Name: Nicholas Guerrero
USCID: 4088107452

Problem 1

(a) Prove that for any distribution P, the minimum possible entropy is 0. In other words, $H(P) \geq 0$. **(6 points)**

$$H(P) = -\sum_{k=1}^{C} P(Y = k) \log P(Y = k)$$

Since the expression $x*\log x$ will always be negative for values between 0 & 1 and since $P(Y = k)$ is always between 0 & 1 (by the definition of a probability distribution). The expression $\sum_{k=1}^{C} P(Y = k) \log P(Y = k)$ will always sum negative numbers. The sum of all negative numbers is always negative. Thus if we multiply a negative number by -1 we always get a positive number or $-\sum_{k=1}^{C} P(Y = k) \log P(Y = k)$ which is just the function $H(P)$. So $H(P)$ must be a positive number. Thus, the lowest positive number that $H(P)$ can be is 0 in the case where the $P(Y = k)$ is 1 so $x\log x$ becomes $1*\log(1) = 0$ and the rest of the summation has probability 0. (This is the case where all samples belong to the same class) ■

(b) Show that a node where all samples belong to the same class achieves minimum entropy. **(6 points)**

If all samples in a node belong to the same class then

$$P(Y = k) = \begin{cases} 1 & k = i \\ 0 & k \neq i \end{cases}$$

$$H(P) = P(Y = 1) \log P(Y = 1) - \sum_{k=1}^{C-1} P(Y = 0) \log P(Y = 0)$$

$$= -1 * \log(1) - \sum_{k=1}^{C-1} 0 * \log(0) = 0 - \sum_{k=1}^{C-1} 0 * \log(0) = 0$$

$0 \log 0$ is defined naturally as $\lim_{z \to 0+} z \log z = 0$

Here the first class k=i has probability 1 and the rest 0 (since all samples belong to the same class), as the expression shows the entropy obtained has a value of 0 the minimum entropy.

(c) Explain why this is desirable for a decision tree leaf. **(4 points)**

This is desirable because the lower the entropy the more certain our algorithm is of predicting the correct class at a node. For example, the minimum entropy of 0 corresponds to one label fully populating a leaf node and thus the algorithm is "certain".

1.2 Next consider a node with an equal number of samples from all classes. Here the probability of each label is the same.

$$P(Y = k) = \frac{1}{C}$$

We'll prove that P is a maximum entropy distribution for the 2-class case ($C = 2$). Clarification: For the remainder of the problem, consider an arbitrary 2-class distribution P.

(a) Define $P_1 = P(Y = 1)$ and $P_2 = P(Y = 2)$. Rewrite $H(P)$ as a function of just P_1. **(6 points)**

$$H(P) = -\sum_{k=1}^{C} P(Y = k) \log P(Y = k)$$

rewritten as a function of just P_1 :

$$H(P_1) = -\sum_{k=1}^{C=2} P_1(Y = k) \log P_1(Y = k)$$

(b) We want to find P such that $H(P)$ is maximized. In this case we can just find a local maximum of $H(P)$. Calculate the values of P_1 and P_2 that maximize entropy. **(10 points)**

$$H(P_1) = -\sum_{k=1}^{C=2} P_1(Y = k) \log P_1(Y = k) = -\left(\frac{1}{2} * log\left(\frac{1}{2}\right) + \frac{1}{2} * log\left(\frac{1}{2}\right)\right) = log_2(2) = 1$$

$$H(P_2) = -\sum_{k=1}^{C=2} P_2(Y = k) \log P_2(Y = k) = -\left(\frac{1}{2} * log\left(\frac{1}{2}\right) + \frac{1}{2} * log\left(\frac{1}{2}\right)\right) = log_2(2) = 1$$

(c) Explain why nodes with high entropy are not desirable for decision trees. **(4 points)**

Nodes with high entropy is not desirable because it is a measure of uncertainty. In other words, if entropy has a value of 1 we can say the algorithm at this node is can do no metter than choosing at random. This is the most uncertain case and thus not desirable.

