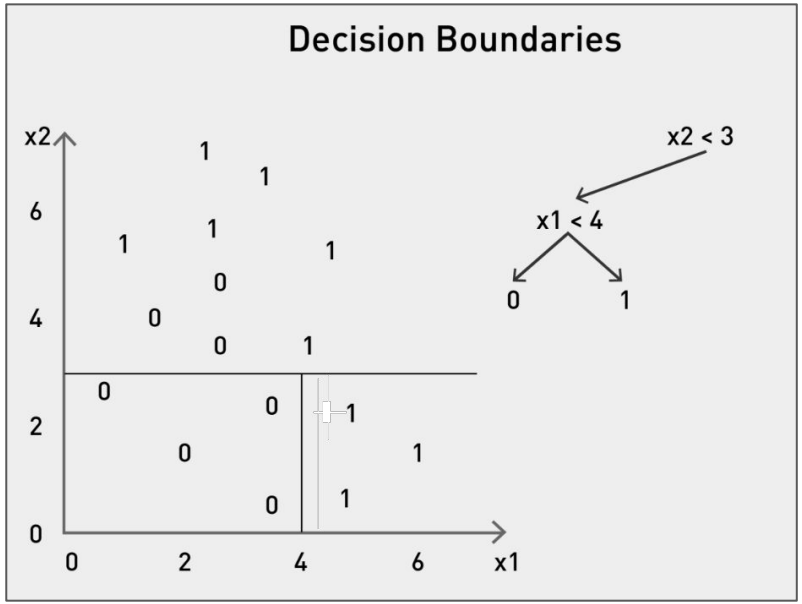
Week 4

Agenda

1. Warm up and breakout
2. Trees
3. Forests
4. Bias/variance tradeoff discussion

Warm up with Quizzes

1. Decision Trees can only represent linear boundaries
2. Which of these are base cases for the decision tree algorithm?
    a. completely pure split of the data
    b. Not data! Default to parent's majority class
    c. All possible splits have zero entropy
    d. All of the above



Warm up with Quizzes (2)

1. Select the most fitting answer: Use ___ with continuous features...
   a. discrete
   b. continuous
   c. thresholds
   d. algorithms
2. Select most fitting answer: ...and ___ for features with many attributes
   a. chaining
   b. binarization
   c. symbology
   d. properties
3. There are typically two points of randomness in decision forests. What are they?
   a. The first point is ___
   b. The second point is ___

Decision Trees

Tree Algorithm

GrowTree (S):

if y $== 0$ for all $< x, y >$ in S: return new leaf (0)

else if y $== 1$ for all $< x, y >$ in S: return new leaf (1)

else:

choose best attribute x_j

S_0 = all $< x, y >$ in S with x_j==0

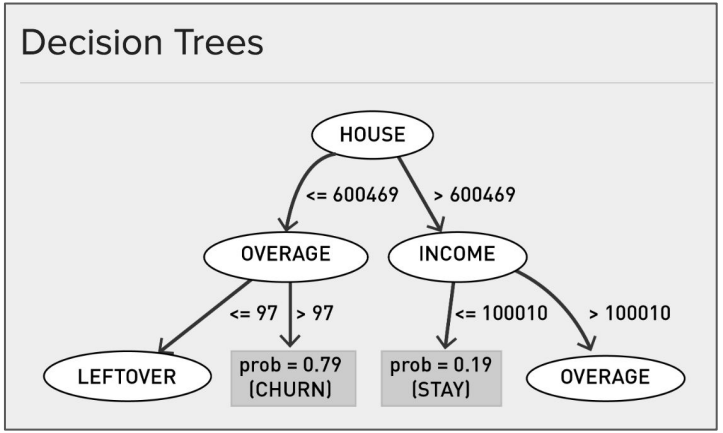
S_1 = all $< x, y >$ in S with x_j==1

return new node $(x_j$, GrowTree (S_0), GrowTree $(S_1))$

- Goal: Find a tree that is consistent with training examples.
- Strategy: Recursively choose most significant attribute as root of subtree.

Compare to KNN, Naive Bayes--

1. Describe the algorithm for training decision trees?
2. What's the training complexity?
   a. Can it be parallelized?
3. What's the prediction complexity?
4. Why is a tree sometimes said to be a 'white box'?



Decision Trees - online learner?

