CIS 419/519: Applied Machine Learning Fall 2021

Homework 1

Handed Out: September 15 *Due: September 27*

Name: QIHANG DAI

PennKey: ahgdyycc

PennID: 78803164

1 Multiple Choice & Written Questions

1. (a) i. increase. no regulariation on x1 and x2 and underfitting in the intercept. the decision boundary would be a line cross the origin. which cant split perfectly.
   ii. increase. the decision boundary would be a line consider only x2 and intercept, which is a line parallel to x1 axis, which cant split perfectly.
   iii. same. there would be a line that perpendicular to x1 axis which can split perfectly.

   (b) i. the intercept can be zero since two class are equal
   ii. class 1 have more possibility. θ_0 should be larger so exp(-θ_0) is smaller and the probability is larger

2. (a) since there is only two point, the boundary should be a perpendicular line to the line connecting two points

   (b) k = 1, so each data point is its own neighbor, for the dataset each data must have it own label thus the decision boundary is acheived

   (c) k = infinite, all data points are neighbors, the family would be a constant model that predict all the same output regardless of input

   (d) when k is inifinite, the bias is high cause underfitting. when k is 1, the variance is high cause overfit

   (e) instead of majority vote, we can use square distance, cubic distance, etc. to weight the vote. the higher order of the distance, it gives more weight on the closer points, which increase the true positive rate.

3. (a) see:

$$P(yes) = \frac{1}{2}$$

$$H(D) = -\frac{1}{2}log_2(\frac{1}{2}) - \frac{1}{2}log_2(\frac{1}{2}) = 1$$

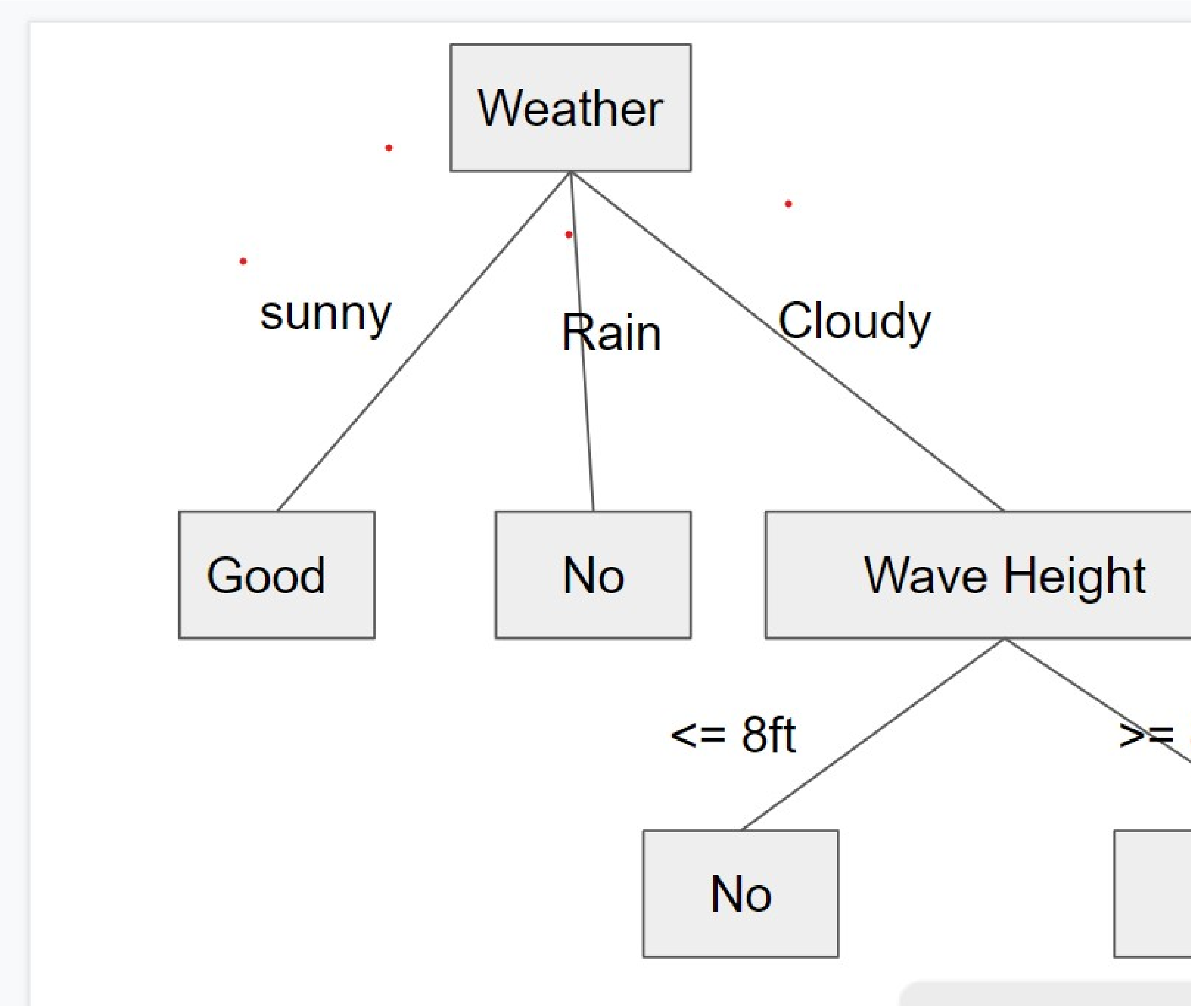
$$IG(D, Weather) = 1 - (\frac{3}{8} * (0) + \frac{2}{8} * (0) + \frac{3}{8}(-\frac{1}{3}log_2\frac{1}{3} - \frac{2}{3}log_2\frac{2}{3})) = 0.65$$

