

Homework 1*Handed Out: September 15**Due: September 27***Name:** QIHANG DAI**PennKey:** ahgdyyc**PennID:** 78803164**1 Multiple Choice & Written Questions**

1. (a)
 - i. increase. no regulariation on x_1 and x_2 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 x_2 and intercept, which is a line parallel to x_1 axis, which cant split perfectly.
 - iii. same. there would be a line that perpendicular to x_1 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 $\exp(-\theta_0)$ 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) $k = \text{infinite}$, 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:

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

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

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

$$IG(D, WT) = 1 - (\frac{2}{8}(-\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2}) \tag{1}$$

$$- \frac{3}{8}(-\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3}) \tag{2}$$

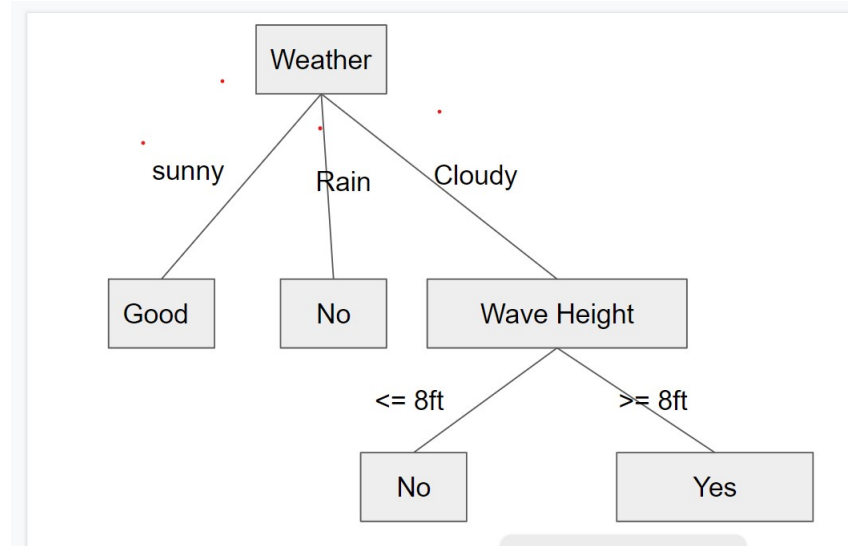
$$- \frac{3}{8}(-\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3})) \tag{3}$$

$$= \frac{6}{8}(1 - \frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3}) = 0.0675 \tag{4}$$

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

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 optimizerring, 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.

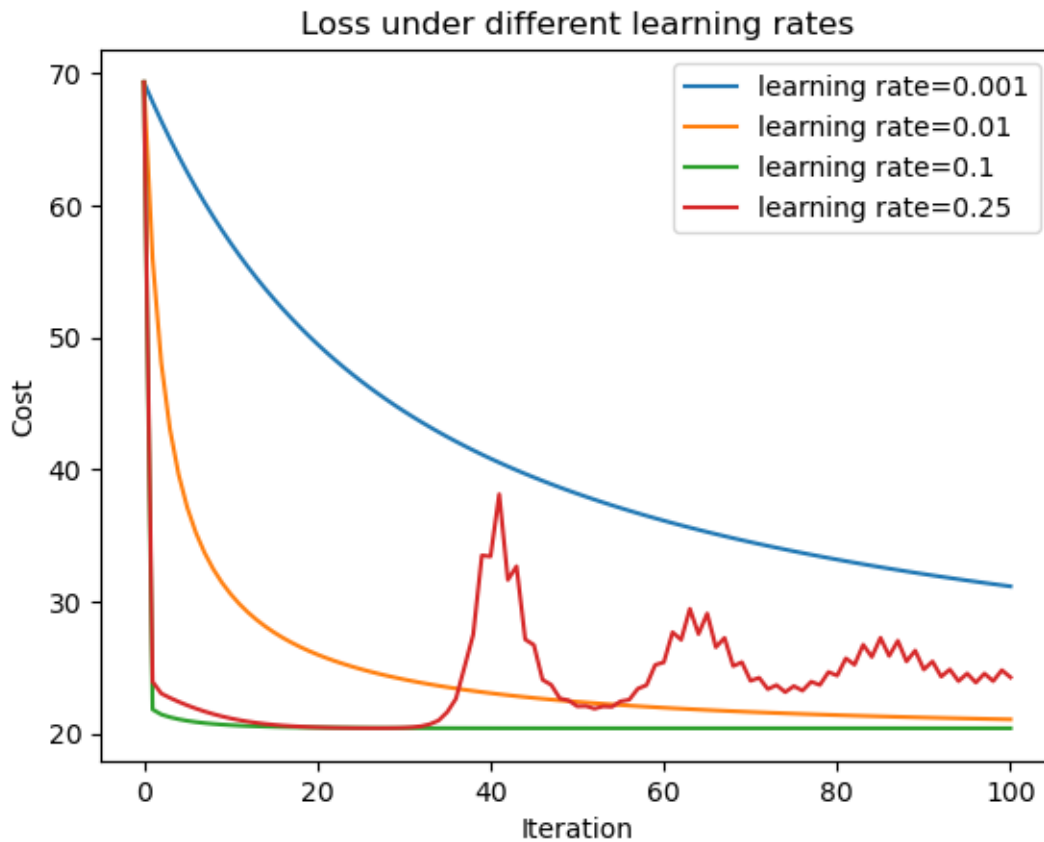
5. ans:

