Announcements

• Midterm is graded! Average: 39, stdev: 6
• HW2 is out today
• HW2 is due Thursday, May 3, by 5pm in my mailbox
Decision Tree Classifiers

- Is $X_2 < 3$?
  - Yes
  - No
- Is $X_1 < 6$?
  - Yes
  - No
- Is $X_1 < 3$?
  - Yes
  - No
- Is $X_2 < 7$?
  - Yes
  - No
Decision Tree Classifiers

- **Root.** \( X_2 < 3 \)
  - **b.** \( X_1 < 3 \)
    - No
    - Yes
  - **a.** \( X_1 < 6 \)
    - No
    - Yes
  - **c.** \( X_2 < 7 \)
    - No
    - Yes
Decision Tree Classifiers

Each node is based on a **feature**

At each node, there is a **decision** based on the value of the feature

Each **leaf** is a **class** (same class may appear on multiple leaves)
An Example: Data

Iris data set
Features Petal Length (PL),
Petal Width (PW)
An Example: Data

Iris data set
Features Petal Length (PL), Petal Width (PW)

Is PW < 0.75?
Yes
No
An Example: Data

Iris data set
Features Petal Length (PL), Petal Width (PW)
An Example: Data

Iris data set
Features Petal Length (PL), Petal Width (PW)

- Is PW < 0.75?
  - Yes
  - No

- Is PL < 4.45?
  - Yes
  - No

- Is PW < 1.75?
  - Yes
  - No
An Example: Data

Iris data set
Features Petal Length (PL), Petal Width (PW)
In-Class Problem

Draw a decision tree for the following training data set:

( (0, 0), 1),    ( (0, 1), 0),    ( (1, 0), 0),    ( (1, 1), 0)
In-Class Problem

Draw a decision tree for the following training data set.

( (0, 0), 1),  ( (0, 1), 0),  ( (1, 0), 0),  ( (1, 1), 0)

Is $X_1 > 0.5$

Yes

No

Is $X_2 > 0.5$?

Yes

No
How to build a good decision tree?

1. How to choose a feature and a threshold value at the next node?
2. When do we stop?
Choosing a Feature (ID3 decision trees)

Choose a feature, and a threshold that reduces uncertainty the most
How do we measure uncertainty?
Measure of Uncertainty: Entropy

Let $X$ be a random variable (r.v) that takes values $v_1, \ldots, v_k$ with probabilities $p_1, \ldots, p_k$. Then, entropy of $X$, denoted by $H(X)$, is defined as:

$$H(X) = - \sum_{i=1}^{k} p_i \log p_i$$

**Example:**

1. $X$ is a 0/1 random variable, $P(X=1) = 0$. What is $H(X)$?
2. $X$ is a 0/1 random variable, $P(X=1) = 0.5$. What is $H(X)$?
3. $X$ is a r.v which takes values $1, \ldots, k$ with probabilities $1/k$ each. What is $H(X)$?
Measure of Uncertainty: Entropy

Let $X$ be a random variable (r.v) that takes values $v_1, \ldots, v_k$ with probabilities $p_1, \ldots, p_k$. Then, entropy of $X$, denoted by $H(X)$, is defined as:

$$H(X) = - \sum_{i=1}^{k} p_i \log p_i$$

Heatmap of entropy of a random variable with three values

Each point in triangle is a distribution $(p_1, p_2, p_3)$ where $p_1 + p_2 + p_3 = 1$
Conditional Entropy

Let $X$ and $Z$ be two random variables. The conditional entropy of $X$ given $Z$ is defined as:

$$H(X|Z) = \sum_z \Pr(Z = z) H(X|Z = z)$$

Essentially, what is the average entropy of $X$, given that we know $Z$.

**In-Class Problems:**

1. Suppose $X = Z$ (so $Z$ is a perfect predictor for $X$). What is $H(X|Z)$?
2. Suppose $X$ and $Z$ are independent. What is $H(X|Z)$?
Conditional Entropy

Let $X$ and $Z$ be two random variables. The conditional entropy of $X$ given $Z$ is defined as:

$$H(X|Z) = \sum_z \Pr(Z = z) H(X|Z = z)$$

Essentially, what is the average entropy of $X$, given that we know $Z$

Information Gain($Z$) = $H(X) - H(X|Z)$
Choosing a Feature (ID3 Decision Trees)

Suppose all features are discrete.

