Name: Muhan Guan **Student ID:** 1407870

If you "submit images" on Gradescope, leave name and UniMelb ID blank. If you "submit PDF", write name and student ID.

Instructions

- This assignment is based on a total of 12.5 points.
- Submit through Gradescope.
- Due on August 28, 11.59pm AEST.

Problem I

Recall that for scalar $\lambda > 0$, the probability density function of an "exponential" random variable with parameter λ, is $P(x;\lambda) = \lambda \exp(-\lambda x)$. We have n independent samples $x_1, \ldots, x_n$. Each $x_1, \ldots, x_n$ is a scalar. Each x_i is an "exponential" random variable with parameter λ.

1) [2 points] We define the dataset $D = \{x_1, \ldots, x_n\}$. Give a short expression for the log-likelihood function $l(D;\lambda) = \sum_{i=1}^{n} \log P(x_i;\lambda)$. Show all the steps in your derivation.

The log-likelihood function is the logarithm of the product of the individual probabilities since the samples are independent.

$$\begin{aligned} l(D;\lambda) &= \sum_{i=1}^{n} \log P(x_i;\lambda) \\ &= \sum_{i=1}^{n} \log(\lambda \exp(-\lambda x_i)) \\ &= \sum_{i=1}^{n} (\log(\lambda) + \log(\exp(-\lambda x_i))) \\ &= \sum_{i=1}^{n} (\log(\lambda) - \lambda x_i) \\ &= n\log(\lambda) - \lambda \sum_{i=1}^{n} x_i \end{aligned}$$

2) [2 points] What is the maximum likelihood estimator? In other words, what is the value of λ for which the derivative of $l(D;\lambda)$ with respect to λ is zero? Show all the steps in your derivation.

Now we have $l(D;\lambda) = n\log(\lambda) - \lambda \sum_{i=1}^{n} x_i$

$$\frac{\partial l(D;\lambda)}{\partial \lambda} = \frac{\partial (n\log(\lambda) - \lambda \sum_{i=1}^{n} x_i)}{\partial \lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i$$

set the derivative equal to zero:

$\frac{n}{\lambda} - \sum_{i=1}^{n} x_i = 0$, then we can solve $\lambda = \frac{n}{\sum_{i=1}^{n} x_i}$

Check if second derivative is negative:

$$\frac{\partial^2 l(D;\lambda)}{\partial \lambda^2} = -\frac{n}{\lambda^2} < 0 \text{ as } \lambda^2 > 0$$

So that we can confirm that this critical point is a maximum.

Thus the MLE for λ is $\hat{\lambda} = \frac{1}{\bar{x}}$, where $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$

Name: Muhan Guan **Student ID:** 1407870

If you "submit images" on Gradescope, leave name and UniMelb ID blank. If you "submit PDF", write name and student ID.

Problem II

3) [2 points] Assume we use classifiers from a countably finite model class $\mathcal{F}$ of 1000000 classifiers. On a dataset of 20 i.i.d. samples, we obtain an empirical risk (i.e., training error) of 0.35.

Would you recommend using this classifier or not? Assume we want to be certain with probability at least 0.99. Please answer "Yes" or "No", and clearly explain why.

Now we have $\Pr[R[f]-\hat{R}[f]>\varepsilon \text{ for some } f\in\mathcal{F}]\le\delta$, where $\delta=|\mathcal{F}|e^{-2n\varepsilon^2}$. then we take complement of this event using $\Pr(A)=1-\Pr[\text{not } A]$

$\Pr[R[f]-\hat{R}[f]\le\varepsilon \text{ for all } f\in\mathcal{F}]\ge 1-\delta$. Therefore, with the probability at least $1-\delta$, the true risk of any classifiers in $\mathcal{F}$ can be upper bounded by $\hat{R}[f]+\varepsilon$, where $\hat{R}[f]$ is its empirical risk and $\varepsilon=\sqrt{\frac{\log|\mathcal{F}|+\log(1/\delta)}{2n}}$. $1-\delta=0.99$ $\delta=0.01$

$R[f]\le 0.35+\sqrt{\frac{6\times\log 10+\log(1/0.01)}{2\times 20}}\approx 1.029$ The upper bound on the true risk $R[f]$ is loose. With probability ≥ 0.99, the bound exceeds 1 suggests this classifier tend to have high true risk and we cannot guarantee that the classifier has same performance on training data and testing data (it could even predict all labels wrong!)

Conclusion: No, the evidence provided by our upper bound cannot provide any theoretical guarantee for the real performance of this classifier if we deploy it on unseen data (testing data), so I will not recommend using this classifier.

