L7: Tree Models

Shan Wang

Lingnan College, Sun Yat-sen University

2020 Data Mining and Machine Learning LN3119

https://wangshan731.github.io/DM-ML/

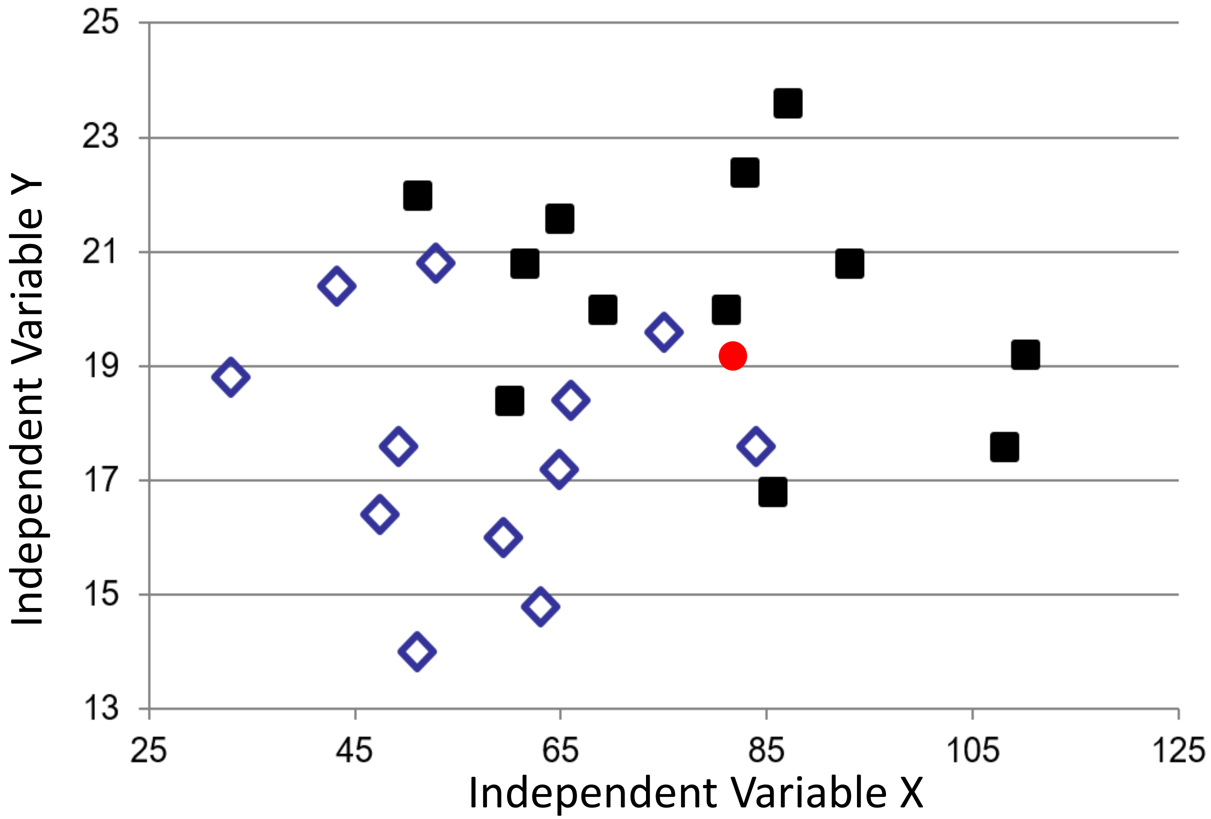
Last lecture

- Linear SVM
  - Model: $y = f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \begin{cases} +1, & \text{if } \boldsymbol{\theta}'\boldsymbol{x} + \theta_0 \geq 0 \\ -1, & \text{if } \boldsymbol{\theta}'\boldsymbol{x} + \theta_0 < 0 \end{cases}$
  - Strategy: maximize margin
  - Algorithm: SMO algorithm
- Regularization
  - Soft margin
- Kernels:
  - Mapping feature vectors to a different space
- Application: diabetes care revisit

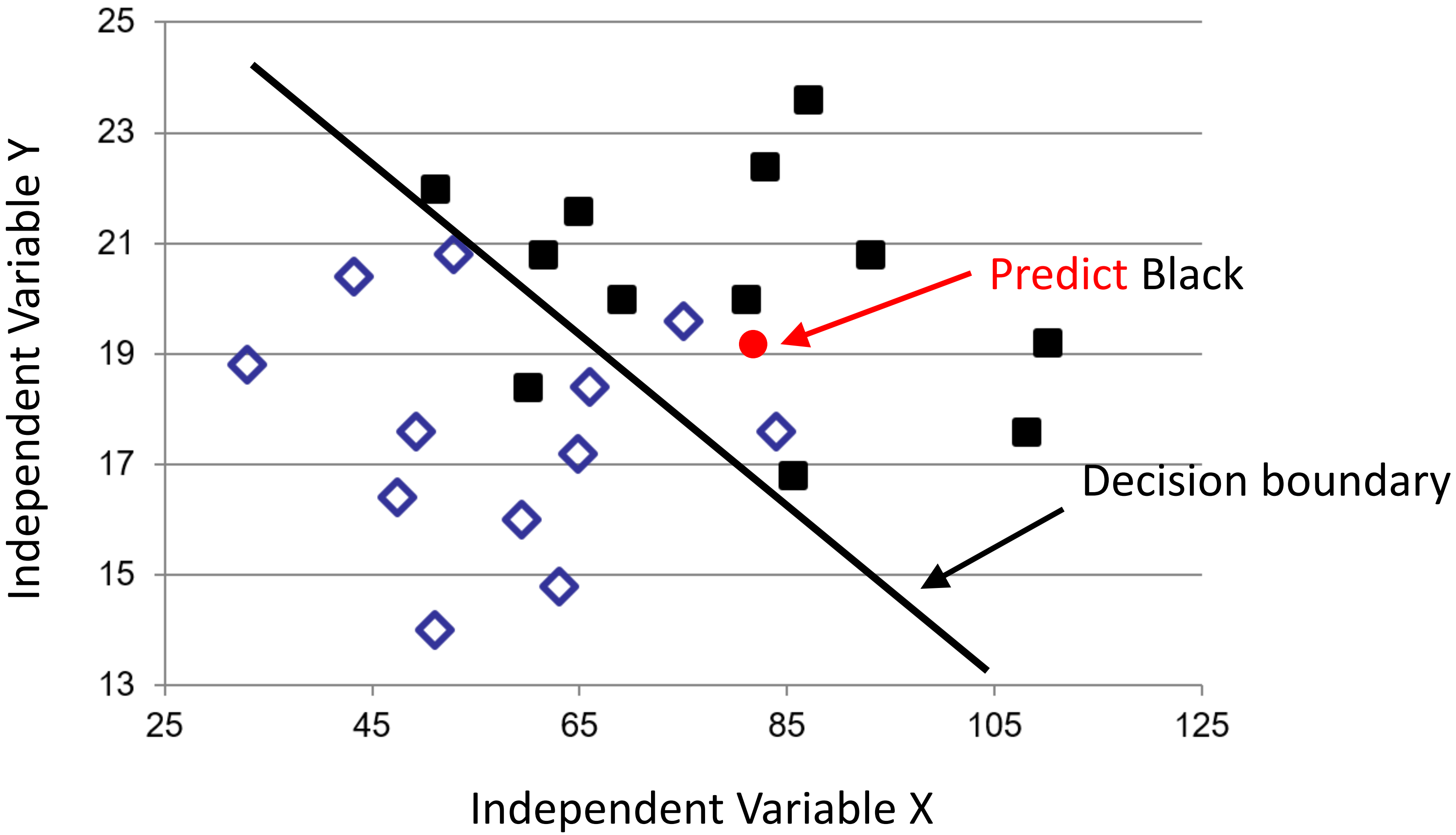
The nature of logistic regression and SVM

- A linear decision boundary in the (mapped) feature space
  - Generally **not** interpretable
  - Do not give a simple explanation of how decision is made

