ECE M146 Discussion 3 Solution
Introduction to Machine Learning
Instructor: Lara Dolecek
TA: Zehui (Alex) Chen

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., $H(IsPoisonous)$? **Solution**: Define the binary entropy function as follows:

$$H_b(p) = -p\log(p) - (1-p)\log(1-p).$$

$$H(IsPoisonous) = H_b(\frac{3}{8}) = -(\frac{5}{8}\log(\frac{5}{8}) + \frac{3}{8}\log(\frac{3}{8})) \approx 0.9544$$

(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:

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

Solution: Use the given equation, we get:

$$\begin{aligned}&H(IsPoisonous|IsColorful)\\ =&\frac{1}{2}H_b(\frac{3}{4}) + \frac{1}{2}H_b(1) \approx \frac{0.81127}{2} + 0 = 0.4056.\end{aligned}$$

(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.

Solution:

$$\begin{aligned} H(IsPoisonous|IsSmelly) &= \frac{2}{8}H_b(1) + \frac{6}{8}H_b(\frac{1}{2}) = 0.75. \\ H(IsPoisonous|IsSmooth) &= \frac{1}{2}H_b(\frac{3}{4}) + \frac{1}{2}H_b(\frac{1}{2}) \approx \frac{0.8113}{2} + \frac{1}{2} = 0.9056. \\ H(IsPoisonous|IsSmall) &= \frac{2}{8}H_b(1) + \frac{6}{8}H_b(\frac{1}{2}) = 0.75. \end{aligned}$$

(d) Calculate the information gain:

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

for

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

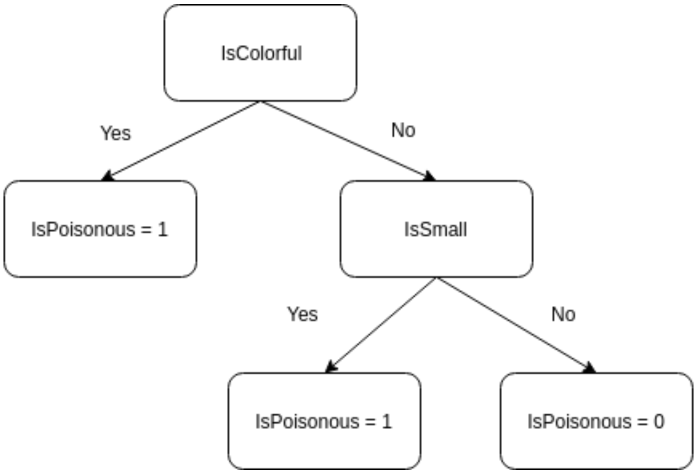
Solution:

$$\begin{aligned} I(IsPoisonous; IsColorful) &= 0.9544 - 0.4056 = 0.5488; \\ I(IsPoisonous; IsSmelly) &= 0.9544 - 0.75 = 0.2044; \\ I(IsPoisonous; IsSmooth) &= 0.9544 - 0.9056 = 0.1407; \\ I(IsPoisonous; IsSmall) &= 0.9544 - 0.75 = 0.2044. \end{aligned}$$

(e) Based on the information gain, determine the first attribute to split on. **Solution**: We choose IsColorful which has the largest information gain.

(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 IsPoisonous 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.

Solution:



After the first split, $H(IsPoisonous|IsColorful = 1) = 0$ so the tree stops growing on that branch. We are left with the samples that have $IsColorful = 0$ which is summarized in the following table. We notice that $H(IsPoisonous|IsSmall) = 0$ for this reduced set. Therefore splitting using the feature IsSmall maximize the information gain.

Sample #	IsSmelly	IsSmooth	IsSmall	IsPoisonous
1	0	0	1	1
2	0	0	0	0
5	0	1	0	0
6	0	1	0	0

(g) Is this mushroom poisonous? Not Poisonous!

2. **Properties of Information Gain**

(a) Show that $I(X;Y) = I(Y;X)$.
Solution:
Let's first rewrite $I(X;Y)$ based on the definition we learned in class.

$$\begin{aligned}
I(X;Y) &= H(X) - H(X|Y) \\
&= \sum_{x\in\mathcal{X}} p(x)\log\frac{1}{p(x)} - \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log\frac{1}{p(x|y)} \\
&= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log\frac{1}{p(x)} - \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log\frac{1}{p(x|y)} \\
&= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log\frac{p(x|y)}{p(x)} \\
&= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log\frac{p(x,y)}{p(x)p(y)} \\
&= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log\frac{p(y|x)}{p(y)} = I(Y;X).
\end{aligned}$$

(b) Using the fact that $\log(\cdot)$ is concave in its argument and the Jensen's Inequality:

If $f(x)$ is a convex function and X is a random variable:$E[f(X)] \geq f(E[X])$.