Tree Algorithm

GrowTree (S):

if y $== 0$ for all $< x, y >$ in S: return new leaf (0)

else if y $== 1$ for all $< x, y >$ in S: return new leaf (1)

else:

choose best attribute x_j

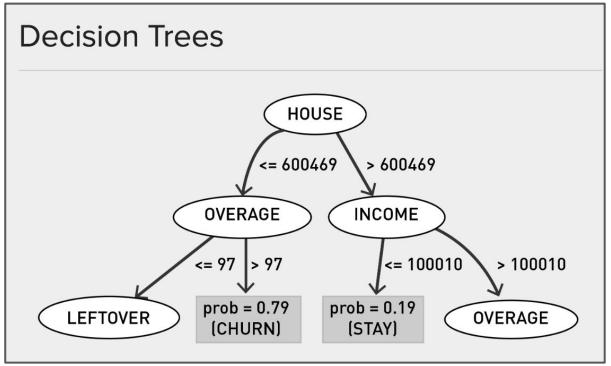
S_0 = all $< x, y >$ in S with x_j==0

S_1 = all $< x, y >$ in S with x_j==1

return new node $(x_j$, GrowTree (S_0), GrowTree $(S_1))$

- Goal: Find a tree that is consistent with training examples.
- Strategy: Recursively choose most significant attribute as root of subtree.

1. Recap: What is an online learner?
2. Is a decision tree an online learner?



Decision Trees - online learner?

Tree Algorithm

GrowTree (S):

if y $== 0$ for all $< x, y >$ in S: return new leaf (0)

else if y $== 1$ for all $< x, y >$ in S: return new leaf (1)

else:

choose best attribute x_j

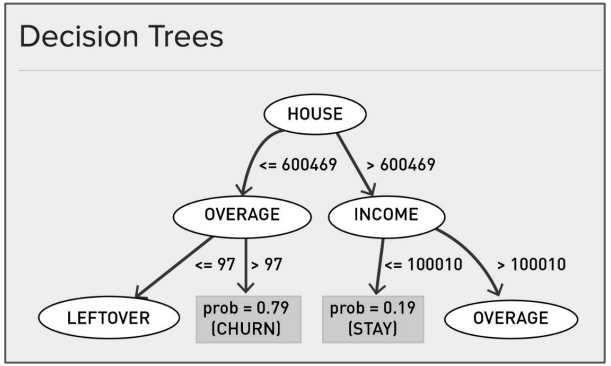
S_0 = all $< x, y >$ in S with x_j==0

S_1 = all $< x, y >$ in S with x_j==1

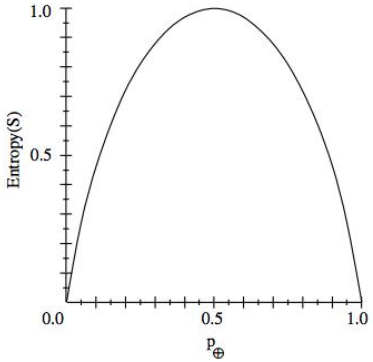
return new node $\left(x_j, \text{GrowTree } (S_0), \text{GrowTree } (S_1)\right)$

- Goal: Find a tree that is consistent with training examples.
- Strategy: Recursively choose most significant attribute as root of subtree.

Decision trees have incremental versions, e.g., VFDT, ID4 (Schlimmer & Fischer paper)



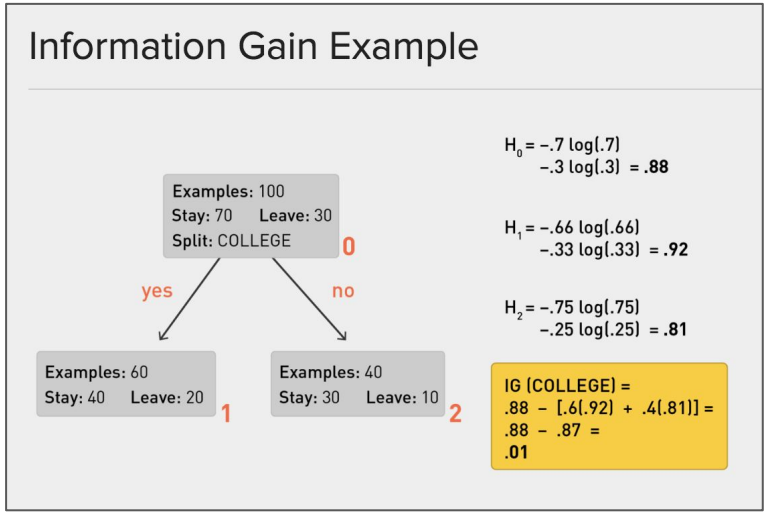
Decision Trees - Features



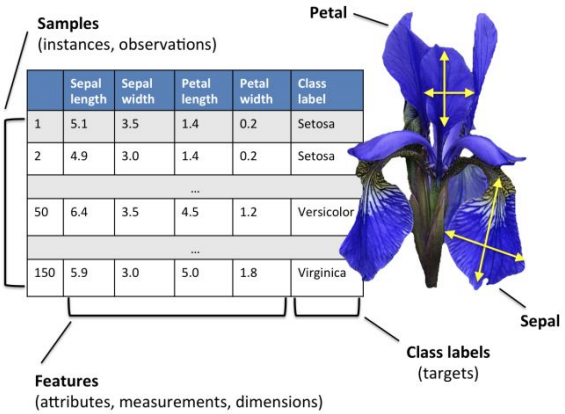
- S is a sample of training examples
- $p_{\oplus}$ is the proportion of positive examples in S
- $p_{\ominus}$ is the proportion of negative examples in S
- Entropy measures the impurity of S

$$Entropy(S) \equiv -p_{\oplus} \log_2 p_{\oplus} - p_{\ominus} \log_2 p_{\ominus}$$