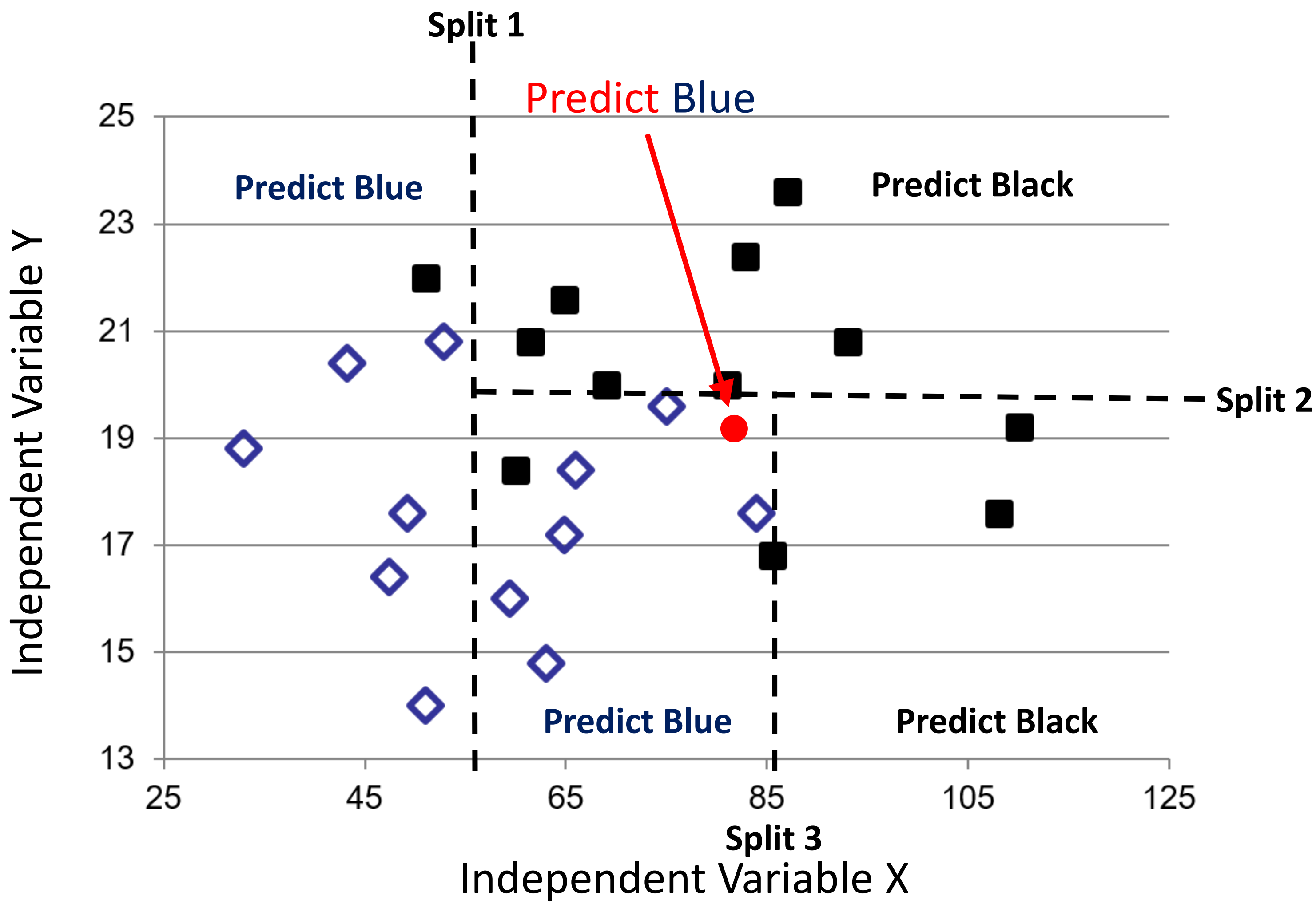
Example: Predict Black or Blue?



Logistic regression or SVM



Another thought



Course outline

- Supervised learning
  - Linear regression
  - Logistic regression
  - SVM and kernel
  - Tree models
- Unsupervised learning
  - Clustering
  - PCA
  - EM
- Deep learning
  - Neural networks
  - Convolutional NN
  - Recurrent NN
- Reinforcement learning
  - MDP
  - ADP
  - Deep Q-Network

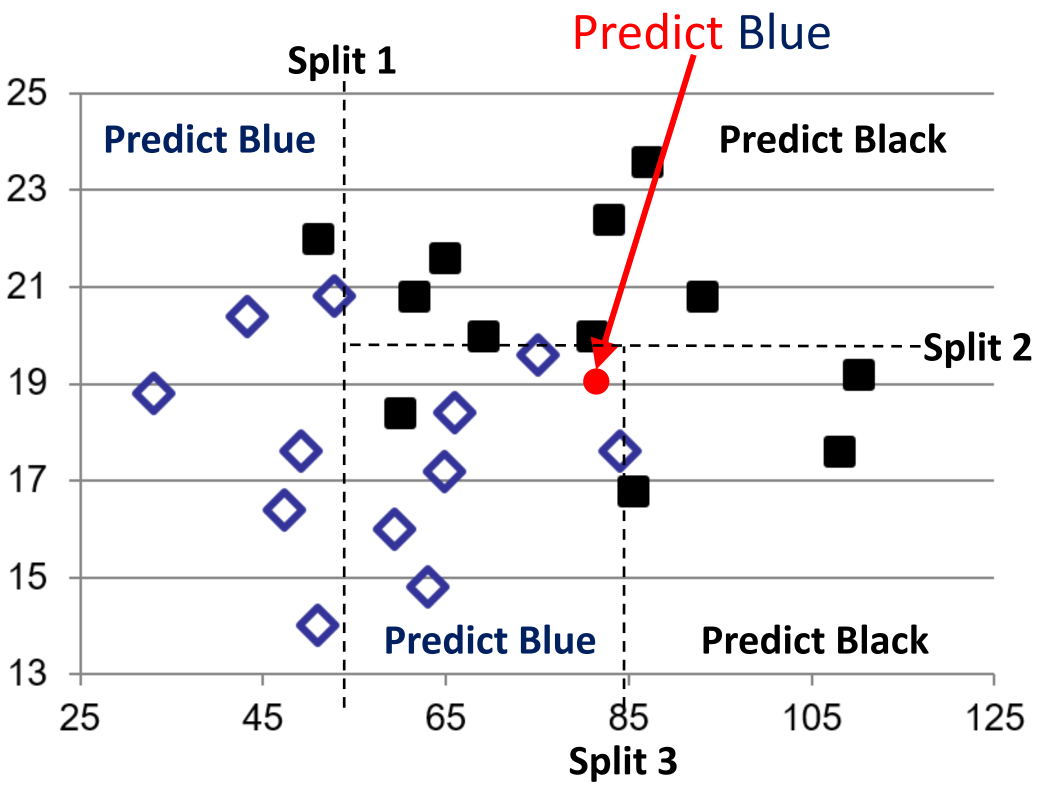
This lecture

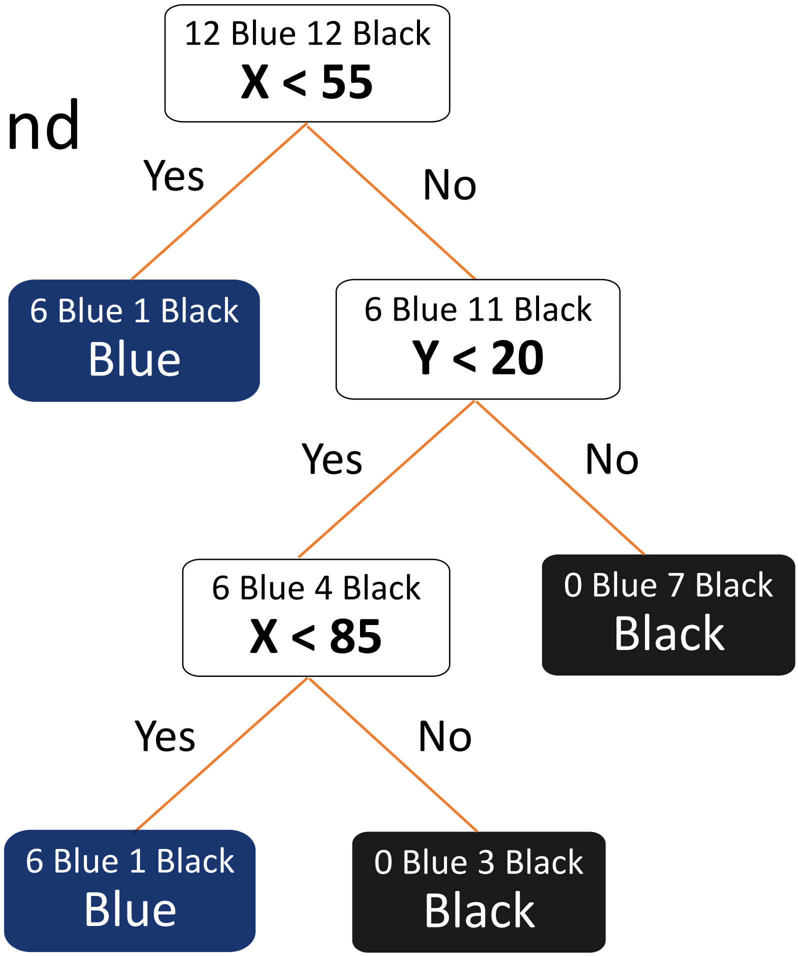
- Decision Tree Model
- Strategy & Algorithm
  - ID.3
  - C4.5
  - CART
- Regularization
- Random forest
- Application: Supreme Court Decisions

Reference: CS420, Weinan Zhang (SJTU), OPIM 326 Daniel Zheng (SMU)

Decision tree

- Build a tree by splitting on independent variables
- To predict the outcome for an observation, follow the splits and at the end…





Decision tree

- Tree components
  - Intermediate node (a feature) for splitting data
  - Leaf node (a class) for label prediction

- Key questions for decision trees
  - How to select node splitting conditions?
  - How to make prediction?
  - How to decide the tree structure?