Show that $I(X;Y) \geq 0$.
Solution:

$$\begin{aligned}
I(X;Y) &= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log\frac{p(x,y)}{p(x)p(y)} \\
&= E\left[\log\frac{p(x,y)}{p(x)p(y)}\right] \\
&= -E\left[\log\frac{p(x)p(y)}{p(x,y)}\right] \\
&\geq -\log E\left[\frac{p(x)p(y)}{p(x,y)}\right] \quad \text{from Jensen's Inequality} \\
&= -\log\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\frac{p(x)p(y)}{p(x,y)} \\
&= -\log 1 = 0
\end{aligned}$$

(c) Prove that

$$I(X;Y) \leq \log(\min(|\mathcal{X}|, |\mathcal{Y}|)),$$

where $|\mathcal{X}|$ and $|\mathcal{Y}|$ are the cardinality of $\mathcal{X}$ and $\mathcal{Y}$, respectively.

Solution:

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ &\leq H(X) \\ &= \sum_{x \in \mathcal{X}} p(x) \log \frac{1}{p(x)} \\ &= E\left[\log(\frac{1}{p(x)})\right] \\ &\leq \log E\left[\frac{1}{p(x)}\right] \quad \text{from Jensen's Inequality} \\ &= \log |\mathcal{X}|. \end{aligned}$$

Similarly, we can get $I(X;Y) \leq \log |\mathcal{Y}|$ because $I(X;Y) = I(Y;X)$. Hence the proof.

3. Show that for a linearly separable data set, the maximum likelihood solution for the logistic regression model is obtained by finding a vector w whose decision boundary $w^T x = 0$ separates the classes and then taking the magnitude of w to infinity.

   Sketch of solution: If the dataset is linearly separable, then we can find w that for all points x_n belongs to class C_1, $w^T x_n > 0$; for all points x_m belongs to class C_2, $w^T x_m < 0$. According to the assumption of logistic regression, if we allow $|w| \to \infty$, for x_n belongs to C_1, $P(C_1|x_n, w) = \sigma(w^T x_n) \to 1$; for x_m belongs to C_2, $P(C_2|x_m, w) = 1 - \sigma(w^T x_m) \to 1$. This would maximize every term in the likelihood function and is therefore the ML solution.

   Hence, for a linearly separable dataset, the learning process may prefer to make $|w| \to \infty$ and use the linear boundary to label the datasets, which can cause severe over-fitting problem.

4. **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 $Var(V)$?

   Solution: Take the derivative with respect to u, set it to zero, and you get

   $$u = \sum_{x_i \in V} (x_i)/|V|$$

   (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.

   Solution:

   $$Var(V) - (\frac{|V_1|}{|V|} Var(V_1) + \frac{|V_2|}{|V|} Var(V_2))$$

   (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?

   Solution: First, we get the average which is 178.33 which results in variance of around 5083.33

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

   Solution:

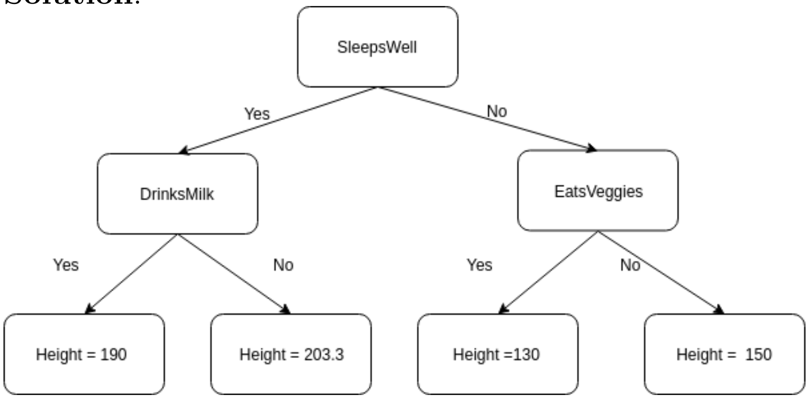
   - DrinksMilk: 666.67
   - SleepsWell: 383.33

- EatsVeggies: 2216.67

SleepsWell is the best attribute.

iii. What is the reduction of variance for the previous attribute?
Solution: 4700

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.
Solution:



v. Now, determine how tall you would get.
Solution:
190

Logistic Loss

$$\frac{1}{\log 2}\log(1+\exp(-y\hat{y})) = \begin{cases} \log(1+e^{-\hat{y}}) & y=1 \\ \log(1+e^{\hat{y}}) & y=-1 \end{cases}$$

Cross Entropy Loss

$$-y\log\sigma(\hat{y}) - (1-y)\log(\overset{\nearrow\ \sigma(-\hat{y})}{1-\sigma(\hat{y})})$$

$$= \begin{cases} -\log\frac{1}{1+e^{-\hat{y}}} & y=1 \\ -\log\frac{1}{1+e^{\hat{y}}} & y=0 \end{cases}$$