4) [2 points] Assume we use classifiers from a countably finite model class $\mathcal{F}$ of 2000000 classifiers. On a dataset of 1000 i.i.d. samples, we obtain an empirical risk (i.e., training error) of 0.1.

Would you recommend using this classifier or not? Assume we want to be certain with probability at least 0.99. Please answer "Yes" or "No", and clearly explain why.

We still use the same formula in (3), $R[f]\le 0.1+\sqrt{\frac{(\log(2\times10^6)+\log(1/0.01))}{2\times 1000}}\approx 0.198$ The bound is tight.

with high probability 0.99, the true risk of this classifier will lower than 0.198 when we deploy it on unseen data. Therefore, such bound can provide performance guarantee for our model and the the classifier is robust for future data.

Conclusion: Yes, because the generalization bound suggests that the true error of this classifier is likely to be low, and the classifier can still perform well on unseen data. (recommend using this classifier)

5) [0.5 points] On another dataset with $p \geq 100$ features, assume we run 3 classification algorithms and obtain the following empirical risks (i.e., training errors). We also include the VC dimension of the 3 algorithms.

Algorithm	Empirical risk	VC dimension
A1	0.15	p
A2	0.15	p^2
A3	0.15	p^3

Which algorithm should we prefer? Write "A1" or "A2" or "A3", and clearly explain your answer.

VC generalisation theorem

We have with high probability $1-\delta$, $R[f_n]\le\hat{R}[f_n]+\sqrt{8\frac{VC(\mathcal{F})\log(2en/1)+\log(4/\delta)}{n}}$

Since all algorithms have the same empirical risk $\hat{R}_n[f_n]$, and they used the same training dataset with the same number of samples n, and they are compared under the same confidence level $1-\delta$. So the algorithm with the lowest VC dimension will have the tightest bound, indicating better performance on unseen data and making it the most preferable choice. Now we have $VC(A_1)<VC(A_2)<VC(A_3)$, so A1 should be preferred because it has the smallest VC dimension, which leads to the tightest bound on the true risk, making it the most reliable choice among the three algorithms.

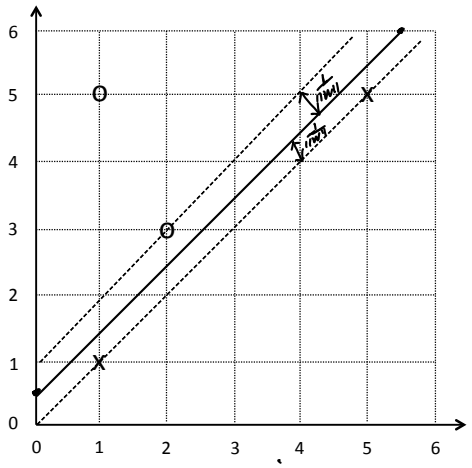
In conclusion, VC dimension is a measure of complexity of model. A model with higher complexity always has stronger ability to fit wider variety of functions, but also has higher risk of overfitting, leading to poorer performance on the prediction of unseen data.

Name: Muhan Guan **Student ID:** 1407870

If you "submit images" on Gradescope, leave name and UniMelb ID blank. If you "submit PDF", write name and student ID.

Problem III

6) [2 points] Suppose we only have 4 training samples in two dimensions: two positive samples $\underline{x}_1 = \begin{bmatrix}1\\1\end{bmatrix}$, $\underline{x}_2 = \begin{bmatrix}5\\5\end{bmatrix}$ and two negative samples $\underline{x}_3 = \begin{bmatrix}2\\3\end{bmatrix}$, $\underline{x}_4 = \begin{bmatrix}1\\5\end{bmatrix}$.



$y = w'x + b$, where $w = \begin{bmatrix}2\\-2\end{bmatrix}$

Figure 1: Training set (positive points shown as X, negative points shown as O)

What is the margin (distance from the separating/decision boundary to a sample) achieved by the maximum margin linear classifier?

Important: We do not require the separating/decision boundary to pass through the origin.

Show all the steps in your derivation. You can draw inside the figure, or use the space below.

