ECE M146 Discussion 3
Introduction to Machine Learning
Instructor: Lara Dolecek
TA: Zehui (Alex) Chen, Ruiyi (John) Wu

1. **Decision Tree Example** You're stuck in a forest with nothing to eat. Suddently, you spot a mushroom but you don't know if its poisonous. Luckly, you've studied some mushrooms as part of a class to fulfill your undergraduate requirements. Your previous knowledge is summarized by the following chart:

Sample #	IsColorful	IsSmelly	IsSmooth	IsSmall	IsPoisonous
1	0	0	0	1	1
2	0	0	0	0	0
3	1	0	1	1	1
4	1	0	0	0	1
5	0	0	1	0	0
6	0	0	1	0	0
7	1	1	0	0	1
8	1	1	1	0	1
9	0	1	1	0	?

(a) What is the entropy of IsPoisonous, i.e., $H(IsPoisonous)$?

(b) Calculate the conditional entropy of IsPoisonous conditioning on IsColorful. To do this, first compute $H(IsPoisonous|IsColorful = 0)$ and $H(IsPoisonous|IsColorful = 1)$, then weight each term by the probabilities $P(IsColorful = 0)$ and $P(IsColorful = 1$, respectively. Namely, calculate the following:

$$\begin{aligned}&H(IsPoisonous|IsColorful)\\=&P(IsColorful = 0)H(IsPoisonous|IsColorful = 0)\\+&P(IsColorful = 1)H(IsPoisonous|IsColorful = 1).\end{aligned}$$

(c) Similarly, calculate

$$H(IsPoisonous|X), \text{for } X \in \{IsSmelly, IsSmooth, IsSmall\},$$

i.e., the conditional entropy of IsPoisonous conditioning on the other three features.

(d) Calculate the information gain:

$$I(IsPoisonous; X) = H(IsPoisonous) - H(IsPoisonous|X),$$

for

$$X \in \{IsColorful, IsSmelly, IsSmooth, IsSmall\}.$$

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

(f) Make the full decision tree. After each split, treat the sets of samples with $X = 0$ and $X = 1$ as two separate sets and redo (b), (c), (d) and (e) on each of them. X is the feature for previous split and is thus excluded from the available features which can be split on next. Terminate splitting if after the previous split, the entropy of IsGoodRetaurant in the current set is 0. For example, if we choose IsSmall as our first feature to split, we get $H(IsGoodRetaurant|IsSmall = 1) = 0$. We thus stop splitting the tree in this branch. Draw the tree and indicate the split at each node.

(g) Now, determine if restaurants 9 and 10 are good or not.

2. **Regression Tree** So far, we have only focused on using tree structures for classification. We can also apply them to regression problems. In decision trees, we define the spread of a discrete dataset by using entropy. For real valued sets, we use variance.

   For each set V, we associate a regression value u that minimizes the variance

$$Var(V) = \sum_{x_i \in V} (x_i - u)^2.$$

   (a) What is the value of u that minimizes $Var(V)$?

   (b) Assume that a decision tree is trying to split V into two sets such that $V_1 \cup V_2 = V$ and $V_1 \cap V_2 = \emptyset$. Write the formula for the reduction in variance.

   (c) Example: You've always been told that drinking milk, getting plenty of sleep, eating your vegetables, and regularly exercising makes you grow up big and strong. Given your habits, you want to find out on average, how tall would you get? You ask your older friends whether they they did these things growing up and compile their answers in the following chart:

Sample #	DrinksMilk	SleepsWell	EatsVeggies	Height(cm)
1	0	1	1	200
2	0	1	0	210
3	0	1	0	200
4	1	1	0	180
5	1	0	1	130
6	1	0	0	150
9	1	1	1	?

   Using this data, you will construct a regression tree to tell how tall you will get.

   i. What is variance of Height?

   ii. Determine the first attribute to split on by determining which attribute gives you the most reduction in variance

   iii. What is the reduction of variance for the previous attribute?

   iv. Make the full decision tree with max depth 2. Draw the tree and indicate the split at each node and the average at each leaf.

   v. Now, determine how tall you would get.

3. **Multi-class Classification Least Squares** In this section, you will determine the parameter matrix $\mathbf{W} \in \mathbb{R}^{m \times p}$ for the Multi-class Least Squares classification.

   Given a data matrix $\mathbf{X} \in \mathbb{R}^{n \times m}$ and target matrix $\mathbf{T} \in \mathbb{R}^{n \times p}$, the sum-of-squares error function can be written as

   $$Er(\mathbf{W}) = \mathbf{Tr}\{(\mathbf{XW} - \mathbf{T})^T(\mathbf{XW} - \mathbf{T})\}$$

   where $\mathbf{Tr}$ is the trace of a matrix. You can assume that $\mathbf{X}$ has full rank.

   We will solve this problem by setting the derivative with respect to $\mathbf{W}$ to be zero and solve for $\mathbf{W}$. To do this we must first know some matrix derivative properties.

   (a) Let $\mathbf{A}, \mathbf{Z}$ be two matrices. Prove

   $$\frac{d\mathbf{Tr}(\mathbf{AZ})}{d\mathbf{Z}} = \mathbf{A}^T$$

   (b) Let $\mathbf{A}, \mathbf{Z}$ be two matrices. Prove

   $$\frac{d\mathbf{Tr}(\mathbf{ZAZ}^T)}{d\mathbf{Z}} = \mathbf{ZA}^T + \mathbf{ZA}$$

   (c) Now, we can take the derivative of $Er(\mathbf{W})$ and set it to zero. Show that this results in

   $$\mathbf{W} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{T}$$