$$\begin{aligned}f_{\hat{\beta}}(x) &= \hat{\beta}^T x = x^T \hat{\beta} \\ \hat{\beta} &= (X^T X)^{-1} X^T Y \\ f_{\hat{\beta}}(x) &= x^T (X^T X)^{-1} X^T Y \\ Y &= (y_1, y_2, \dots, y_n)^T, \\ f_{\hat{\beta}}(x) &= x^T (X^T X)^{-1} X^T (y_1, y_2, \dots, y_n)^T \\ &= \sum_{i=1}^n x^T (X^T X)^{-1} X^T y_i \\ k_i &= x^T (X^T X)^{-1} X^T I_i\end{aligned}$$

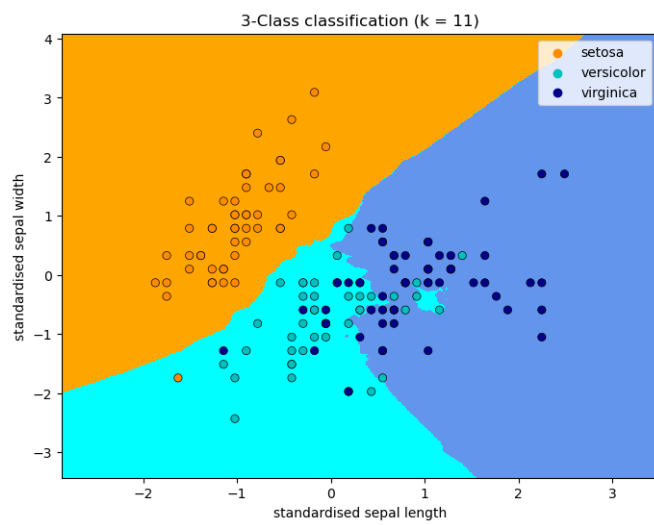
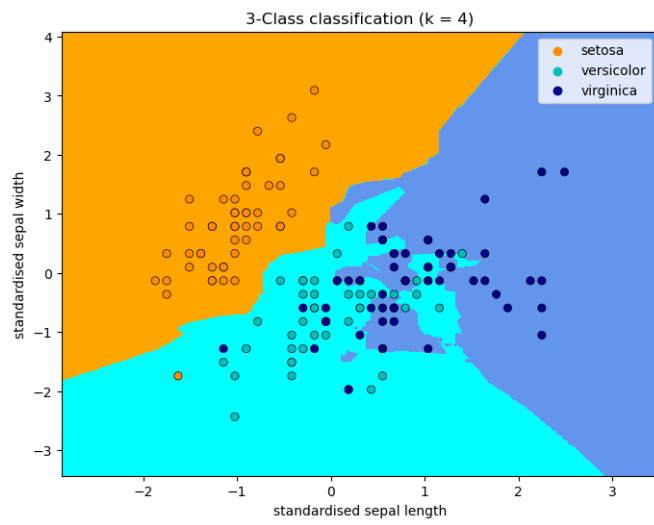
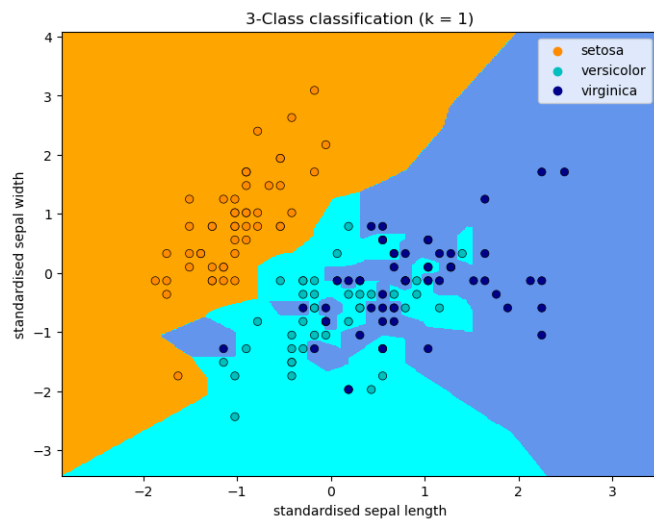
I_i represent $(n \times 1)$ vector where only i th element is 1 and others are 0.