Model

Model

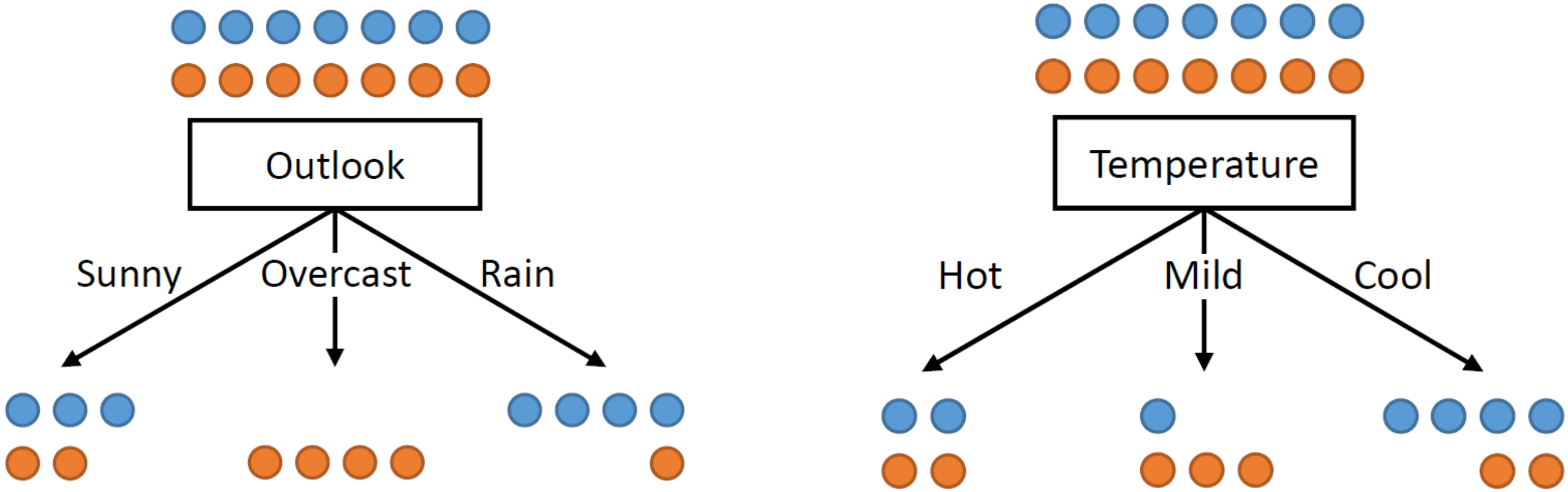
- Problem setting
  - Instance feature space X
  - Instance label space Y
  - Unknown underlying function (target): $f: X \rightarrow Y$
  - Set of function hypothesis $H = \{h | h\colon\ X \rightarrow Y\ \}$
    - Here each hypothesis h is a decision tree
- Input: training data generated from the unknown
  - $\{(\boldsymbol{x}_i, y_i)\} = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_N, y_N)\}$
- Output: a hypothesis $h \in H$ that best approximates

Strategy and algorithm

ID.3

Node splitting

- Which node splitting condition to choose?

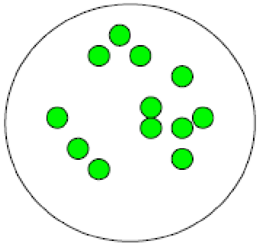


- Choose the features with higher classification capacity
  - Quantitatively, with higher information gain ???

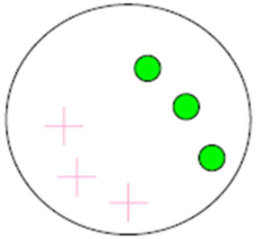
Entropy

- Entropy
  - A measure of the uncertainty
  - Suppose random variable Y has k^{th} possible value with probability p_k
    - Entropy: $H(Y) = -\sum_{k=1}^{K} p_k \log p_k$
- Lager entropy has higher uncertainty
  - What is the entropy of a group in which all examples belong to the same class?
    - $H(Y) = -1 \log 1 = 0$
    - Deterministic
  - What is the entropy of a group with same probability in each class?
    - $H(Y) = \log K$
    - Uniform

Minimum uncertainty



Maximum uncertainty



Conditional Entropy

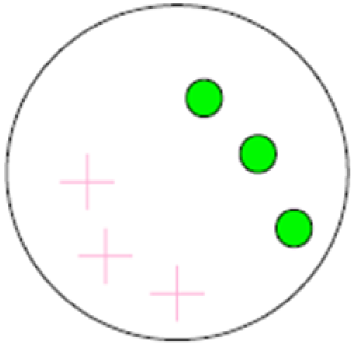
- Given random variable X, Y, suppose random variable Y has i^{th} possible value with probability p_i, and random variable X has j^{th} possible value with probability p_j, and the joint probability is p_{ij}
  - Denote $H(Y|X=x_j) = -\sum_{i=1}^{I} \frac{p_{ij}}{p_j} \log \frac{p_{ij}}{p_j}$ as the entropy of Y when X takes value x_j
- Conditional entropy

$$H(Y|X) = \sum_{j=1}^{J} p_j H(Y|X=x_j)$$

  - Representing the uncertainty of Y under the condition X

Example

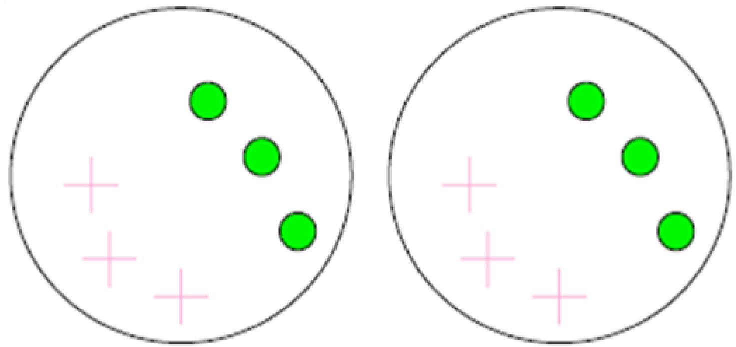
Before



$H(Y) = \log 2$

Which one bring more information? X_1 or X_2?

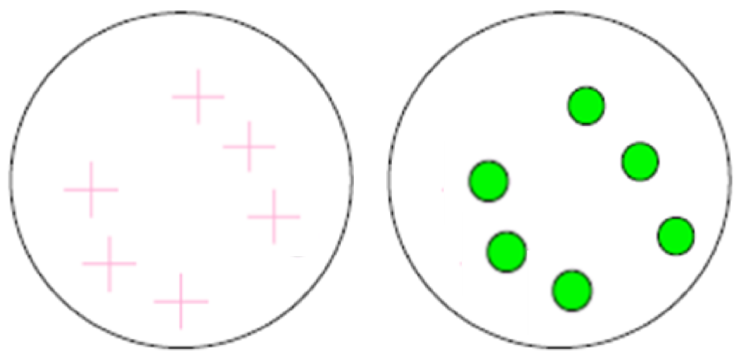
With condition X_1



$X_1 = 1$ $X_1 = 2$

$H(Y|X_1) = \log 2$

With condition X_2



$X_2 = 1$ $X_2 = 2$

$H(Y|X_2) = 0$

Information gain

- Information gain represents, with the feature information X, how much the uncertainty of label Y decreases

$$G(Y,X) = H(Y) - H(Y|X)$$

- The larger information gain, the stronger classification capability of the feature

- The idea of ID.3
  - Every time, choose the feature with largest information gain as the node splitting condition

Example

- Given a dataset of 8 students about whether they like the famous movie *Gladiator*, calculate the entropy in this dataset

- $H(Like) = -\frac{4}{8}\log\frac{4}{8} - \frac{4}{8}\log\frac{4}{8} = \log 2 = 1$

Like
Y
N
Y
N
N
Y
N
Y

Example (cont.)

- Suppose we now also know the gender of these 8 students, what is the conditional entropy on gender?
- The labels are divided into two small dataset based on the gender

Gender	Like
M	Y
F	N
M	Y
F	N
F	N
M	Y
M	N
F	Y

Like-M
Y
Y
Y
N

Like-F
N
N
N
Y

$P(Yes|Male) = 0.75$ $P(Yes|Female) = 0.25$

Example (cont.)

- Suppose we now also know the gender of these 8 students, what is the conditional entropy on gender?

- $P(Yes|Male) = 0.75; \quad P(Yes|Female) = 0.25$
- $H(Like|Male) = -0.25\log(0.25) - 0.75\log(0.75) = 0.81$
- $H(Like|Female) = 0.81$

- $H(Like|Gender) = \Pr(Male) \times H(Like|Male) + \Pr(Female) \times H(Like|Female) = 0.81$

- $G(Like, Gender)$
  $= H(Like) - H(Like|Gender)$
  $= 1 - 0.81 = 0.19$

Gender	Like
M	Y
F	N
M	Y
F	N
F	N
M	Y
M	N
F	Y

Example (cont.)

- Suppose we now also know the major of these 8 students, what is the conditional entropy on gender?
- The labels are divided into two small dataset based on the gender

Like-M
Y
N
N
Y

Like-H
N
N

Like-C
Y
Y

$P(Yes|Math) = 0.5$ $P(Yes|Econ) = 0$ $P(Yes|CS) = 1$

Major	Like
Math	Y
Econ	N
CS	Y
Math	N
Math	N
CS	Y
Econ	N
Math	Y

Example (cont.)

- Suppose we now also know the major of these 8 students, what is the conditional entropy on gender?

