Gary Chen
804361027

Homework 3

EE M146
April 24, 2019

Problem 1

a) **What is $H(IsGoodRestaurant)$?**

We find that $P(IsGoodRestaurant) = \frac{5}{8}$, which we use in the definition of entropy:

$$\begin{aligned} H(P) &= -P\log(P) - (1-P)log(1-P) \\ &= \boxed{0.9544} \end{aligned}$$

b) **What is $H(IsGoodRestaurant|HasOutdoorSeating)$?**

We start with the definition of conditional entropy:

$$H(A|X) = \sum_i H(P(A|X=i))P(X=i) \tag{1}$$

We find the probabilities of interest:

i	$P(IGR\|HOS=i)$	$P(HOS=i)$
0	3/4	1/2
1	1/2	1/2

Which yields:

$$\begin{aligned} H(IGR|HOS) &= H(P(IGR|HOS=1))P(HOS=1) + H(P(IGR|HOS=0))P(HOS=0) \\ &= H\left(\frac{1}{2}\right)\left(\frac{1}{2}\right) + H\left(\frac{3}{4}\right)\left(\frac{1}{2}\right) \\ &= \boxed{0.9056} \end{aligned}$$

c) **What is $H(IsGoodRestaurant|X)$, for $X \in \{HasBar, IsClean, HasGoodAtmosphere\}$?**

Considering the probabilities of interest:

X	$P(IGR\|X=1)$	$P(IGR\|X=0)$	$P(X=1)$	$P(X=0)$
$HasBar$	1/2	2/3	1/4	3/4
$IsClean$	1	1/4	1/2	1/2
$HasGoodAtmosphere$	1	1/4	1/2	1/2

Putting all these probabilities into equation 1, we find:

X	$H(IsGoodRestaurant\|X)$
$HasBar$	0.9387
$IsClean$	0.4056
$HasGoodAtmosphere$	0.4056

d) **Calculate the information gain**

Information gain is defined to be:

$$IG(A|X) = H(A) - H(A|X) \tag{2}$$

We already found $H(IGR) = 0.9544$ from part a), so we only need to subtract away the conditional entropies derived in parts b) and c).

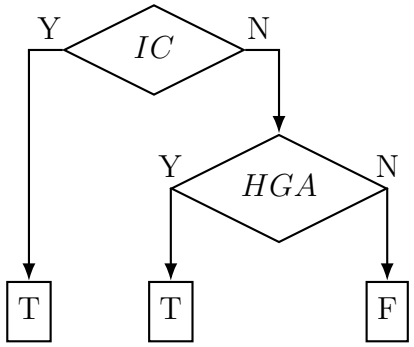
X	$IG(IsGoodRestaurant\|X)$
$HasOutdoorSeating$	0.0488
$HasBar$	0.0157
$IsClean$	0.5488
$HasGoodAtmosphere$	0.5488

e) **Based on the information gain, determine the first attribute to split on.**

We have two attributes that share the max information gain, without going through and optimizing the decision path, we'll choose IsClean as the first split.

f) **Make full decision tree**

Abbreviating the conditions where $IC \equiv IsClean$ and $HGA \equiv HasGoodAtmosphere$.



g) **Determine if restaurants 9 and 10 are good or not**

Based on our decision tree, both restaurants 9 and 10 should be good restaurants. Restaurant 10 is clean, and that immediately indicates a good restaurant. Restaurant 10 is not clean, but does have a good atmosphere, which also indicates a good restaurant.

Problem 2

Supposed we want to minimize

$$J(w_0, w_1) = \sum_{n=1}^{N} \alpha_n (w_0 + w_1 x_{n,1} - y_n)^2 \tag{3}$$

Here $\alpha_n > 0$. Prove that equation 3 has a global optimal solution
From HW2, we know:

$$\frac{\partial J}{\partial w_0} = 2 \sum_{n=1}^{N} \alpha_n (w_0 + w_1 x_{n,1} - y_n)$$
$$\frac{\partial J}{\partial w_1} = 2 \sum_{n=1}^{N} \alpha_n x_{n,1} (w_0 + w_1 x_{n,1} - y_n)$$

Where we then find find second derivatives to construct the Hessian (H).

