = 81, n = 15, o = 27$. However, an alternative solution is $d = 5, e = 2, f = 6, g = 12$.

Q4. New NN hack **(7 marks)**



We wish to use a neural network, with network diagram given above to the left, to solve a binary classification problem depicted on the right. The green decision boundaries in the figure depict the hyperbola $xy = 4$. The neural network has as parameters a 2×2 matrix $A \in \mathbb{R}^{2\times 2}$, a 2D vector $\mathbf{w} \in \mathbb{R}^2$ and a bias term $b \in \mathbb{R}$. The output of the neural network is $\text{sign}(\mathbf{w}^\top \phi(\mathbf{x}) + b)$ where $\phi(\mathbf{x}) = (A\mathbf{x})^2$ where the square operation is applied coordinate-wise (i.e. The square activation). Find out values of the parameters $A, \mathbf{w}, b$ so that the neural network gives output $+1$ in the yellow region and -1 in the blue region.

The classifier desired is $\text{sign}(4 - xy)$ (and not $\text{sign}(xy - 4)$) as we want the $+1$ label in the yellow region and -1 in the blue region. Elementary calculations give $xy = \frac{1}{4}((x+y)^2 - (x-y)^2)$. This tells us that one way to solve the problem is to have $\mathbf{w} = \left(-\frac{1}{4}, \frac{1}{4}\right), b = 4$ and $A = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$ (note that $A = \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}$ would work too). Other solutions exist too that shift the constants between the matrix and the classifier e.g. $\mathbf{w} = (-1,1), b = 4$ and $A = \frac{1}{2} \cdot \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$.

Q5. Ranked Expectations **(3 marks)**

Suppose X is a 2×2 rank-one matrix-valued random variable i.e. whenever an event happens, a 2×2 rank-one matrix is generated by our observer friend as the value taken by the random variable. It is known that all entries in the random variable are between 0 and 1. Then which of the following properties are satisfied by $\mathbb{E}[X]$ (the expected value of this matrix) -- tick all that apply.

[✓] $\mathbb{E}[X]$ is a 2×2 matrix

[ ] $\mathbb{E}[X]$ must be a rank-one matrix

[✓] $\mathbb{E}[X]$ cannot have rank more than 2

[✓] All entries in $\mathbb{E}[X]$ are between 0 and 1.

Justify your choices briefly below.

Expectations have the same shape as the random variable hence the first option is correct. The second option is not correct all the time – e.g. if the random variable takes the value $\begin{bmatrix}1 & 0\\0 & 0\end{bmatrix}$ with probability 0.5 and $\begin{bmatrix}0 & 0\\0 & 1\end{bmatrix}$ with probability 0.5, then both values in the support are rank 1 but the expectation is $\frac{1}{2}\cdot\begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}$ which is rank 2. Since the expectation is a 2×2 matrix, it can never have a rank more than 2 so the third option is correct. The fourth option is correct since if a random variable always lies in a certain interval, its expectation also lies in that same interval.

Q6. Total Confusion **(6 marks)**

Melbo has designed a linear classifier for a binary classification problem that gives its label prediction as $\text{sign}(\mathbf{w}^\top\mathbf{x} + b)$ where $\mathbf{x}$ is the 3D feature vector of a data point, $\mathbf{w}$ is the 3D model vector and b is the scalar bias term. The classifier was tested on 1000 test data points and the following results were obtained as the confusion matrix. Note that y denotes the true label of a test point and $\hat{y}$ denotes the label predicted by the classifier. The entries in the confusion matrix show how many points of a particular class were classified in a particular manner by the classifier. There are only two classes namely $-1, +1$.

	$\hat{y}=1$	$\hat{y}=-1$
$y=1$	100	700
$y=-1$	100	100

Find the accuracy of the classifier

$$\text{acc} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}} = \frac{200}{1000} = 0.2$$

Find the precision of the classifier

$$\text{prec} = \frac{\text{TP}}{\text{TP}+\text{FP}} = \frac{100}{100+100} = 0.5$$

Find the recall of the classifier

$$\text{rec} = \frac{\text{TP}}{\text{TP}+\text{FN}} = \frac{100}{100+700} = 0.125$$

Find the F1-score of the classifier

$$F_1 = \frac{2\cdot\text{prec}\cdot\text{rec}}{\text{prec}+\text{rec}} = \frac{2\cdot 0.5\cdot 0.125}{0.5+0.125} = \frac{0.125}{0.625} = 0.2$$

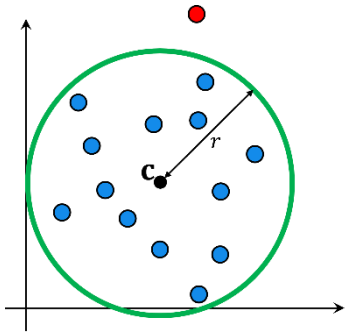
You may have noticed that the classifier is not a very good one. Suggest a simple change to the model parameters that guarantees that the new classifier's accuracy goes up to at least 75% (maybe more). You are not allowed to (re)train on additional data or change the training algorithm. All you can do is make modifications directly to the model parameters $\mathbf{w}, b$ learnt by Melbo.

Set $\hat{\mathbf{w}} = -\mathbf{w}, \hat{b} = -b$ so that the predictions of the classifier are flipped. The new confusion matrix would look like the one given on the right and the new classifier would have 80% accuracy.

	$\hat{y} = 1$	$\hat{y} = -1$
$y = 1$	700	100
$y = -1$	100	100

Q7. Threshold Tuning **(6 marks)**

Melba is designing an anomaly detection algorithm. To do so Melba first embedded all data points as 2D vectors and found the mean vector $\mathbf{c}$. Melba now wants to find a radius threshold r such that when a data point with 2D embedding $\mathbf{x}$ comes along, Melba can declare an anomaly if $\|\mathbf{x} - \mathbf{c}\|_2 \geq r$. Melba has the following data that contains both normal and anomalous points. To simplify things, we have just given the Euclidean distance of the various data points from the center and their label (A for anomalous and N for normal).



$\|\mathbf{x} - \mathbf{c}\|_2$	1	2	3	4	5	6	7	8	9	10
label	N	N	A	N	N	N	A	N	N	A

We call a prediction a **false negative** if an anomalous point was predicted as normal. We call a prediction a **false positive** if a normal point was predicted as anomalous. Anomalous points (correctly) predicted as anomalous are called **true positives** and normal points (correctly) predicted as normal are called **true negatives**. The TPR (**true positive rate**) of a classifier is the number of true positives divided by the total number of anomalous points. The FPR (**false positive rate**) of a classifier is the number of false positives divided by the total number of normal points.

Is there a radius threshold that achieves 100% accuracy?

( ) True

(✓) False

Justify your answer briefly

There exists no real number $\theta \in \mathbb{R}$ such that all anomalous data points all have distances more than θ and all normal points have distances less than θ.

For which radius thresholds will Melba's classifier achieve 100% TPR? Justify your answer briefly.

To achieve 100% TPR, a classifier must classify every anomalous data point as anomalous. Since the closest anomalous data point is distance 3 away from the center, any radius threshold ≤ 3 will achieve 100% TPR.

Among radius thresholds that achieve 100% TPR, find one that minimizes the FPR. Justify your answer briefly.

Radius thresholds that achieve 100% TPR must misclassify at least 5 normal points as anomalous. If the radius threshold drops below 2 it will classify 6 or 7 normal points as anomalous. Thus, the minimum possible FPR is $\frac{5}{7}$ for thresholds achieving 100% TPR – this is achieved by any radius threshold in the range $(2,3]$.