Problem 2

2.1 Calculate the classification error over the validation set. Show your work. **(6 points)**

$$right\ child\ of\ x_2 :\ prediction\ =\ 1 \quad error :\ 0\ /\ 13$$

$$left\ child\ of\ x_3 :\ prediction\ =\ 0 \quad error :\ 1\ /\ 13$$
$$right\ child\ of\ x_3 :\ prediction\ =\ 1 \quad error :\ 1\ /13$$

$$left\ child\ of\ x_4 :\ prediction\ =\ 0 \quad error :\ 1/\ 13$$
$$right\ child\ of\ x_4 :\ prediction\ =\ 1 \quad error :\ 2/\ 13$$

$$total\ error\ = \frac{5}{13}\ =\ \sim 38.46\%\ error\ over\ validation\ set$$

2.2 Calculate the classification error if you were to prune the x_2 node. Show your work. **(7 points)**

$$Given\ prune\ of\ x_2$$

$$count\ of\ class\ 0 :\ 4$$
$$count\ of\ class\ 1 :\ 3$$

$$Majority\ class\ is\ class\ 0$$

$$new\ error\ rate\ is\ \frac{1}{13} + \frac{1}{13} + \frac{3}{13} = \frac{5}{13} = \sim 38.46\%\ error\ over\ validation\ set$$

2.3 Calculate the classification error if you were to prune the x_3 node. Show your work. **(7 points)**

Given prune of x_3

count of class 0 : 3
count of class 1 : 3

Majority class is class 1 *or* 0 (*doesnt matter flip coin*)

$$\textit{new error rate is } \frac{1}{13} + \frac{2}{13} + \frac{3}{13} = \frac{6}{13} = \sim 46.15\% \textit{ error over validation set}$$

2.4 Calculate the classification error if you were to prune the x_4 node. Show your work. **(7 points)**

Given prune of x_4

count of class 0 : 4
count of class 1 : 1

Majority class is class 0

$$\textit{new error rate is } \frac{1}{13} + \frac{1}{13} + \frac{1}{13} = \frac{3}{13} = \sim 23.08\% \textit{ error over validation set}$$

2.5 Based on classification error, should this tree be pruned? If so, which node should be pruned? **(5 points)**

The tree should be pruned, the only node that improves the error rate over the validation set is pruning the node x_4

Problem 3

Show that, if data is normalized with unit L2 norm, that is, $\|\boldsymbol{x}\| = 1$ for all $\boldsymbol{x}$ in the training and test sets, changing the distance function from the Euclidean distance to the cosine distance will NOT affect the nearest neighbor classification results. **(10 points)**

Since if we sum and square all features in the dataset we obtain a resultant value is 1. All values after normalization will be between -1 & 1. This implies that the denominator of the second term in the cosine distance formula will be 1 (we take L2 norm).

Thus,

$$C(x, x') = 1 - \frac{\sum_{d=1}^{D}(x_d \cdot x'_d)}{||x||_2 \, ||x'||_2} = 1 - \sum_{d=1}^{D}(x_d \cdot x'_d)$$

$$E(x, x') = \sum_{d=1}^{D}(x_d - x'_d)^2 = \sum_{d=1}^{D} x_d^2 - 2x_d x'_d + x'^2_d = \sum_{d=1}^{D} x_d^2 - 2\sum_{d=1}^{D} x_d x'_d + \sum_{d=1}^{D} x'^2_d$$

$$= 1 - 2\sum_{d=1}^{D} x_d x'_d + 1 = 2 - 2\sum_{d=1}^{D} x_d x'_d = 2\left(1 - \sum_{d=1}^{D}(x_d \cdot x'_d)\right) = 2 * C(x, x')$$

As shown, since the two distance functions are ratios of one another (by 2). All points will retain the same classification results in the nearest neighbors algoirthm.

3.2 Assume we have a dataset, each data $\boldsymbol{x}$ is a 100 dimensional binary vector, i.e. $\boldsymbol{x} \in \{0,1\}^{100}$, and each $\boldsymbol{x}$ is assigned a label $\in \{0,1\}$. You can assume that all data points are distinct, i.e. $\forall \boldsymbol{x}_i, \boldsymbol{x}_j$ in the dataset, $\boldsymbol{x}_i \neq \boldsymbol{x}_j$.

(a) Can we have a decision tree to classify the dataset with zero classification error w.r.t. their labels? Explain your answer. **(5 points)**

We can. Consider the case of 2-dimensional binary vector x. This vector we can classify with zero error by splitting on each feature with a value of 0.5. For the more complex case of a 100 dimensional binary vector is trivial. This would simply require a higher depth tree where we would split on each of the 100 features with a value of 0.5.