$$\frac{\partial^2 J}{\partial w_0^2} = 2 \sum_{n=1}^{N} \alpha_n$$
$$\frac{\partial^2 J}{\partial w_1^2} = 2 \sum_{n=1}^{N} \alpha_n w_{n,1}^2$$
$$\frac{\partial^2 J}{\partial w_0 w_1} = 2 \sum_{n=1}^{N} \alpha_n x_{n,1}$$

$$H = 2 \begin{bmatrix} \sum_{n=1}^{N} \alpha_n & \sum_{n=1}^{N} \alpha_n x_{n,1} \\ \sum_{n=1}^{N} \alpha_n x_{n,1} & \sum_{n=1}^{N} \alpha_n x_{n,1}^2 \end{bmatrix} \tag{4}$$

We take H and find if it is positive semi-definite, where all the eigenvalues are non-negative. For a vector $\vec{z} = \begin{bmatrix} z_1 \\ z_2 \end{bmatrix}$:

$$
\begin{aligned}
z^T H z &= 2 \begin{bmatrix} z_1 & z_2 \end{bmatrix} \begin{bmatrix} \sum_{n=1}^N \alpha_n z_1 + \alpha_n x_{n,1} z_2 \\ \sum_{n=1}^N \alpha_n x_{n,1} z_1 + \alpha_n x_{n,1}^2 z_2 \end{bmatrix} \\
&= 2 \sum_{n=1}^N \left(z_1(\alpha_n(z_1 + x_{n,1} z_2)) + z_2(\alpha_n(x_{n,1} z_1 + x_{n,1}^2 z_2)) \right) \\
&= 2 \sum_{n=1}^N \alpha_n (z_1^2 + 2 x_{n,1} z_1 z_2 + x_{n,1}^2 z_2^2) \\
&= 2 \sum_{n=1}^N \alpha_n (z_1 + x_{n,1} z_2)^2
\end{aligned}
$$

Where we see that since $\alpha_n > 0$, the matrix H is positive semi-definite meaning that there is a local minimum. To determine if it is a global minimum, we consider the possible local minima.

$$
\begin{aligned}
\frac{\partial J}{\partial w_0} &= 2 \sum_{n=1}^N \alpha_n (w_0 + w_1 x_{n,1} - y_n) = 0 \\
&= w_0 \sum_{n=1}^N \alpha_n + w_1 \sum_{n=1}^N \alpha_n x_{n,1} = \sum_{n=1}^N \alpha_n y_n \\
\frac{\partial J}{\partial w_1} &= 2 \sum_{n=1}^N \alpha_n x_{n,1} (w_0 + w_1 x_{n,1} - y_n) = 0 \\
&= w_1 \sum_{n=1}^N \alpha_n x_{n,1}^2 + w_0 \sum_{n=1}^N \alpha_n x_{n,1} = \sum_{n=1}^N \alpha_n x_{n,1} y_n
\end{aligned}
$$

We can solve the linear set of equations to find w_0 and w_1:

$$
\begin{bmatrix} \sum_{n=1}^N \alpha_n & \sum_{n=1}^N \alpha_n x_{n,1} \\ \sum_{n=1}^N \alpha_n x_{n,1} & \sum_{n=1}^N \alpha_n x_{n,1}^2 \end{bmatrix} \begin{bmatrix} w_0 \\ w_1 \end{bmatrix} = \begin{bmatrix} \sum_{n=1}^N \alpha_n y_n \\ \sum_{n=1}^N \alpha_n x_{n,1} y_n \end{bmatrix}
$$

To find:

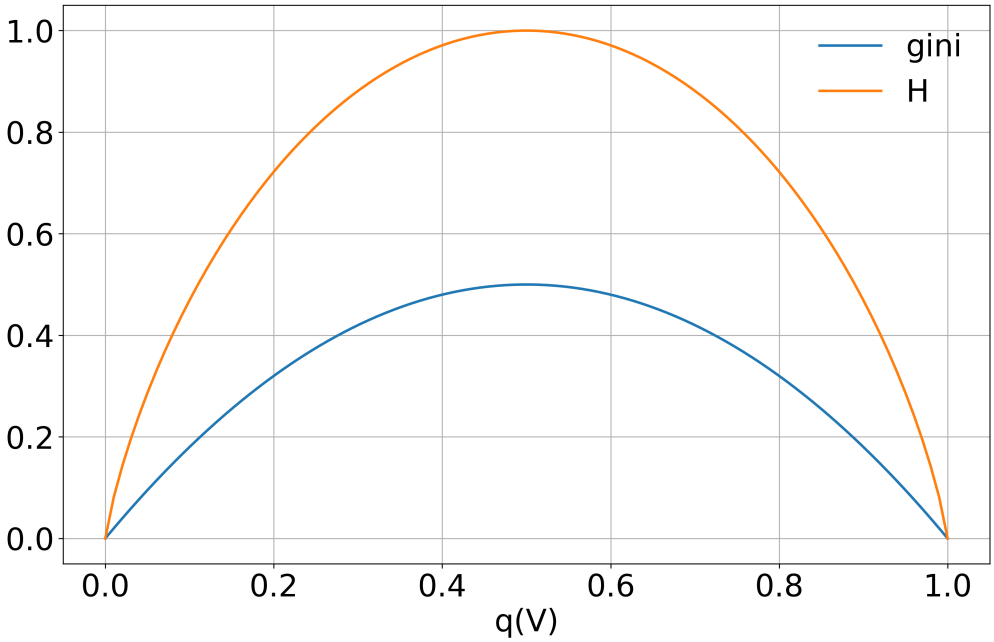
$$w_0 = \frac{\sum_{n=1}^{N}\alpha_n y_n \sum_{n=1}^{N}\alpha_n x_{n,1}^2 - \sum_{n=1}^{N}\alpha_n x_{n,1} y_n \sum_{n=1}^{N}\alpha_n x_{n,1}}{\left(\sum_{n=1}^{N}\alpha_n x_{n,1}\right)^2 - \sum_{n=1}^{N}\alpha_n \sum_{n=1}^{N}\alpha_n x_{n,1}^2},$$

$$w_1 = \frac{\sum_{n=1}^{N}\alpha_n \sum_{n=1}^{N}\alpha_n x_{n,1} y_n - \sum_{n=1}^{N}\alpha_n y_n \sum_{n=1}^{N}\alpha_n x_{n,1}}{\left(\sum_{n=1}^{N}\alpha_n x_{n,1}\right)^2 - \sum_{n=1}^{N}\alpha_n \sum_{n=1}^{N}\alpha_n x_{n,1}^2}.$$

As long as $x_{n,1} \neq 0$, which would yield a division by zero, there is a unique local minimum, which means this is the global minimum.

Problem 3

a) **Plot gini and H**



We can see that both impurity measures behave in very similar ways. They're both symmetric about $q(V) = 0.5$ with a maximum and go to zero on the ends $q(V) = 0, 1$.

b) **Show that if $i(q(V))$ is concave in $q(V)$, then $I(V_1, v_2, V) \geq 0 \forall V_1 \cup V_2 = V, V_1 \cap V_2 = 0$**

First, we consider:

$$\begin{aligned} p(V_1, V)q(V_1) &= \frac{|V_1|}{|V|} \frac{|\{i : i \in V_1, y_i = 1\}|}{V_1} \\ &= \frac{|\{i : i \in V_1, y_i = 1\}|}{|V|} \\ &= \frac{|\{i : i \in V, y_i = 1\}|}{|V|} \frac{|\{i : i \in V_1, y_i = 1\}|}{|\{i : i \in V, y_i = 1\}|} \\ &= \lambda q(V) \end{aligned}$$

Similarly, we find that $P(V_2, V)q(V_2) = (1 - \lambda)q(V)$. Starting from the definition of concavity:

$$\begin{aligned} \Rightarrow i(\lambda q_1 + (1 - \lambda)q_2) &\geq \lambda i(q_1) + (1 - \lambda)i(q_2) \\ i(p(V_1, V)q(V_1) + p(V_2, V)q(V_2) &\geq p(V_1, V)i(q(V_1)) + p(V_2, V)i(q(V_2)) \\ i(q(V)) &\geq p(V_1, V)i(q(V_1)) + p(V_2, V)i(q(V_2)) \end{aligned}$$

We may immediately see from the definition of $I(V_1, V_2, V)$ that it must be ≥ 0:

$$\begin{aligned} I(V_1, V_2, V) &= i(q(V)) - (p(V_1, V)i(q(V_1)) + p(V_2, V)i(q(V_2))) \\ &\geq 0 \end{aligned}$$

c) **Show that entropy is concave**

$$\begin{aligned} H(p) &= -(p \log(p) + (1 - p) \log(1 - p)) \\ \frac{\partial H(p)}{\partial p} &= -(\log(p) + 1 - \log(1 - p) - 1) \\ &= -(\log(p) - \log(1 - p)) \\ \frac{\partial^2 H(p)}{\partial p^2} &= -\left(\frac{1}{p} + \frac{1}{1 - p}\right) \end{aligned}$$