- $P(Yes|Math) = 0.5, P(Yes|Econ) = 0, P(Yes|CS) = 1$
- $H(Like|Math) = -0.5\log(0.5) - 0.5\log(0.5) = 1$
- $H(Like|Econ) = 0$
- $H(Like|CS) = 0$

- $H(Like|Major) = \Pr(Math) \times H(Like|Math) + \Pr(Econ) \times H(Like|Econ) + \Pr(CS) \times H(Like|CS) = 0.5 \times 1 = 0.5$

- $$G(Like, Major) \\ = H(Like) - H(Like|Major) \\ = 1 - 0.5 = 0.5$$

Major	Like
Math	Y
Econ	N
CS	Y
Math	N
Math	N
CS	Y
Econ	N
Math	Y

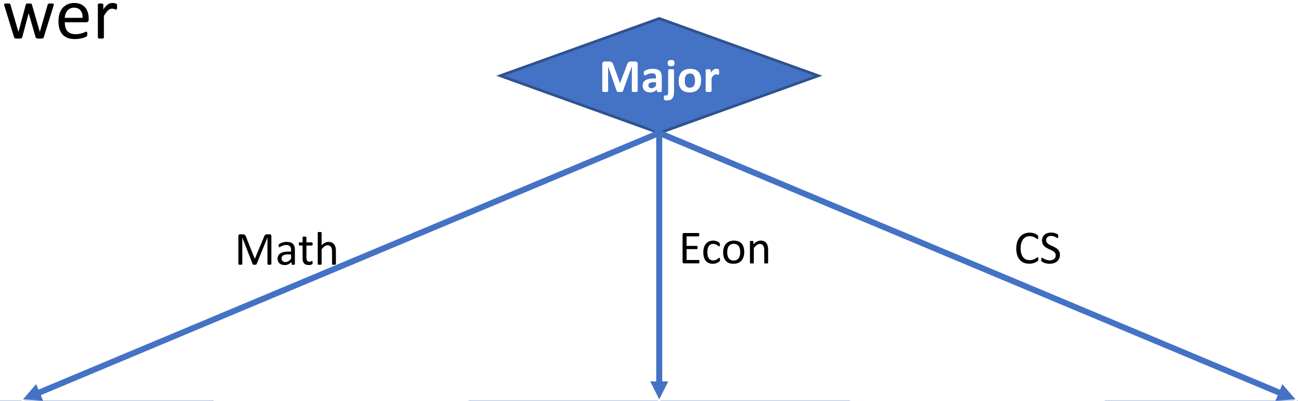
Example (cont.)

- Compare Major and Gender
- $G(Like, Gender)$
$$= H(Like) - H(Like|Gender)$$
$$= 1 - 0.81 = 0.19$$
- $G(Like, Major)$
$$= H(Like) - H(Like|Major)$$
$$= 1 - 0.5 = 0.5$$
- Major is the better feature to predict the label "like"

Gender	Major	Like
M	Math	Y
F	Econ	N
M	CS	Y
F	Math	N
F	Math	N
M	CS	Y
M	Econ	N
F	Math	Y

Example (cont.)

- Major is used as the decision condition and it splits the dataset into three small one based on the answer

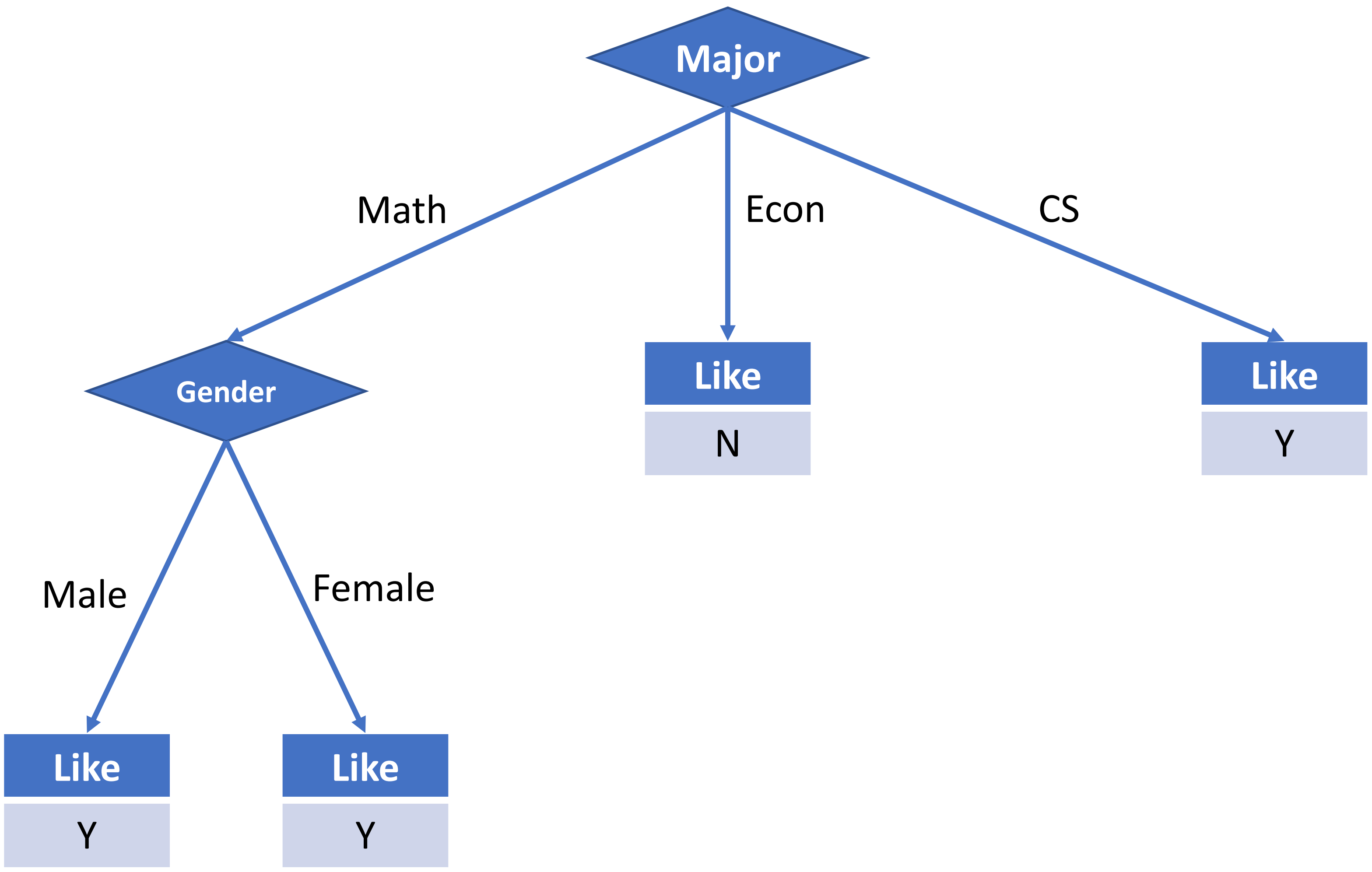


Gender	Like
M	Y
F	N
F	N
F	Y

Gender	Like
F	N
M	N

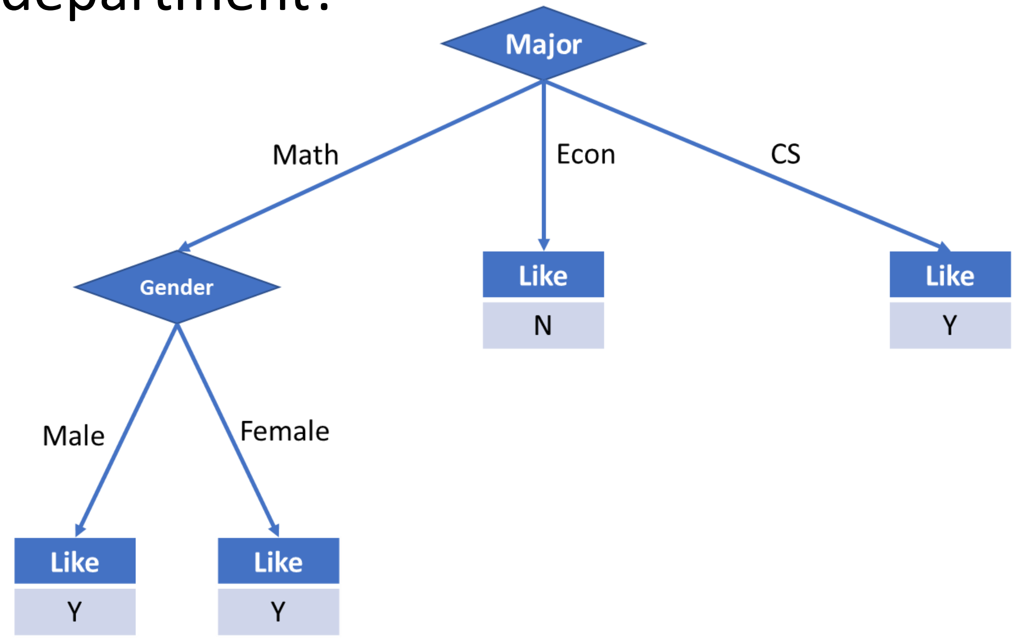
Gender	Like
M	Y
M	Y

Example (cont.)



Example (cont.)

- In the stage of testing, suppose there come a female students from the CS department, how can we predict whether she like the movie Gladiator?
  - Based on the major of CS, we will directly predict she like the movie.
  - What about a male student and a female student from math department?