$$\begin{aligned} IG(D, WT) &= 1 - (\frac{2}{8}(-\frac{1}{2}log_2\frac{1}{2} - \frac{1}{2}log_2\frac{1}{2}) && (1) \\ &- \frac{3}{8}(-\frac{1}{3}log_2\frac{1}{3} - \frac{2}{3}log_2\frac{2}{3}) && (2) \\ &- \frac{3}{8}(-\frac{1}{3}log_2\frac{1}{3} - \frac{2}{3}log_2\frac{2}{3})) && (3) \\ &= \frac{6}{8}(1 - \frac{1}{3}log_2\frac{1}{3} - \frac{2}{3}log_2\frac{2}{3}) = 0.0675 && (4) \end{aligned}$$

$$IG(D, Wh) = 1 - (\frac{3}{8}(0) + \frac{5}{8}(-\frac{1}{5}log_2(\frac{1}{5}) - \frac{4}{5}log_2(\frac{4}{5}))) = 0.55 \qquad (5)$$

thus we choose Weather as the root node to split the data

(b) pic:



(c) yes its a good day

(d) no. ID3 use IG to do greedily optimizering, heruestic is not guaranteed to be optimal. ID3 can also be overfitting.

4. ans:
For real-valued input, we cant pick a set of thresholds to do binary split. Thus we can calculate different information gain based on different set of thresholds and pick the one with the highest information gain.
For the optimizer, along with the greedily choose the best IG, we can also publish the errorate of the node to gain a better performance, or use Gain ratio to avoid overfitting.