We see that the second derivative of $H(p)$ is always non-positive for positive p.

d) **Show that the gini index is concave**

$$\begin{aligned} gini &= 2p(1-p) \\ &= 2(p-p^2) \\ \frac{\partial gini}{\partial p} &= 2(1-2p) \\ \frac{\partial^2 gini}{\partial p^2} &= 2(-2) \\ &= -4 \end{aligned}$$

Problem 4

a) **For $k \in \{1;3;5;7\}$, what values of k minimize leave-one-out cross-validation error for each dataset? What is the resulting validation error?**

k	Example 1 errors	Example 2 errors
1	10	2
3	6	4
5	4	6
7	4	12

We find that for Example 1, $k = 5, 7$ both yield only 4 errors, while for Example 2, $k = 1$ only has 2 errors.

b) **In general, what are the drawbacks of using too big a k? What about too small a k? To see this, calculate the leave-one-out cross-validation error for the dataset in Figure 1 using $k = 1$ and $k = 13$.**

The size of k determines the characteristic size of the class grouping. If you have a k that is large, you may tend towards the size of your data set, in that case, you are washing out the features of the classes and may end up averaging over the entire set itself. Having too small of a k may cause mislabeled data to skew your results. Considering Example 1 again, we find that we have 10 errors for $k = 1$ due to tightly clustered mislabeled data; when we go to $k = 13$, we are averaging over all of the data points, which gives us an error rate of 14 where every point is mislabeled.

Problem 5

1. **Which examples can be fully separated using a depth 2 decision tree?**

Decision trees 1 and 3 can be separated with depth 2 trees.

2. **Which example would have the most complex decision tree in terms of the number of splits? Explain why.**

   Decision tree 2 would be the most complex because the allowed splits are along the coordinate axes, but the split is actually on a diagonal. Not only is there an angle to the split, the ratio between the distance of class separation and length of separation is small, meaning to be able to define the boundary, many small steps must be made in order to properly separate the data.

3. **If you used a depth 4 decision tree, is one more example now separable? If so, show how you can separate it using a depth 4 decision tree.**

   Decision tree 4 is separable given a depth 4 tree.