(b) Can we specify a 1-NN over the dataset to result in exactly the same classification as our decision tree? Explain your answer. **(5 points)**

Yes, we can specify our model by choosing at least 200 points. Say these 200 points are 0.5 ± 0.1
for each of the dimensions (two points above and below 0.5 on each dimension). Then our training data will be classified by which of these points it is nearest to, which will be one of the 200 points that constitute our model.

Problem 4

(a) We are given a dataset **X** that has only 1 feature. $X = [X_1]$ i.e. X_{nxp} where $p = 1$. Write the ridge regression loss function (MSE + regularization) for this problem. What is the value of β_1 in terms of **X** and **Y**? **(1 point)**

ridge regression loss function:

$$\varepsilon(\beta_1) = RSS(\beta_1) + \eta||\beta_1||_2^2 = \sum_n (\beta_1 X_1 - Y_1)^2 + \eta||\beta_1||_2^2$$
$$= (\beta_1 X_1 - Y_1)^T * (\beta_1 X_1 - Y_1) + \eta||\beta_1||_2^2$$
$$= \left(\beta_1^T X_1^T - Y_1^T\right) * (\beta_1 X_1 - Y_1) + \eta||\beta_1||_2^2$$
$$= Y_1^T * Y_1 - Y_1^T X_1 \beta_1 - \beta_1^T X_1^T Y_1 + \beta_1^T X_1^T X_1 \beta_1 + \eta||\beta_1||_2^2$$
$$= Y_1^T * Y_1 - 2Y_1^T X_1 \beta_1 + \beta_1^T X_1^T X_1 \beta_1 + \eta||\beta||_2^2$$

$$\nabla RSS(\beta_1) = \nabla\left(Y_1^T * Y_1 - 2Y_1^T X_1 \beta_1 + \beta_1^T X_1^T X_1 \beta_1 + \eta||\beta_1||_2^2\right)$$
$$= 0 - 2Y_1^T X_1 + 2X_1^T X_1 \beta_1 + 2\eta\beta_1 = 0$$
$$\implies$$
$$-Y_1^T X_1 + X_1^T X_1 \beta_1 + \eta\beta_1 = 0$$
$$\left(X_1^T X_1 + \eta I\right)\beta_1 = Y_1^T X_1$$
$$\beta_1 = \left(X_1^T X_1 + \eta I\right)^{-1} * Y_1^T X_1 \leftarrow \textit{result is } 1x1 \textit{ vector (scalar) since}$$

X_1^T *is* 1 *x n vector*
X_1 *is n x* 1 *vector*
Y_1^T*is* 1 *x n vector*

(b) We expand the feature set by adding $k - 1$ copies of this feature to our dataset as new columns to get a new dataset $X = [X_1 X_2 X_3 ... X_k]$ i.e. X_{nxp} where $p = k$ now. We fit ridge regression again. Find the new coefficients β_2 where β_2 is a $kx1$ vector. Write the loss function and show your work. Summarize how the magnitude compares with β_1 in (a). **(4 points)**

ridge regression loss function: $\varepsilon(\beta_2) = RSS(\beta_2) + \eta||\beta_2||_2^2 = \sum_n (\beta_2 X - Y)^2 + \eta||\beta_2||_2^2$

$$= (\beta_2 X - Y)^T * (\beta_2 X - Y) + \eta||\beta_2||_2^2$$
$$= \left(\beta_2^T X^T - Y^T\right) * (\beta_2 X - Y) + \eta||\beta_2||_2^2$$
$$= Y^T * Y - Y^T X \beta_2 - \beta_2^T X^T Y + \beta_2^T X^T X \beta_2 + \eta||\beta_2||_2^2$$
$$= Y^T * Y - 2Y^T X \beta_2 + \beta_2^T X^T X \beta_2 + \eta||\beta_2||_2^2$$

$$\nabla RSS(\beta_2) = \nabla\left(Y^T * Y - 2Y^T X\beta_2 + \beta_2^T X^T X\beta_2 + \eta||\beta_2||_2^2\right)$$
$$= 0 - 2Y^T X + 2X^T X\beta_2 + 2\eta\beta_2 = 0$$
$$\Longrightarrow$$
$$-Y^T X + X^T X\beta_2 + \eta\beta_2 = 0$$
$$\left(X^T X + \eta I\right)\beta_2 = Y^T X$$
$$\beta_2 = \left(X^T X + \eta I\right)^{-1} * Y^T X \leftarrow \textit{result is k x k vector (matrix) since}$$