1. Throwback: What is feature selection?
2. What is entropy? What is information gain?
3. How is entropy used for feature selection in decision trees?
4. Can it be used for feature selection with other algorithms (e.g. Naive Bayes)?



Ensembling

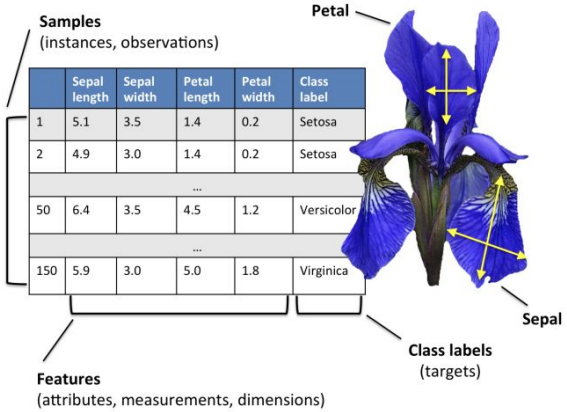


	Sepal length	Sepal width	Petal length	Petal width	Class label
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
...					
50	6.4	3.5	4.5	1.2	Versicolor
...					
150	5.9	3.0	5.0	1.8	Virginica

Random Forest Algorithm

- Generate artificial training set through bootstrapping with replacement.
- Build decision tree.
  - Randomly choose subset of features; consider those as split points.
  - No pruning!
- Repeat process to create multiple trees.
- Run test case through all trees.
- Predict by taking vote among trees.

- **Recap Bootstrapping (with replacement)**
- **Bagging**
  1. What is bagging?
  2. What is the intuition as to why bagging is effective?
  3. What is the complexity of training a bagging classifier? Can it be parallelized?
  4. What are different ways to combine predictions?



	Sepal length	Sepal width	Petal length	Petal width	Class label
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
...					
50	6.4	3.5	4.5	1.2	Versicolor
...					
150	5.9	3.0	5.0	1.8	Virginica

Boosting Classifier: AdaBoost Algorithm

1. Set weight for each training example = 1/n.
2. Train a classifier where objective respects the weights.
   - Run classifier over training examples.
3. Reduce weights for correct examples; increase weights for misclassified examples.
4. Return to step 2.
   - Second classifier is trained with objective that respects importance weights placed on each feature.

Boosting

1. What is boosting?
2. In terms of implementation, how do you incorporate example weights into the training of a tree?
3. How would you describe the complexity of training a boosting classifier? Is it parallelizable?
4. One of the best approaches -> often the winning "team" in Kaggle

Bias and Variance Errors

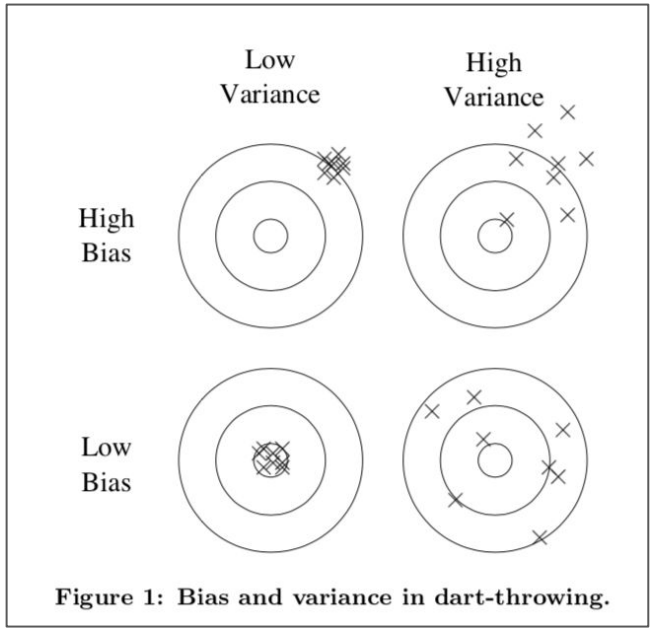


Figure 1: Bias and variance in dart-throwing.

Bias / Variance Tradeoff

1. What do bias and variance intuitively refer to?
2. Why is there typically a tradeoff?

Discuss when boosting or bagging may reduce variance and when bias