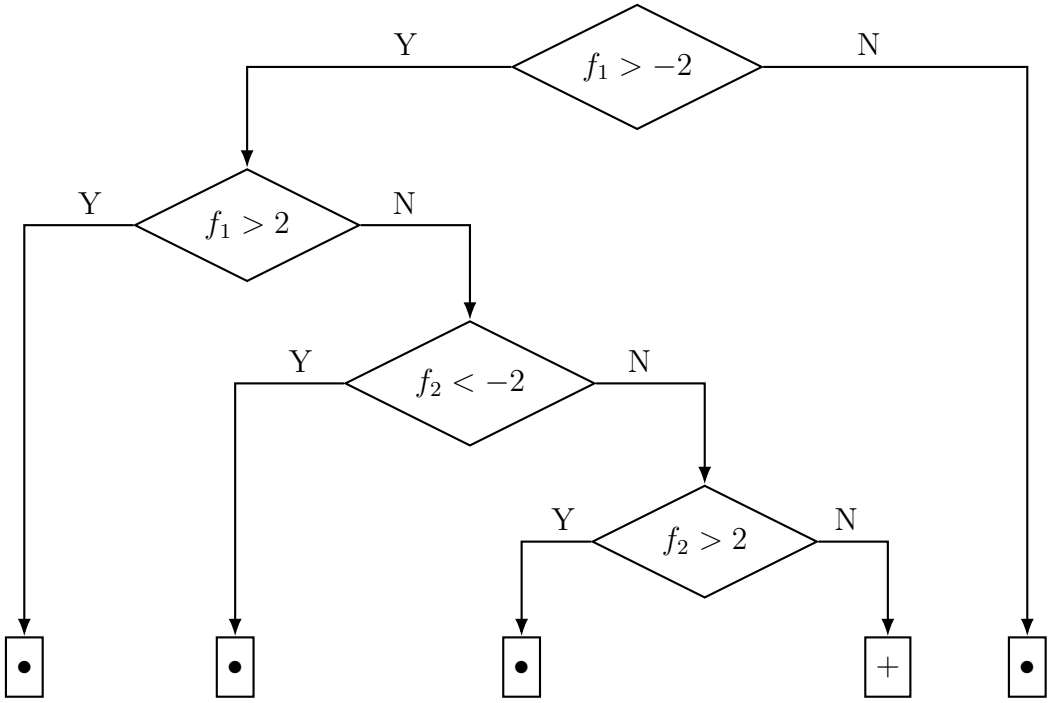


Figure 1: Depth 4 decision tree for example 4.

Problem 6

a) **Provide the accuracy of majority voting where the classifier always outputs the class with the majority in the training set on both the training and the testing set.**

The majority classifier for both training and testing datasets is death. The training dataset has a 61.5% accuracy with this classification, while the test dataset has a 61.0% accuracy.

b) **Use sklearn.tree.DecisionTreeClassifier to train a decision tree classifier on the training data. What is the training and testing accuracy of this model?**

	Accuracy
train	0.986
test	0.701

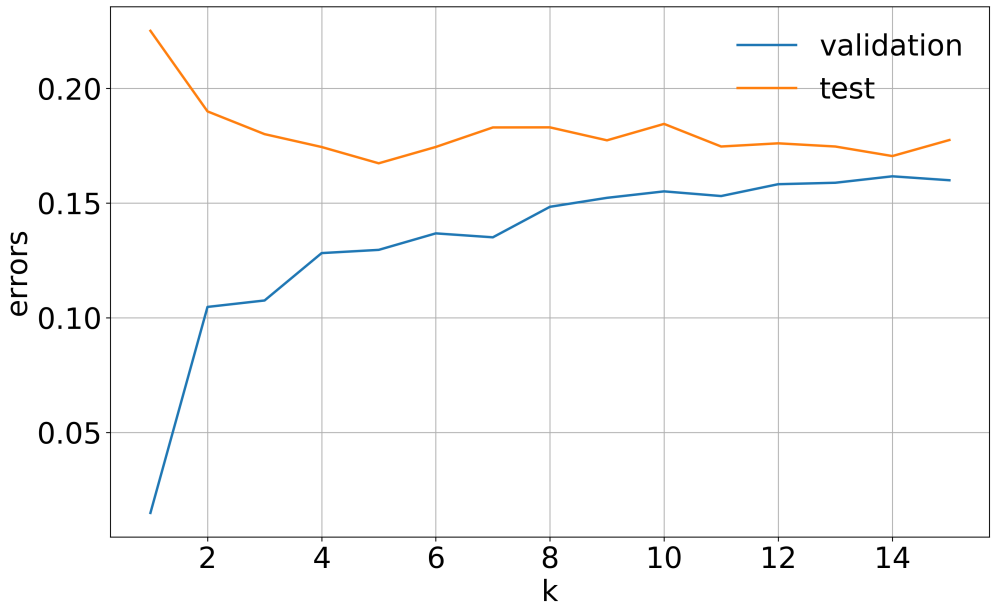
c) **Use sklearn.neighbors.KNeighborsClassifier to train a K-nn classifier the training data. Try it out using $k = 1, 3, 5$. What is the training and testing accuracy of this model?**

k	training accuracy	testing accuracy
1	0.985	0.729
3	0.897	0.768
5	0.869	0.780

d) **Using 10 fold-cross validation, give the validation accuracy for the already trained models.**

Model	k	Cross-validation score
Tree	-	0.796
KNN	1	0.775
	3	0.820
	5	0.833

e) **Find the 10-fold cross-validation error on Knn classifiers trained with "k" set to $1, 2, \ldots, 15$. Provide a plot of the validation error and testing error for each k. Try to explain you observation on how validation error and testing error change with respect to k.**



We observe overfitting with low values of k given the low error on the validation, with higher error on the test. As k increases, we find that the two errors start to converge, meaning that we are starting to average over inconsistencies.

f) **To predict the outcome of the Titanic dataset, which classifier will you use, the k-nearest neighbor or the decision tree?**

I would use the k-nearest neighbor model because the decision tree tends to overfit. While the KNN method can overfit, controlling the k size allows us to mitigate that behavior, while even $k = 1$ yields better results than the decision tree model.