ID.3 algorithm

- Algorithm framework
  - Start from the root node with all data
  - For each node, calculate the information gain of all possible features
  - Choose the feature with the highest information gain
  - Split the data of the node according to the feature
  - Do the above recursively for each leaf node, until
    - There is no (**enough**) information gain for the leaf node
    - Or there is no feature to select
- Testing
  - Pass the example through the tree to the leaf node for a label

Strategy and algorithm

C4.5

C4.5 algorithm

- The algorithm framework of C4.5 is similar to ID.3
  - The only difference is the criteria of C4.5 is information gain ratio
- In ID.3, we use information gain as the feature selection criteria
  - The feature which takes more possible values has advantages
- Information gain ratio will correct this problem
  - $G(Y,X) = H(Y) - H(Y|X)$
  - $G_R(Y,X) = \frac{G(Y,X)}{H(X)}$

Strategy and algorithm

CART

CART

- ***Classification and Regression Trees (CART)***
  - Binary Tree
    - Each node represents whether a feature satisfies some condition, the left branch is “yes”, and the right branch is “no”
  - Can repeatedly use the same feature (with different splitting)
  - It can handle both of discrete label and continuous label
    - Classification tree
    - Regression tree

CART: Classification tree

- The generation of a classification tree is similar to ID.3
  - At each node, for all possible values of all features, select the one with smallest Gini Index
- Gini index
  - In a data set D, there are K classes, and D_k is the subset of data belonging to k^{th} class. The Gini index of D

$$Gini(D) = 1 - \sum_{k=1}^{K} \left(\frac{|D_k|}{|D|}\right)^2$$

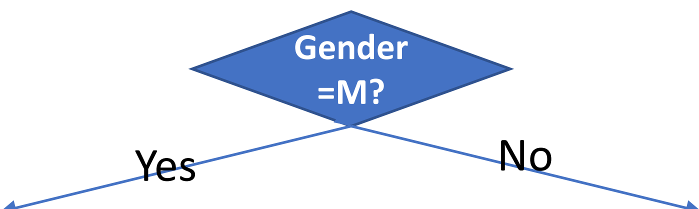
  - The smaller Gini index, the smaller uncertainty of the data set
    - Deterministic: $Gini(D) = 0$
    - Uniform (2-class): $Gini(D) = \frac{1}{2}$

CART: Classification tree (cont.)

- Gini index of data set D under condition $X = x$
  - Let D_1 denote the set of data which satisfy the condition
  - Let D_2 denote the set of remaining data
- The Gini index under condition $X = x$ is
  - $Gini(D, X = x) = \frac{|D_1|}{|D|} Gini(D_1) + \frac{|D_2|}{|D|} Gini(D_2)$
- At each node, select the feature condition with smallest Gini index
- The condition
  - Discrete feature: $X = x$
  - Continuous feature: $X < x$

Example

- $Gini(D, Gender = M) = \frac{4}{8}\left(1 - \left(\frac{3}{4}\right)^2 - \left(\frac{1}{4}\right)^2\right) + \frac{4}{8}\left(1 - \left(\frac{3}{4}\right)^2 - \left(\frac{1}{4}\right)^2\right) = 0.375$
- $Gini(D, Major = Math) = \frac{4}{8}\left(1 - \left(\frac{2}{4}\right)^2 - \left(\frac{2}{4}\right)^2\right) + \frac{4}{8}\left(1 - \left(\frac{2}{4}\right)^2 - \left(\frac{2}{4}\right)^2\right) = 0.5$
- $Gini(D, Major = Econ) = \frac{2}{8}\left(1 - \left(\frac{2}{2}\right)^2\right) + \frac{6}{8}\left(1 - \left(\frac{2}{6}\right)^2 - \left(\frac{4}{6}\right)^2\right) = 0.44$
- $Gini(D, Major = CS) = \frac{2}{8}\left(1 - \left(\frac{2}{2}\right)^2\right) + \frac{6}{8}\left(1 - \left(\frac{2}{6}\right)^2 - \left(\frac{4}{6}\right)^2\right) = 0.44$
- Pick $(Gender = M)$ as the first splitting condition



Major	Like
Math	Y
CS	Y
CS	Y
Econ	N

Major	Like
Econ	N
Math	N
Math	N
Math	Y

Gender	Major	Like
M	Math	Y
F	Econ	N
M	CS	Y
F	Math	N
F	Math	N
M	CS	Y
M	Econ	N
F	Math	Y

CART: Regression tree

- The output of regression tree is the predicted value
  - The selection criteria of node splitting condition is Squared Error
- Given a condition
  - Discrete feature: $X = x$
  - Continuous feature: $X < x$
- The ta set D is divided into two sub sets D_1 and D_2
  - D_1 satieties the condition while D_2 does not
  - $SE(D, X = x) = \sum_{i \in D_1} (y_i - \bar{y}_{D_1})^2 + \sum_{i \in D_2} (y_i - \bar{y}_{D_2})^2$
    - Where $\bar{y}_{D_1} = \frac{1}{|D_1|} \sum_{i \in D_1} y_i$, $\bar{y}_{D_2} = \frac{1}{|D_2|} \sum_{i \in D_2} y_i$
- At each node, find the condition $X = x$ or $X < x$, which minimizes $SE(D, X = x)$ or

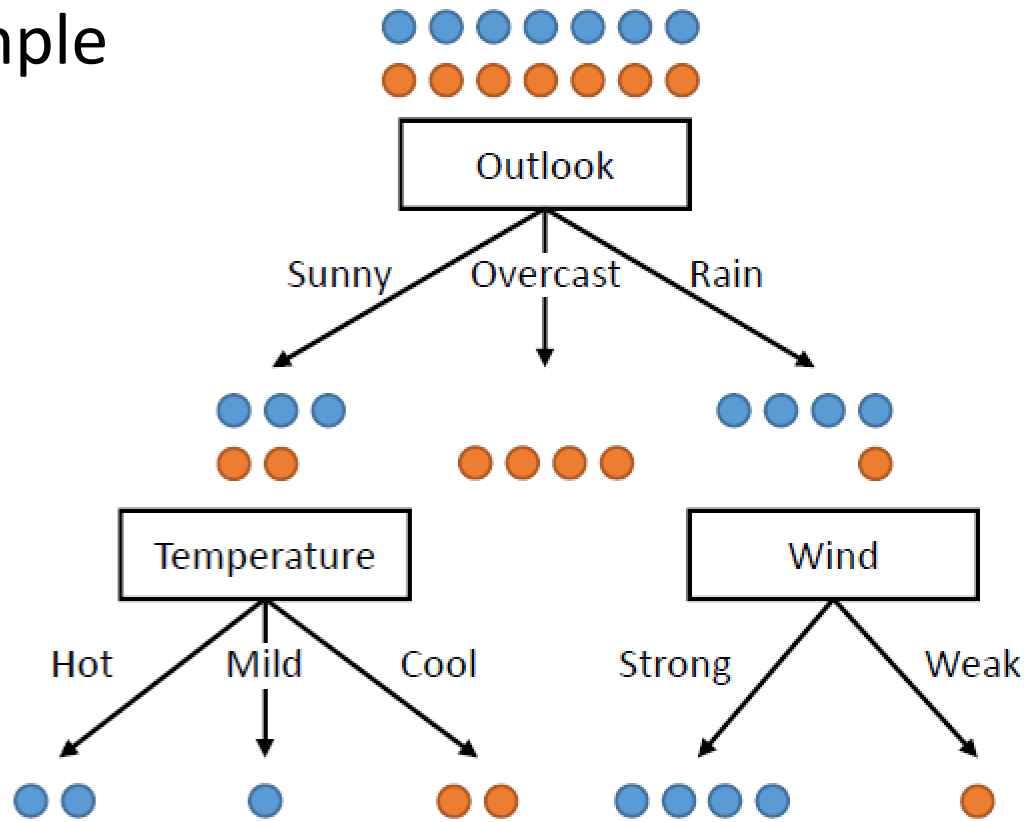
Regularization

Overfitting

- Tree model can approximate any finite data by just growing a leaf node for each instance
  - A very deep tree
- It will result in overfitting, and the generated tree is lack of generalization ability
- How to solve this problem?
  - Pruning, a kind of regularization
  - i.e., control the tree size while ensure the prediction ability

Overfitting (cont.)

- An example



- How about this tree, yielding perfect partition?

Overfitting (cont.)

- How to solve this problem?
  - Pruning, a kind of regularization
  - i.e., control the tree size while ensure the prediction ability
- The cost function to minimize takes two parts into account
  - The prediction error
  - The complexity of the tree