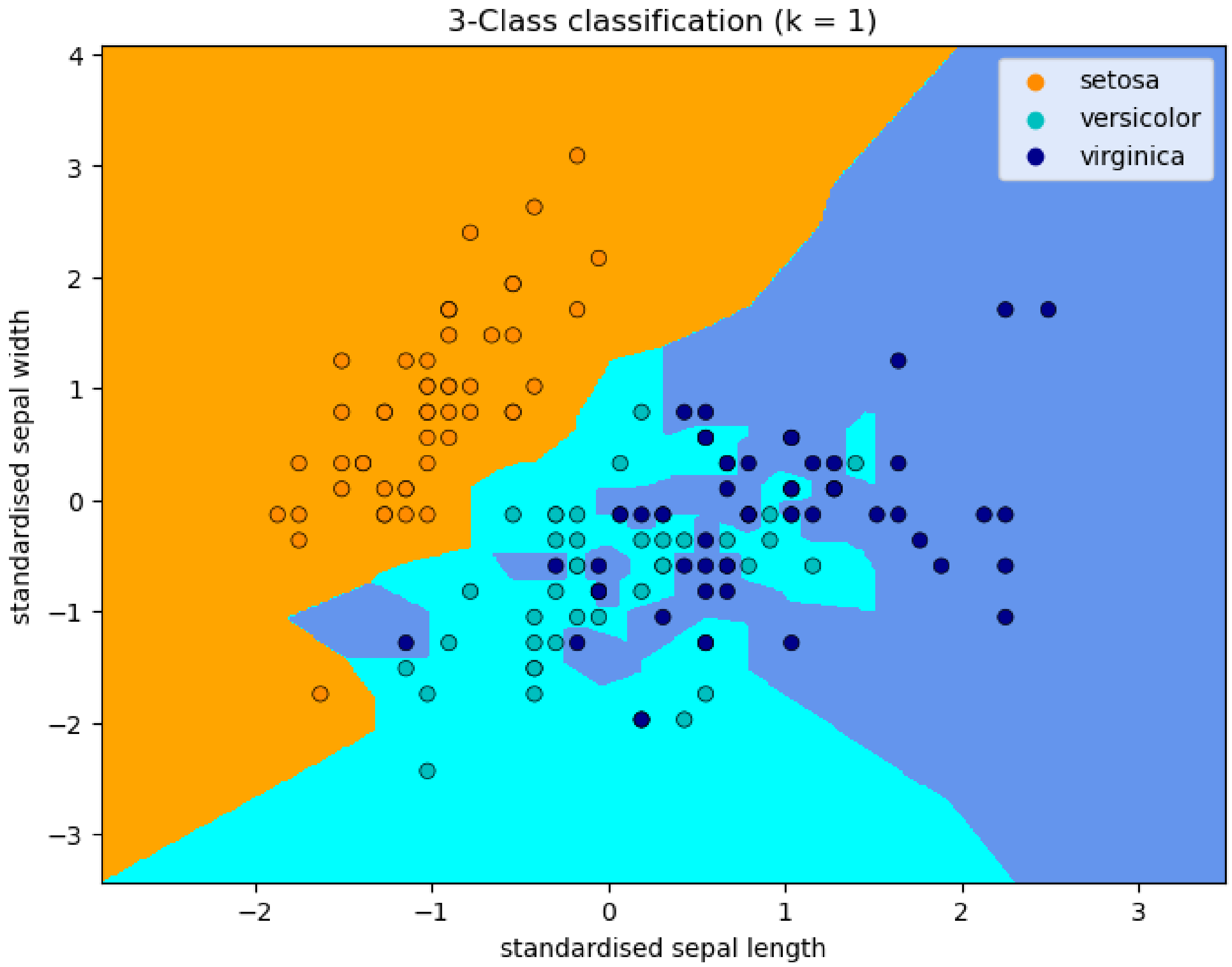
5. ans:

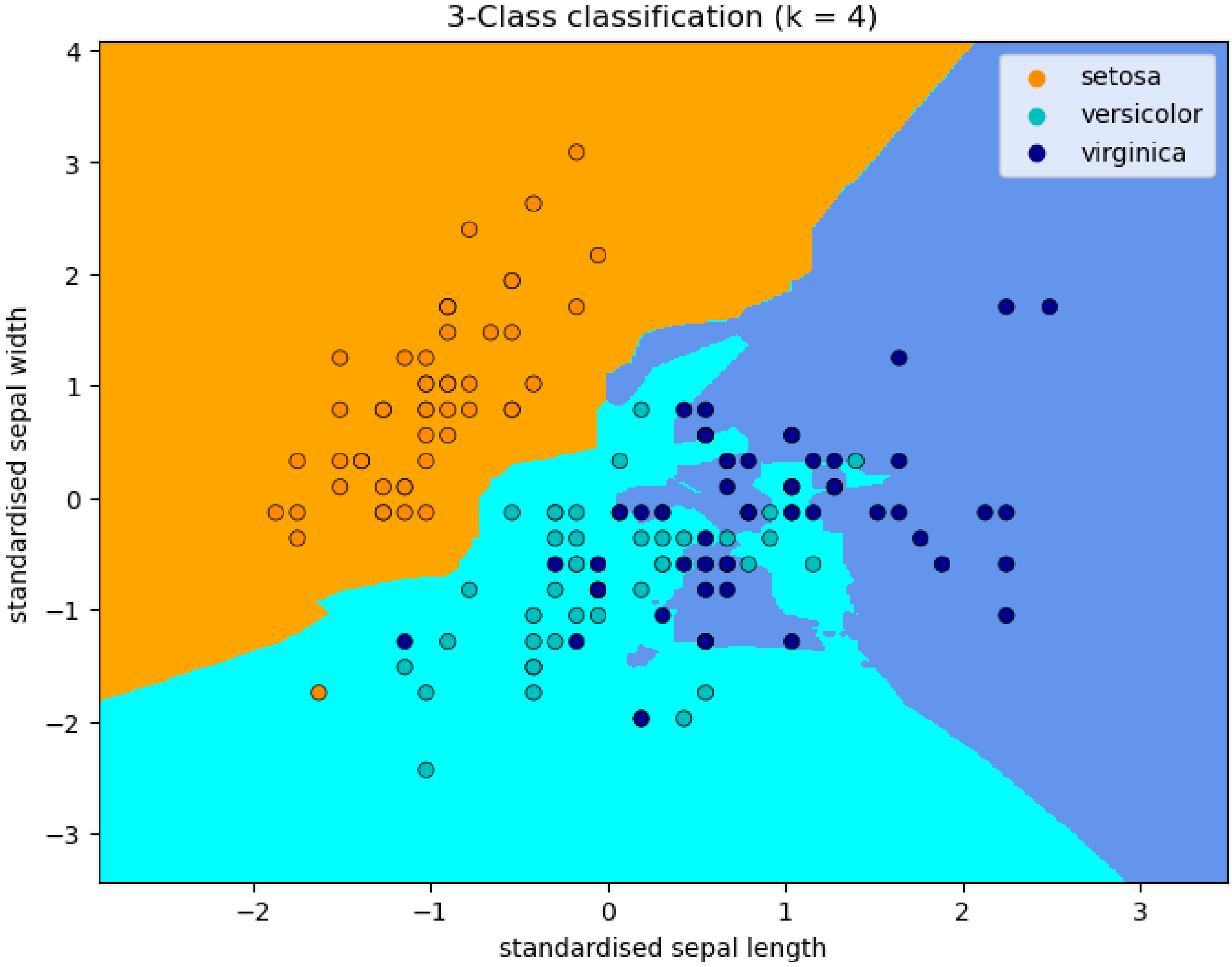
$$
\begin{gathered}
f_{\hat{\beta}}(x) = \hat{\beta}^T x = x^T \hat{\beta} \\
\hat{\beta} = (X^T X)^{-1} X^T Y \\
f_{\hat{\beta}}(x) = x^T (X^T X)^{-1} X^T Y \\
Y = (y1, y2, ..., yn)^T, \\
f_{\hat{\beta}}(x) = x^T (X^T X)^{-1} X^T (y1, y2, ..., yn)^T \\
= \sum_{i=1}^{n} x^T (X^T X)^{-1} X^T yi \\
k_i = x^T (X^T X)^{-1} X^T I_i
\end{gathered}
$$