We can write the hard-margin SVM objective:

$$\operatorname{argmin}_{w_1,w_2,b} \tfrac{1}{2}(w_1^2+w_2^2)$$

s.t. $\begin{cases} |x(w_1+w_2+b)| \geq 1 & (1)\\ |x(5w_1+5w_2+b)| \geq 1 & (2)\\ -1\,|(2w_1+3w_2+b)| \geq 1 & (3)\\ -1\,|(w_1+5w_2+b)| \geq 1 & (4)\end{cases}$

$\Rightarrow$

$(1)+(2) \Rightarrow 6w_1+6w_2+2b \geq 2$ (5)

$(3)+(4) \Rightarrow -(3w_1+8w_2+2b) \geq 2$ (6)

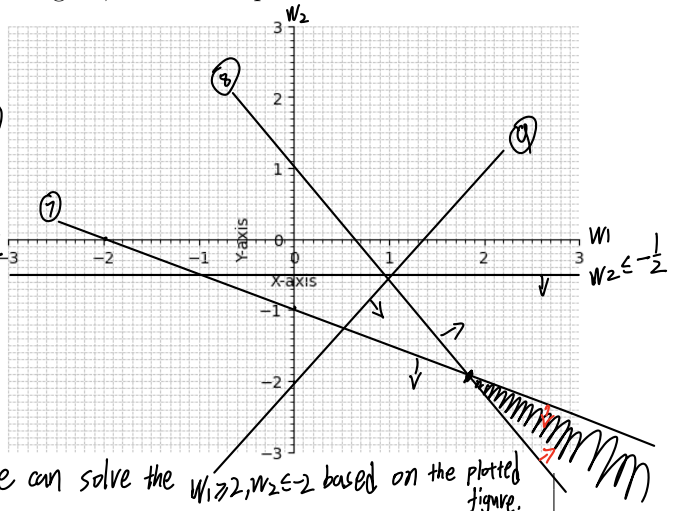
$(1)+(3) \Rightarrow -w_1-2w_2 \geq 2$ (7)

$(1)+(4) \Rightarrow -4w_2 \geq 2 \Rightarrow w_2 \leq -\frac{1}{2}$

$(2)+(3) \Rightarrow 3w_1+2w_2 \geq 2$ (8)

$(2)+(4) \Rightarrow 4w_1 \geq 2 \Rightarrow w_1 \geq \frac{1}{2}$

$(5)+(6) \Rightarrow 3w_1-2w_2 \geq 4$ (9)



The intersecting point of (7) and (8) is $\begin{cases}-w_1-2w_2=2\\3w_1+2w_2=2\end{cases} \Rightarrow \begin{cases}w_1=2\\w_2=-2\end{cases}$, then we can solve the $w_1 \geq 2, w_2 \leq -2$ based on the plotted figure.

The arrow '→' represents the selectable range for w_1 and w_2.

To achieve the optimal goal we choose $w_1=2$, $w_2=-2$, then the maximum margin is $\frac{1}{\|w\|} = \frac{1}{\sqrt{2^2+(-2)^2}} = \frac{1}{2\sqrt{2}}$

The optimal decision boundary and margin are shown inside the above figure.

Name: Muhan Guan **Student ID:** 1407870

If you "submit images" on Gradescope, leave name and UniMelb ID blank. If you "submit PDF", write name and student ID.

Problem IV

7) [2 points] Consider the following dataset with 12 samples.

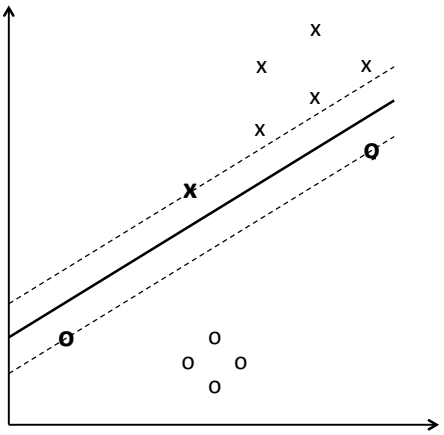


Figure 2: Training set (positive points shown as X, negative points shown as O), maximum margin linear classifier, and the 3 support vectors (in bold)

Imagine we remove some points from the above training set and learn a support vector machine with the remaining points. How many points can we remove (at most) from the above training set so that the separating/decision boundary remains exactly the same as in the above chart?

Write the number of points (for instance: 0, 1, 2, 3, . . . , or 12) and clearly explain your answer. You can draw inside the figure, or use the space below.

We can remove at most 9 points from above trainy set while remain the decision boundary exactly the same as in the above chart.

Support vectors are the point that lie on the margin lines and directly influence the position of the decision boundary. When we solve this usiy lagrangian, the $w^* = \sum_{i=1}^{n} \lambda_i y_i x_i$ and b^* can be recovered from dual solution $y_i(b^* + \sum_{i=1}^{n} \lambda_i^* y_i x_i' x_j) = 1$. We can note that the w^* and b^* are only decided by the trainy points with corresponding $\lambda_i > 0$, those trainiy points are support vectors and we discard all the other points with corresponding $\lambda_i = 0$.

It is obvious that there are 9 points are non-support vectors ($\lambda_j = 0$) and 3 points are support vectors($\lambda_j > 0$) the non-support vectors do not influence the decision boundary, hence removiy them does not change the position.