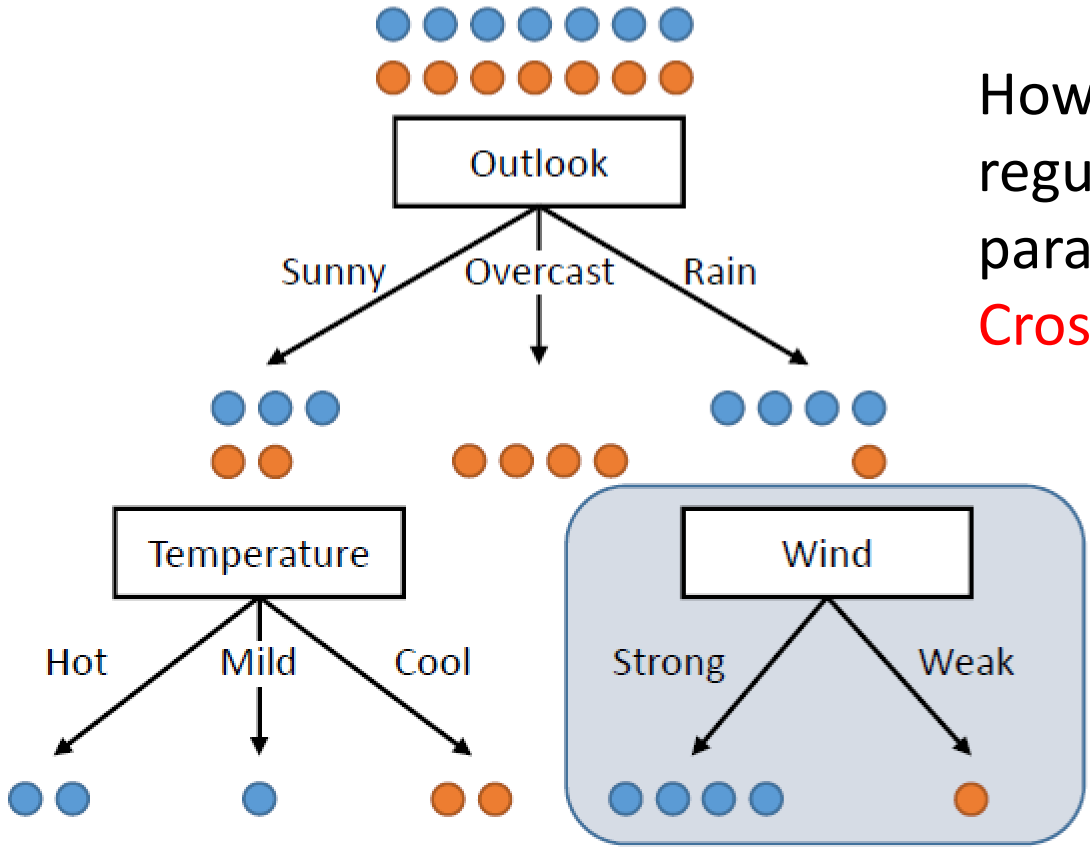
Regularization

$$R_\lambda(T) = R(T) + \lambda|T|$$

- T represents the tree model, $|T|$ is the number of left nodes
- $R(T)$ is the empirical error of the tree model T
  - For example, we can use empirical entropy or Gini index
- Empirical entropy
  - $R(T) = \sum_{t=1}^{|T|} |D_t|\, H(D_t)$
    - D_t is the data set at left node t; $H(D_t)$ is its empirical entropy
- Gini index
  - $R(T) = \sum_{t=1}^{|T|} |D_t|\, Gini(D_t)$

Pruning

- Whether to prune the "Wind" node?
  - Calculate the $R_\lambda(T)$ difference of the trees with/without "Wind" node



How to select regularization parameter λ?

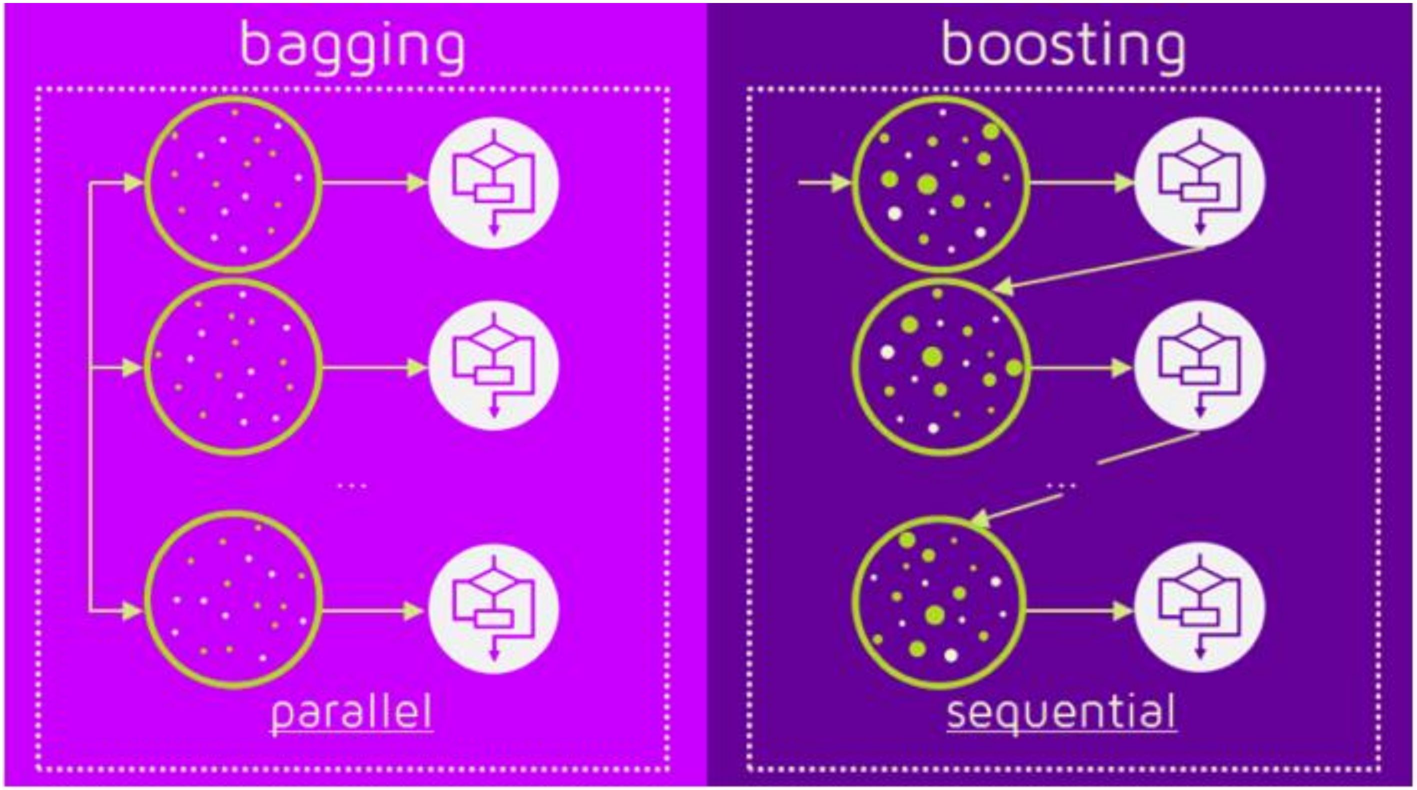
Cross validation!

Random Forest

Ensemble learning

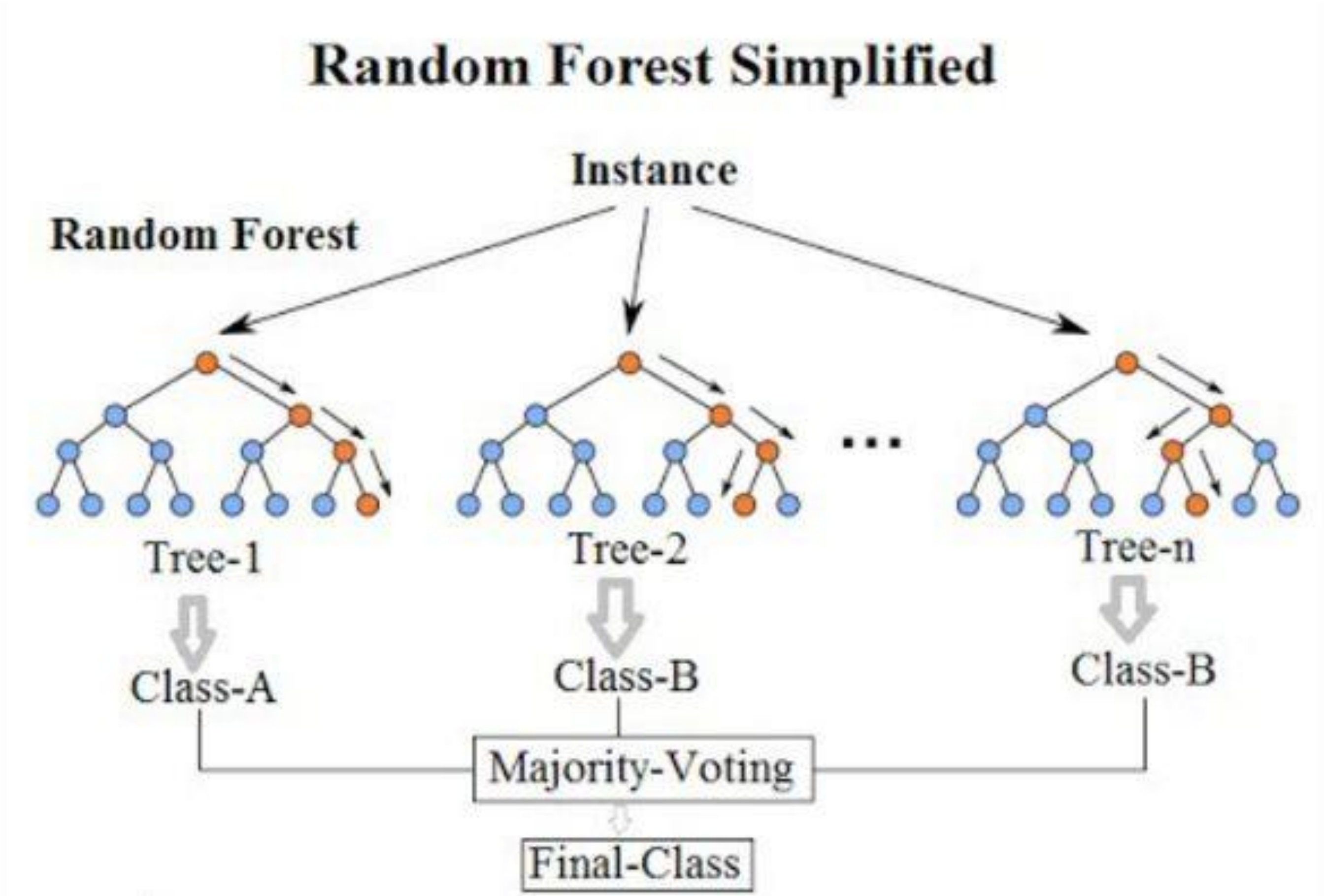
- **Ensemble learning**
  - multiple models are strategically generated and combined
  - converts weak learners to strong learners
  - two main methods: boosting and bagging
- Boosting
  - train weak learners sequentially
  - each learner tries to correct its predecessor
- Bagging (**B**ootstrap **agg**regat**ing**)
  - combining the results of multiple parallel models
  - Bootstrapping the data sets

Bagging v.s. Boosting



Random forest

- Random forest
  - A kind of bagging on decision tree



Random forest (cont.)