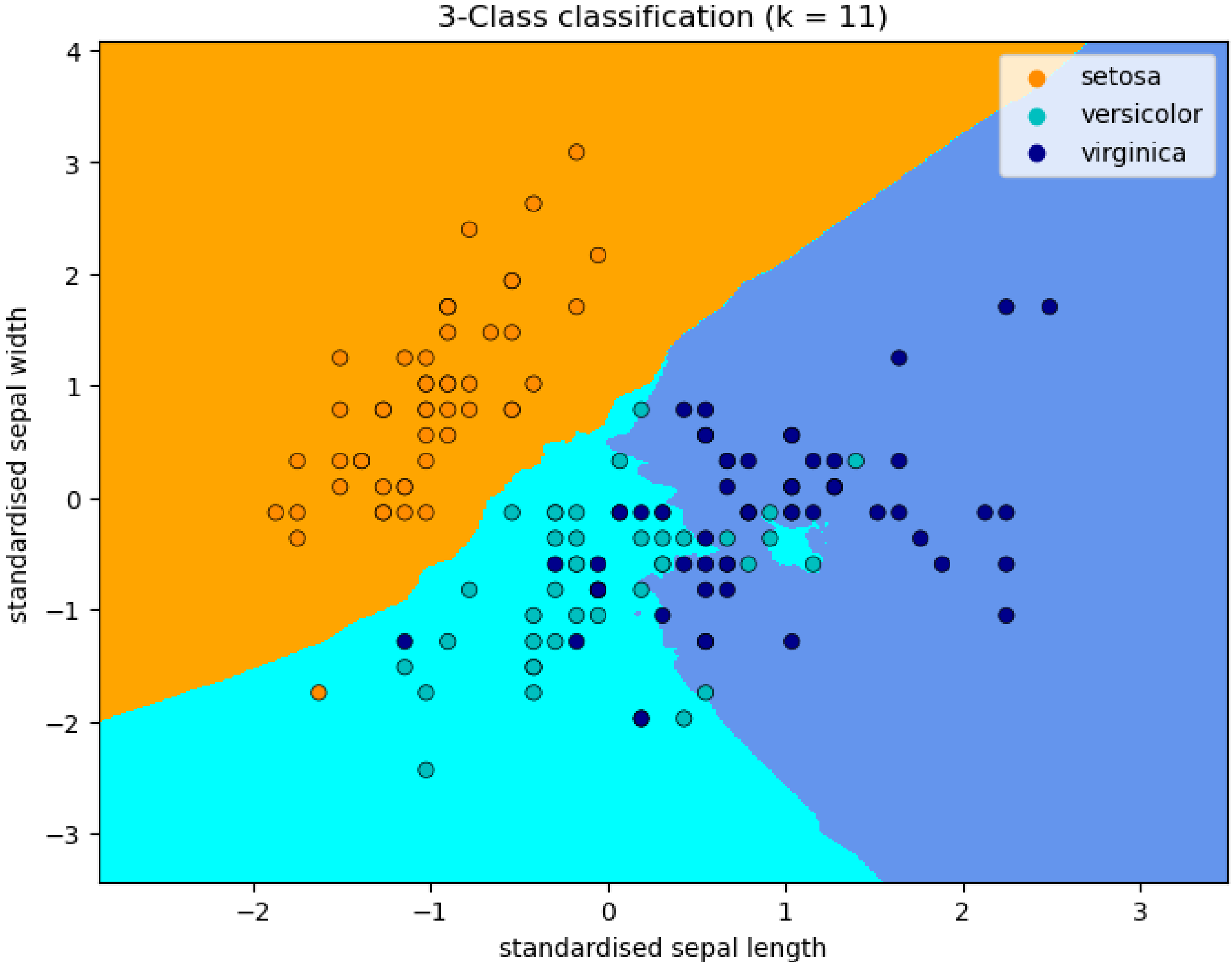
I_i represent (n x 1) vector where only ith element is 1 and others are 0.