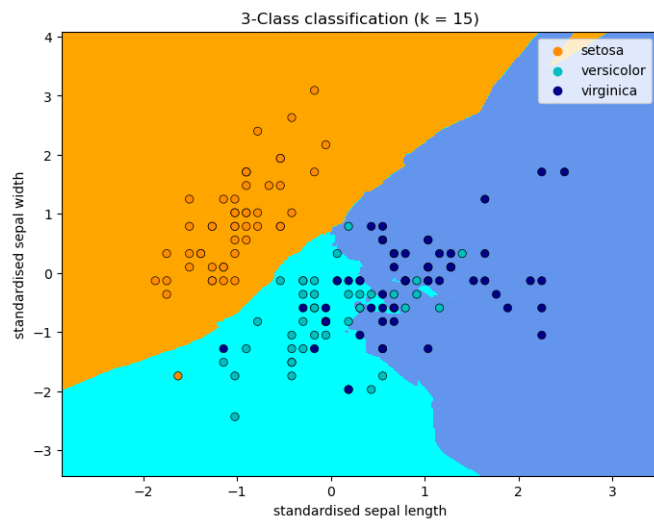
2 Python Programming Questions

Q1.2

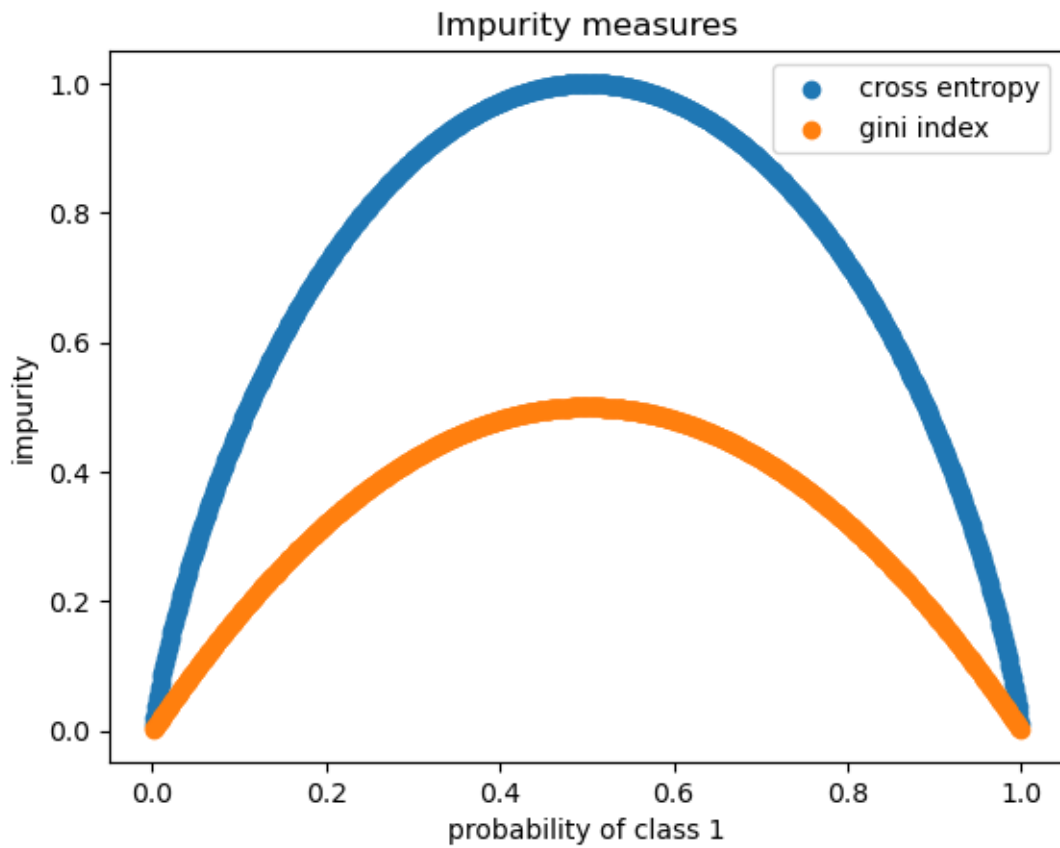


Q2.3 here





Q3.1



Q4.5

4.5. Performance Table [6 pts, manually graded]

Repeat the process for two other sets of features and present a performance table (like the one shown below) that compares the F1 scores and confidence intervals of the three sets of features, indicating which one is your chosen best set. Remember that each set should include the 13 features mentioned earlier. As mentioned earlier, submit this table along with the written homework solutions as this is manually graded.

S.No.	Features	Best CCP Alpha	Mean Cross-validation F1 Score	Cross-validation F1 Score Confidence Interval
1	Set 1	0.000277777777777778	0.33360924894302285	[0.21662747278240435, 0.45059102510364135]
2	Set 2	0.00044667783361250707	0.35345351082718246	[0.26418657801498585, 0.44272044363937907]
3	Set 3	0.0048017867113344504	0.35528841710990394	[0.27897334342557506, 0.4316034907942328]

Q5.3

5.3. Performance Table [6 pts, manually graded]

Using the 3 best features from 4.5, repeat the process with logistic regression, ultimately finding Cross-validation F1 Score Confidence Intervals. Submit this table along with the written homework solutions as this is manually graded. (Note: although F1 scores are calculated, the confidence interval may appear as one end being negative. Although this is not ideal, it is still possible if the fold number is small.)

S.No.	Features	Best Alpha	Mean Cross-validation F1 Score	Cross-validation F1 Score Confidence Interval
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)