- Data set size: N, feature number M
- What data used to train each tree?
  - A bootstrap sample of size N from training data
  - Samples are repeatedly drawn, with **replacement**
- What features considered in each node splitting
  - Randomly select $m < M$ features
  - Pick the best feature & split-point among the m features
- Output predicted value
  - Regression: average
  - Classification: majority vote

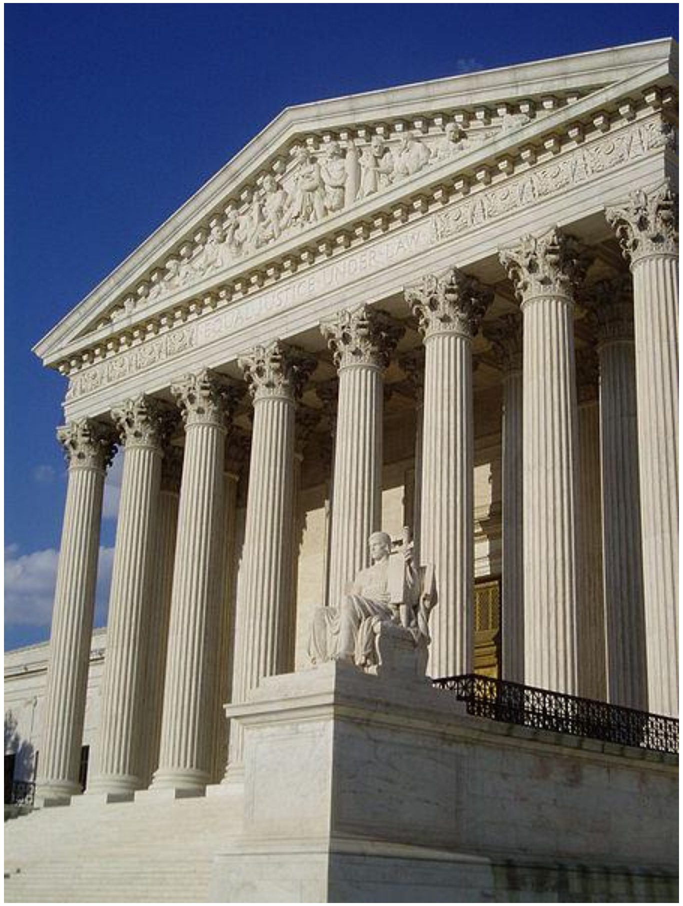
How many trees we need?
How many features to select?

Application

Supreme Court Decisions

The American Legal System

- The legal system of the United States operates at the state level and at the federal level
- Federal courts hear cases beyond the scope of state law
- Federal courts are divided into:
    - District Courts
        - Makes initial decision
    - Circuit Courts
        - Hears appeals from the district courts
    - Supreme Court
        - Highest level – makes final decision



The Supreme Court of the United States

- Consists of nine justices, appointed by the President
- Decides the most difficult and controversial cases
  - Often involve interpretation of Constitution
  - Significant social, political and economic consequences

How a Case Gets to the US Supreme Court

Notable Decisions

- Wickard v. Filburn (1942)
  - Congress allowed to intervene in industrial/economic activity
- Roe v. Wade (1973)
  - Legalized abortion
- Bush v. Gore (2000)
  - Decided outcome of presidential election!
- National Federation of Independent Business v. Sebelius (2012)
  - Patient Protection and Affordable Care Act ("ObamaCare") upheld the requirement that individuals must buy health insurance

Predicting Supreme Court Decisions

- Legal academics and political scientists regularly make predictions of Supreme Court decisions from detailed studies of cases and individual justices
- In 2002, Andrew Martin, a professor of political science at Washington University in St. Louis, decided to instead predict decisions using a statistical model built from data
- Together with his colleagues, he decided to test this model against a panel of experts

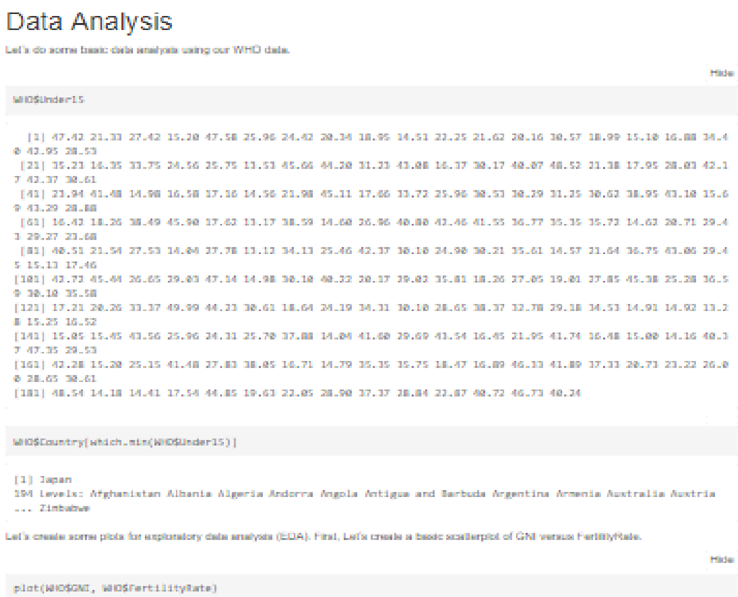
Data

- Cases from 1994 through 2001
- In this period, same nine justices presided
    - Breyer, Ginsburg, Kennedy, O'Connor, Rehnquist (Chief Justice), Scalia, Souter, Stevens, Thomas
    - Rare data set—longest period of time with the same set of justices in over 180 years
- We will focus on predicting Justice Stevens' decisions
    - Started out moderate, but became more liberal
    - Self-proclaimed conservative

Variables

- Dependent Variable
  - Did Justice Stevens vote to reverse the lower court decision?
  - 1 = reverse, 0 = affirm
- Independent Variables: Properties of the case
  - Circuit court of origin (1^{st}–11^{th}, DC, FED)
  - Issue area of case (e.g., civil rights, federal taxation)
  - Type of petitioner, type of respondent (e.g., US, an employer)
  - Ideological direction of lower court decision (conservative or liberal)
  - Whether petitioner argued that a law/practice was unconstitutional

Let’s get our hands dirty!



Data Analysis

Let's do some basic data analysis using our WHO data.

Hide

```
WHO$Under15
```

```
  [1] 47.42 21.33 27.42 15.20 47.58 25.96 24.42 20.34 18.95 14.51 22.25 21.62 20.16 30.57 18.99 15.10 16.88 34.4
0 42.95 28.53
 [21] 35.23 16.35 33.75 24.56 25.75 13.53 45.66 44.20 31.23 43.08 16.37 30.17 40.07 48.52 21.38 17.95 28.03 42.1
7 42.37 30.61
 [41] 23.94 41.48 14.98 16.58 17.16 14.56 21.98 45.11 17.66 33.72 25.96 30.53 30.29 31.25 30.62 38.95 43.10 15.6
9 43.29 28.88
 [61] 16.42 18.26 38.49 45.90 17.62 13.17 38.59 14.60 26.96 40.80 42.46 41.55 36.77 35.35 35.72 14.62 20.71 29.4
3 29.27 23.68
 [81] 40.51 21.54 27.53 14.04 27.78 13.12 34.13 25.46 42.37 30.10 24.90 30.21 35.61 14.57 21.64 36.75 43.06 29.4
5 15.13 17.46
[101] 42.72 45.44 26.65 29.03 47.14 14.98 30.10 40.22 20.17 29.02 35.81 18.26 27.05 19.01 27.85 45.38 25.28 36.5
9 30.10 35.58
[121] 17.21 20.26 33.37 49.99 44.23 30.61 18.64 24.19 34.31 30.10 28.65 38.37 32.78 29.18 34.53 14.91 14.92 13.2
8 15.25 16.52
[141] 15.05 15.45 43.56 25.96 24.31 25.70 37.88 14.04 41.60 29.69 43.54 16.45 21.95 41.74 16.48 15.00 14.16 40.3
7 47.35 29.53
[161] 42.28 15.20 25.15 41.48 27.83 38.05 16.71 14.79 35.35 35.75 18.47 16.89 46.33 41.89 37.33 20.73 23.22 26.0
0 28.65 30.61
[181] 40.54 14.18 14.41 17.54 44.85 19.63 22.05 28.90 37.37 28.84 22.87 40.72 46.73 40.24
```

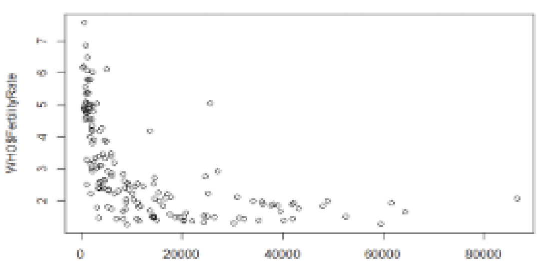
```
WHO$Country[which.min(WHO$Under15)]
```

```
[1] Japan
194 Levels: Afghanistan Albania Algeria Andorra Angola Antigua and Barbuda Argentina Armenia Australia Austria
... Zimbabwe
```

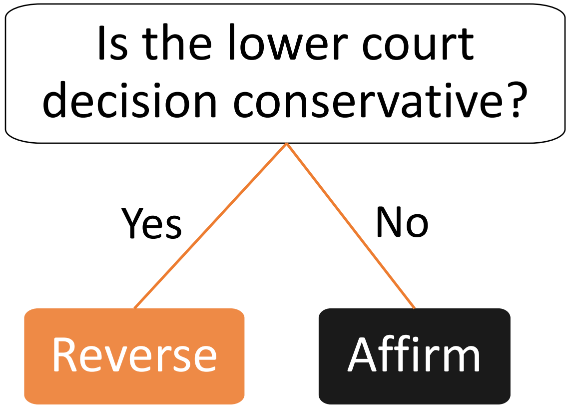
Let's create some plots for exploratory data analysis (EDA). First, Let's create a basic scatterplot of GNI versus FertilityRate.