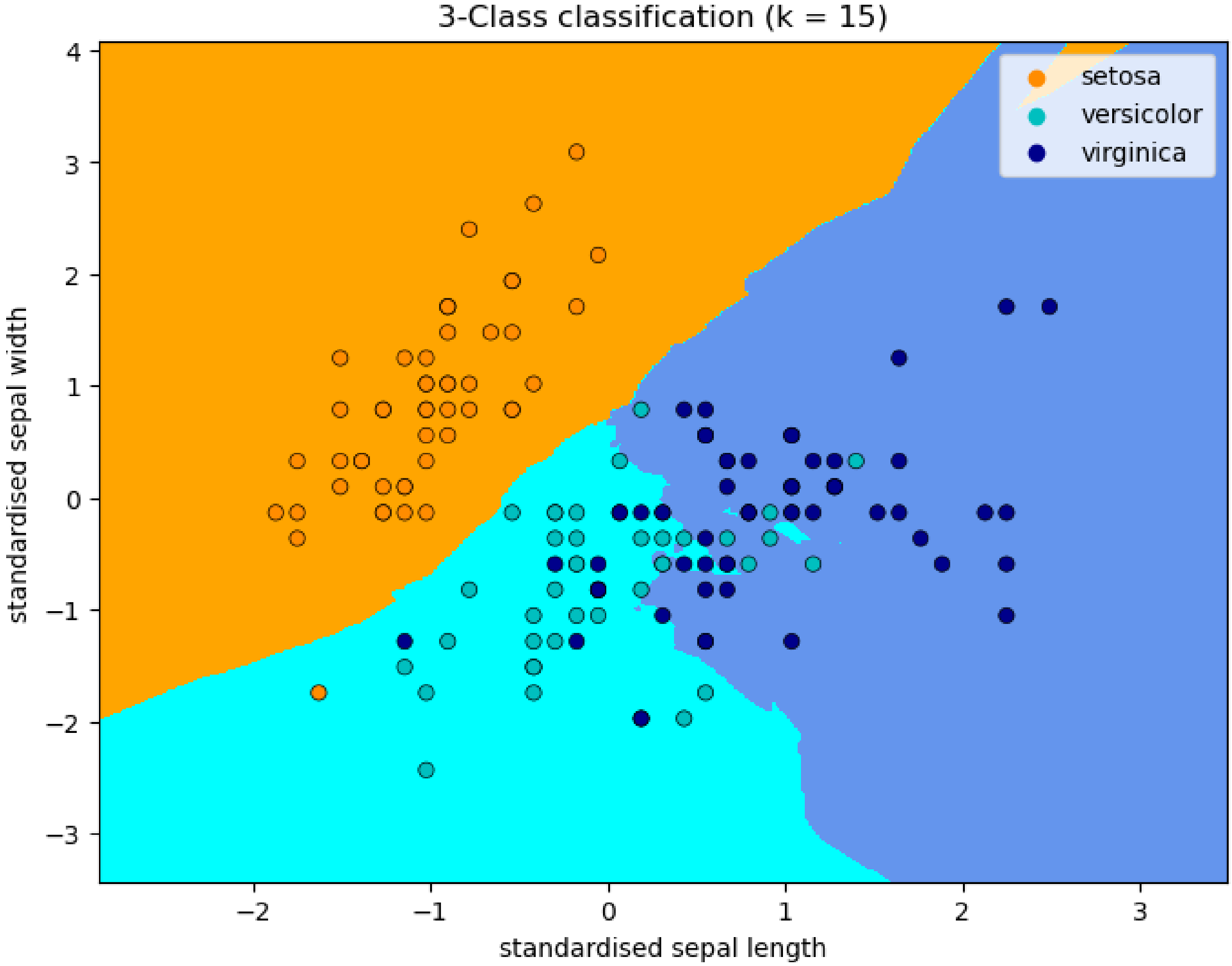
2 Python Programming Questions

Q2.3 here

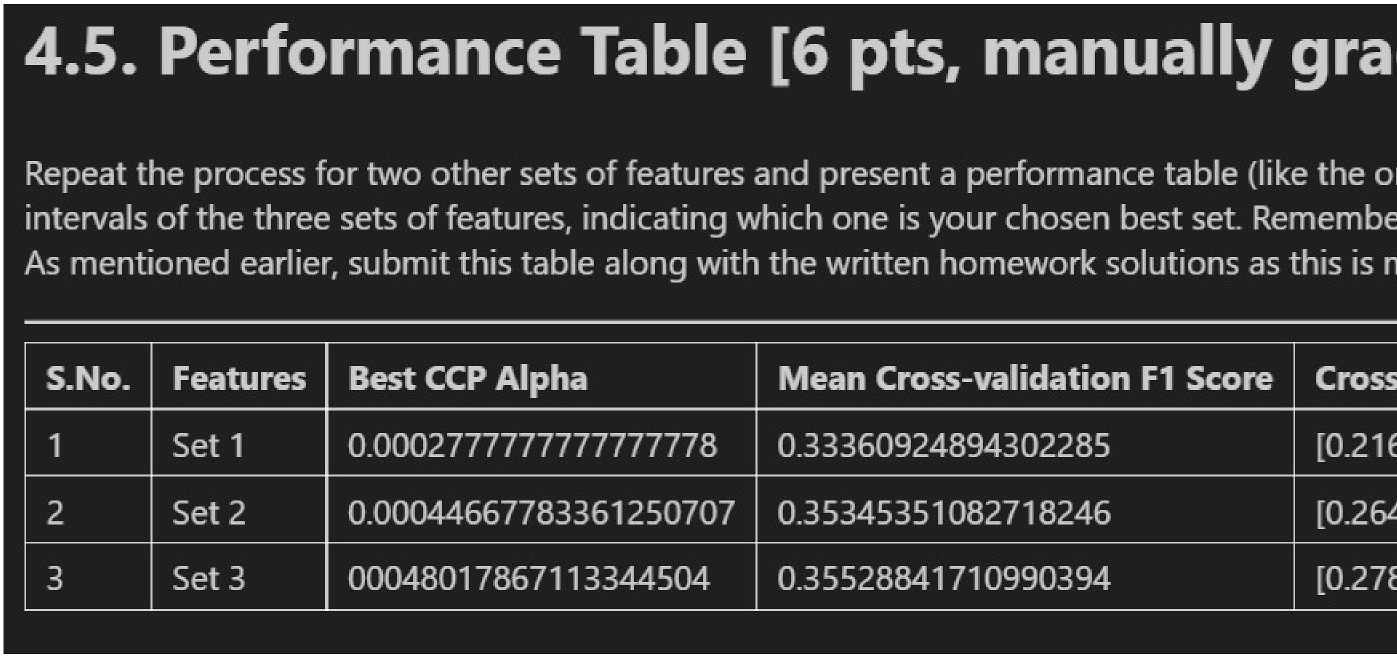

3-Class classification (k = 1)
setosa
versicolor
virginica
standardised sepal width
standardised sepal length
4
3
2
1
0
−1
−2
−3
−2
−1
0
1
2
3


3-Class classification (k = 4)
setosa
versicolor
virginica
standardised sepal width
standardised sepal length
4
3
2
1
0
−1
−2
−3
−2
−1
0
1
2
3


3-Class classification (k = 11)
setosa
versicolor
virginica
standardised sepal width
4
3
2
1
0
−1
−2
−3
−2
−1
0
1
2
3
standardised sepal length


3-Class classification (k = 15)
setosa
versicolor
virginica
standardised sepal width
standardised sepal length
4
3
2
1
0
−1
−2
−3
−2
−1
0
1
2
3