X^T *is k x n vector*
X *is n x k vector*
Y^T*is k x n vector*

Compared the β_1, β_2 has a k-1 more elements. Thus the magnitude of β_2 is larger by a factor of k

(c) We will try to carry over our understanding to higher dimensional vectors. Consider a general dataset X_{nxp} where $p > 1$. Write the loss function and the value of β using the matrix notation as discussed in class. **(1 point)**

ridge regression loss function: $\varepsilon(\beta) = RSS(\beta) + \eta||\beta||_2^2 = \sum_n (\beta X - Y)^2 + \eta||\beta||_2^2$

$$= (\beta X - Y)^T * (\beta_2 X - Y) + \eta||\beta||_2^2$$
$$= \left(\beta^T X^T - Y^T\right) * (\beta X - Y) + \eta||\beta||_2^2$$
$$= Y^T * Y - Y^T X\beta - \beta^T X^T Y + \beta^T X^T X\beta + \eta||\beta||_2^2$$
$$= Y^T * Y - 2Y^T X\beta + \beta^T X^T X\beta + \eta||\beta||_2^2$$

$$\nabla RSS(\beta) = \nabla\left(Y^T * Y - 2Y^T X\beta + \beta^T X^T X\beta + \eta||\beta||_2^2\right)$$
$$= 0 - 2Y^T X + 2X^T X\beta + 2\eta\beta = 0$$
$$\Longrightarrow$$
$$-Y^T X + X^T X\beta + \eta\beta = 0$$
$$\left(X^T X + \eta I\right)\beta = Y^T X$$
$$\beta = \left(X^T X + \eta I\right)^{-1} * Y^T X \leftarrow \textit{result is p x p vector since}$$

X^T *is p x n vector*
X *is n x p vector*
Y^T*is p x n vector*

(d) We augment the features in X with a copy of all features in the dataset. $X' = [XX]$ such that X' has dimension $nx2p$, how will the value of β change? Write the new loss function and calculate new coefficients. Show your work. **(4 points)**

ridge regression loss function: $\varepsilon(\beta) = RSS(\beta) + \eta||\beta||_2^2 = \sum_n (\beta X' - Y)^2 + \eta||\beta||_2^2$

$= (\beta X' - Y)^T * (\beta_2 X' - Y) + \eta||\beta||_2^2$

$= (\beta^T X'^T - Y^T) * (\beta X' - Y) + \eta||\beta||_2^2$

$= Y^T * Y - Y^T X'\beta - \beta^T X'^T Y + \beta^T X'^T X'\beta + \eta||\beta||_2^2$

$= Y^T * Y - 2Y^T X'\beta + \beta^T X'^T X'\beta + \eta||\beta||_2^2$

$\nabla RSS(\beta) = \nabla\left(Y^T * Y - 2Y^T X'\beta + \beta^T X'^T X'\beta + \eta||\beta||_2^2\right)$

$= 0 - 2Y^T X' + 2X'^T X'\beta + 2\eta\beta = 0$

$\Longrightarrow$

$-Y^T X' + X'^T X'\beta + \eta\beta = 0$

$\left(X'^T X' + \eta I\right)\beta = Y^T X'$

$\beta = \left(X'^T X' + \eta I\right)^{-1} * Y^T X'$ ← *result is 2p x 2p vector (matrix) since*

X^T *is 2p x n vector*

X *is n x 2p vector*

Y^T*is 2p x n vector*

β changed by a factor of 2 in the row and column direction

(e) **Lasso and Multicollinearity :** Consider a scenario that we were using LASSO instead of ridge regression. Can you predict how the new β will be different from old β when multicollinearity is introduced? Give a brief summary. No proof or calculation required. **(2 points)**

In the equation, since L_1 regularization is now used when the derivative is taken and the loss function is set equal to zero, we no longer have a coefficient of 2 on the term $2\eta\beta$. Instead we will obtain an expression $\eta * sigmoid(\beta)$. Due to this new expression in our calculation penalize the absolute value of our weights as opposed to the squared magnitude of our weights and so β have higher magnitude than if we used L_2 regularization.