**Goal:** Choose a feature that reduces uncertainty the most

We choose the feature that maximizes information gain, where

\[ \text{Information Gain}(Z) = H(X) - H(X|Z) \]
Computing the Information Gain

**Root:** (5 R, 2 B)

\[
H(X) = -(\frac{5}{7} \log(\frac{5}{7}) - (\frac{2}{7} \log(\frac{2}{7}))
= 0.59
\]

**Node a:**

\[
\begin{align*}
H(X|Z=0) &= -(\frac{2}{2} \log(\frac{2}{2}) - (\frac{0}{2} \log(\frac{0}{2})) \\
H(X|Z=1) &= -(\frac{3}{5} \log(\frac{3}{5}) - (\frac{2}{5} \log(\frac{2}{5})) \\
P(Z=0) &= (2+0)/(5+2) = 2/7 \\
P(Z=1) &= (3+2)/(5+2) = 5/7
\end{align*}
\]

\[
H(X|Z) = P(Z=0)H(X|Z=0) + P(Z=1)H(X|Z=1) = 0.48
\]
Choosing a Feature: Example

Recall, Information Gain(Z) = H(X) - H(X|Z)

Root: (5 R, 2 B)
H(X) = 0.59

a: (2 R, 0 B), (3 R, 2 B)
H(X|Z) = 0.48

b: (4 R, 0 B), (1 R, 2 B)
H(X|Z) = 0.27

c: (5 R, 1 B), (0 R, 1 B)
H(X|Z) = 0.38

d: (2 R, 2 B), (3 R, 0 B)
H(X|Z) = 0.39

Pick solution b
Choosing a Feature: Example 2

Recall, Information Gain(Z) = H(X) - H(X|Z)

**Root:** (2 R, 2 B)
H(X) = 0.69

**a:** (1 R, 1 B), (1 R, 1 B)
H(X|Z) = 0.69

**b:** (1 R, 1 B), (1 R, 1 B)
H(X|Z) = 0.69

What is the information gain?
Choosing a Feature: Example 2

Recall, Information Gain(Z) = H(X) - H(X|Z)

Information gain is 0, but the correct tree is in the figure

**Lesson:** We shouldn’t necessarily stop if the information gain is 0
How to build a decision tree?

1. How to choose a feature and a threshold value at the next node?
2. When do we stop?
Stopping Rule 1

Stop when every node is pure (contains data from one class)
But this rule may overfit the training data
An Example

Decision tree if we stop when pure

But perhaps the blue point is an outlier or an error
So we **overfit** the training data
What is Overfitting?

Data (features and labels) come from some underlying true data distribution $D$. We never see the true distribution -- we only see samples (training sets, test sets, etc)

When we choose a classifier, we can talk about its training error:

$$\hat{err}(h) = \frac{1}{n} \sum_{i=1}^{n} 1(h(x_i) \neq y_i)$$

We can also talk about its true error:

$$err(h) = P_{(x,y) \sim D}(h(x) \neq y)$$

For a fixed classifier, as we get more and more data, the training error should approach the true error
As we make the tree more complicated, training error decreases. But at some point, true error stops improving and may even get worse!
Overfitting

True underlying distribution has some structure, which we would like to capture. Training data captures some of this structure, but may have some chance structure of its own, which we should try not to model into a classifier.

- True Decision Boundary
- Decision Boundary obtained from Training Data
How to Avoid Overfitting?

Avoid overfitting by **pruning** the tree using a **validation dataset**
How to Prune the Tree?

1. Split training data into training set S and validation set V

2. Build a tree T using training set S

3. Prune using V:
   
   Repeat:  
   
   For each node v in T:  
   
   Replace the subtree rooted at v by a single note that predicts the majority label in v (according to S) to get tree T′
   
   If the error of T′ on V is less than error of T on V, then: T = T′
   
   Continue until there is no such node v in T
If the blue point is an error, then the subtree at $v$ will hopefully have higher validation error than predicting red at $v$. 
As we make the tree more complicated, training error decreases. But at some point, true error stops improving and may even get worse! This will be reflected in the validation error, if we have sufficient validation data.