Q4.5

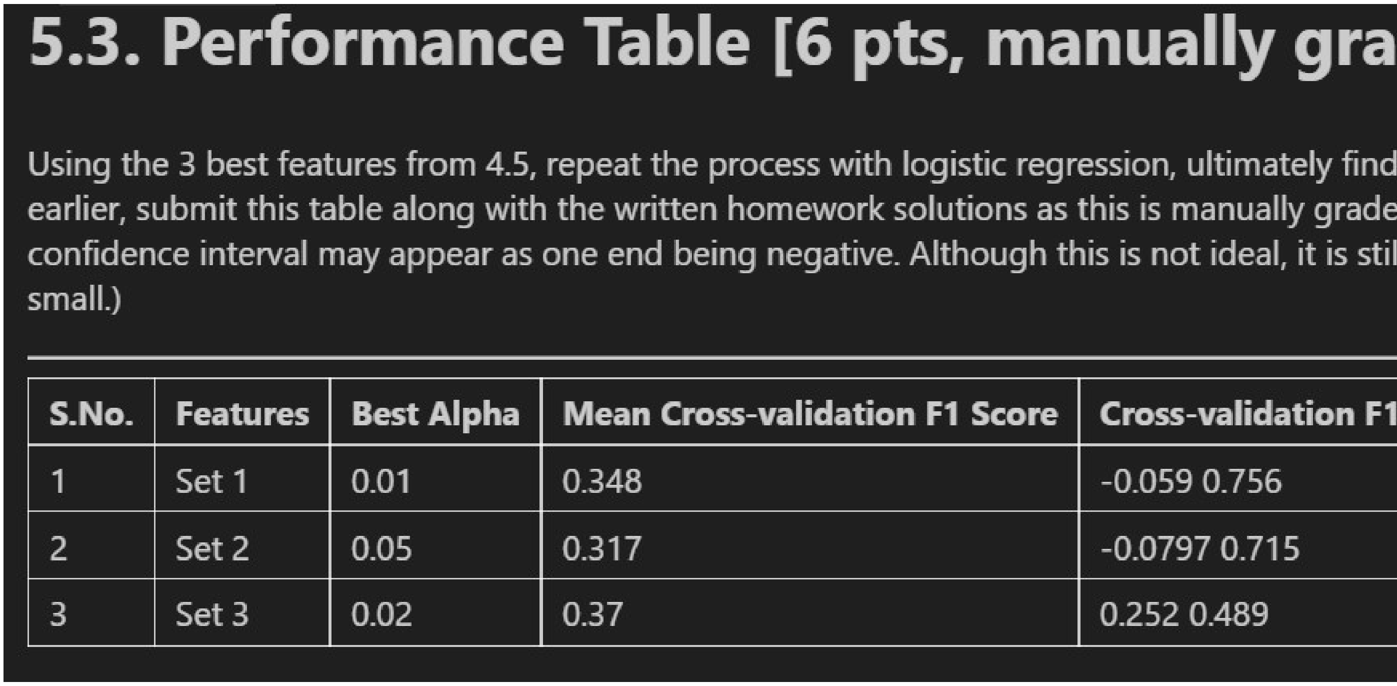


4.5. Performance Table [6 pts, manually gra

Repeat the process for two other sets of features and present a performance table (like the o
intervals of the three sets of features, indicating which one is your chosen best set. Rememb
As mentioned earlier, submit this table along with the written homework solutions as this is m

S.No.	Features	Best CCP Alpha	Mean Cross-validation F1 Score	Cross
1	Set 1	0.0002777777777777778	0.33360924894302285	[0.216
2	Set 2	0.00044667783361250707	0.35345351082718246	[0.264
3	Set 3	00048017867113344504	0.35528841710990394	[0.278

Q5.3



5.3. Performance Table [6 pts, manually gra

Using the 3 best features from 4.5, repeat the process with logistic regression, ultimately find
earlier, submit this table along with the written homework solutions as this is manually grade
confidence interval may appear as one end being negative. Although this is not ideal, it is stil
small.)

S.No.	Features	Best Alpha	Mean Cross-validation F1 Score	Cross-validation F1
1	Set 1	0.01	0.348	-0.059 0.756
2	Set 2	0.05	0.317	-0.0797 0.715
3	Set 3	0.02	0.37	0.252 0.489

TODO: Place your report for Q4.2 here

TODO: Place your paragraph for Q4.2.1 here

(if you are attempting 4.3, remember to include your confidence intervals in the performance table)