Hide

```
plot(WHO$GNI, WHO$FertilityRate)
```



Final Tree for Justice Stevens

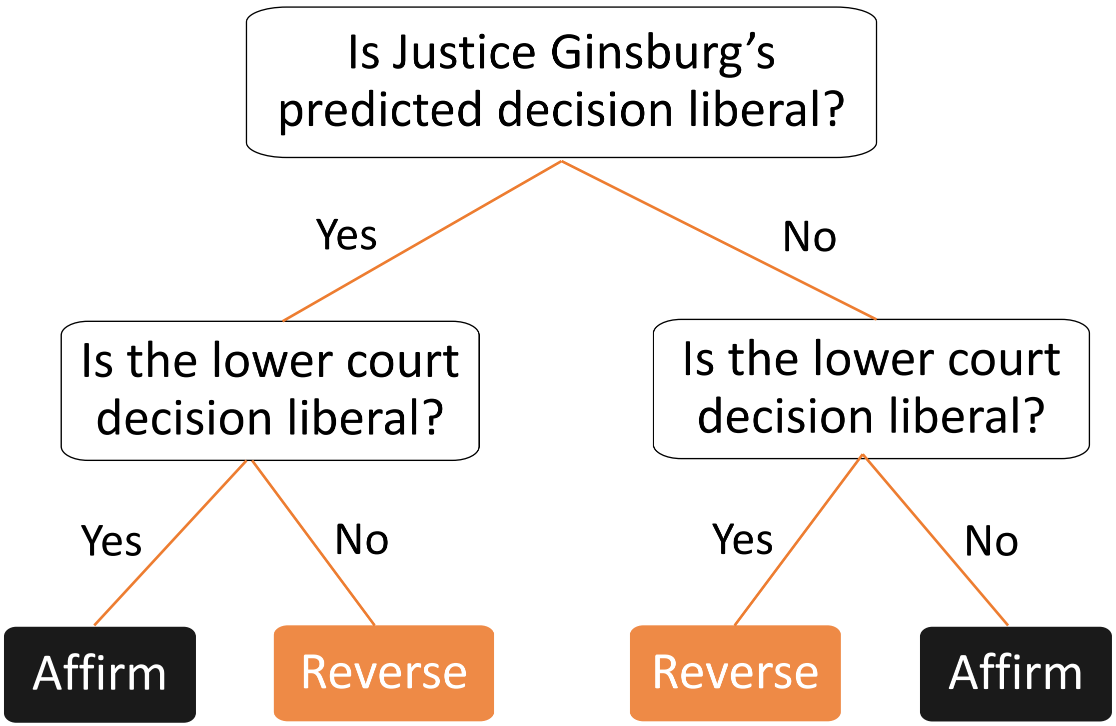


- Self-proclaimed conservative
- Was his claim supported by this tree model?

Martin's Model

- Used 628 previous Supreme Court cases that occurred between 1994 and 2001
- Made predictions for 68 cases that would be decided in the October 2002 term, before it started
- Two stage approach based on CART:
  - First stage: one tree to predict a unanimous liberal decision, other tree to predict unanimous conservative decision
    - Around 50% of cases resulted in a unanimous decision
    - If conflicting predictions or predict no, move to next stage
  - Second stage consists of predicting decision of each individual justice, and using majority decision as prediction

Tree for Justice Souter



"Make a liberal decision" "Make a conservative decision"

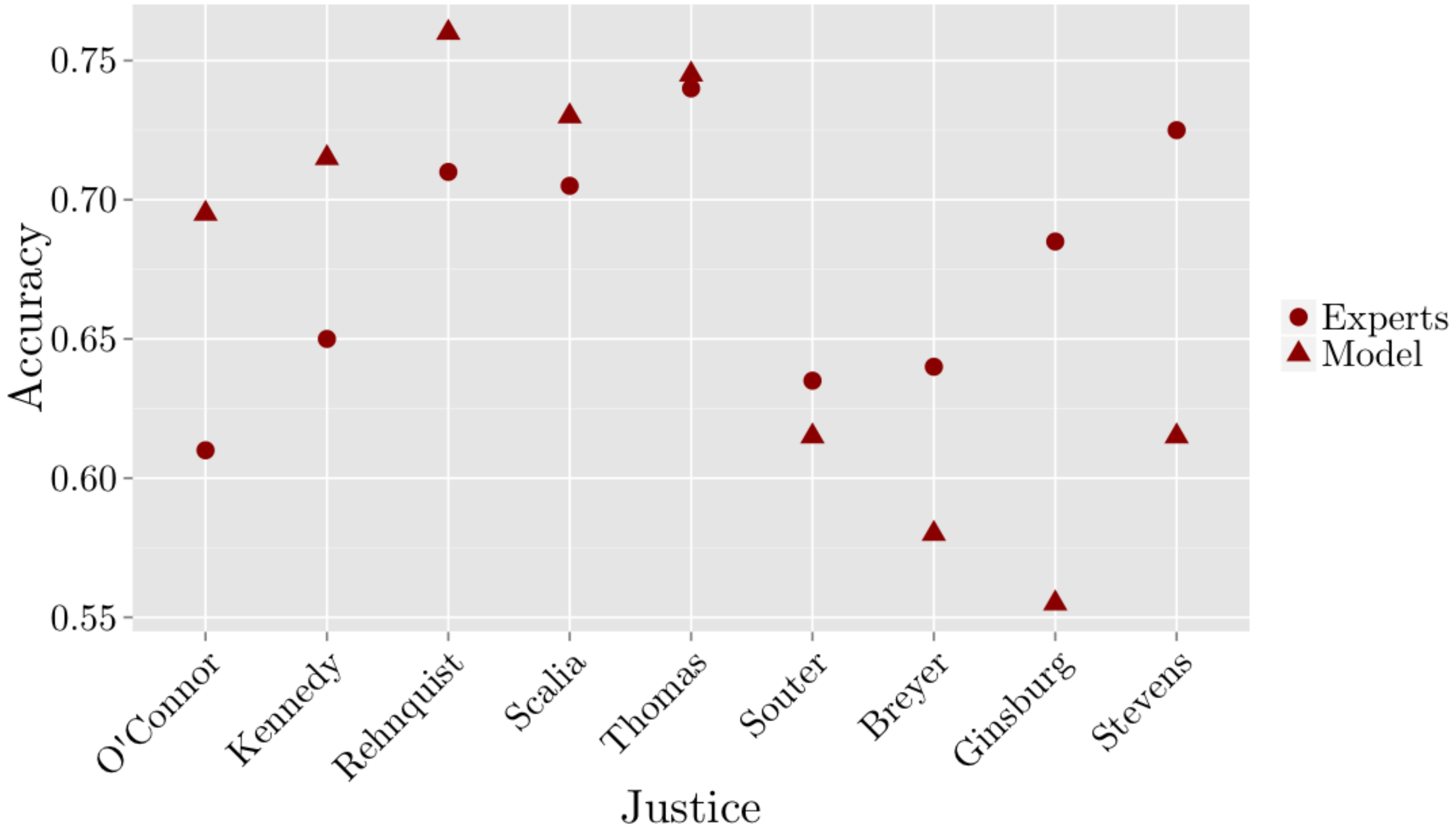
The Experts

- Martin and his colleagues recruited 83 legal experts
  - 71 academics and 12 attorneys
  - 38 previously clerked for a Supreme Court justice, 33 were chaired professors and 5 were current or former law school deans
- Experts only asked to predict within their area of expertise; more than one expert to each case
- Allowed to consider any source of information, but not allowed to communicate with each other regarding predictions

The Results

- For the 68 cases in October 2002
- Overall case predictions
  - Model accuracy: 75%
  - Experts accuracy: 59%
- Individual justice predictions
  - Model accuracy: 67%
  - Experts accuracy: 68%

TheIndividual Justice Predictions Results



Expert vs. Analytics

- Predicting Supreme Court decisions is very valuable to firms, politicians and non-governmental organizations
- A model that predicts these decisions is both more accurate and faster than experts
  - CART model based on very high-level details of case beats experts who can process much more detailed and complex information

Lecture 7 wrap-up

- ✓ Decision Tree Model
- ✓ Strategy & Algorithm
  - ✓ ID.3
  - ✓ C4.5
  - ✓ CART
- ✓ Regularization
- ✓ Random forest
- ✓ Application: Supreme Court Decisions

Next lecture

- Supervised learning
  - Linear regression
  - Logistic regression
  - SVM and kernel
  - Tree models
- Unsupervised learning
  - Clustering
  - PCA
  - EM
- Deep learning
  - Neural networks
  - Convolutional NN
  - Recurrent NN
- Reinforcement learning
  - MDP
  - ADP
  - Deep Q-Network

Questions?

Shan Wang (王杉)

https://wangshan731.github.io/