Random Forest

Yao-Yi Chiang
Computer Science and Engineering
University of Minnesota
yaoyi@umn.edu

Slides adopted from Machine Learning 10601, Recitation 8, Oct 21, 2009 Oznur Tastan
(http://people.sabanciuniv.edu/otastan/)

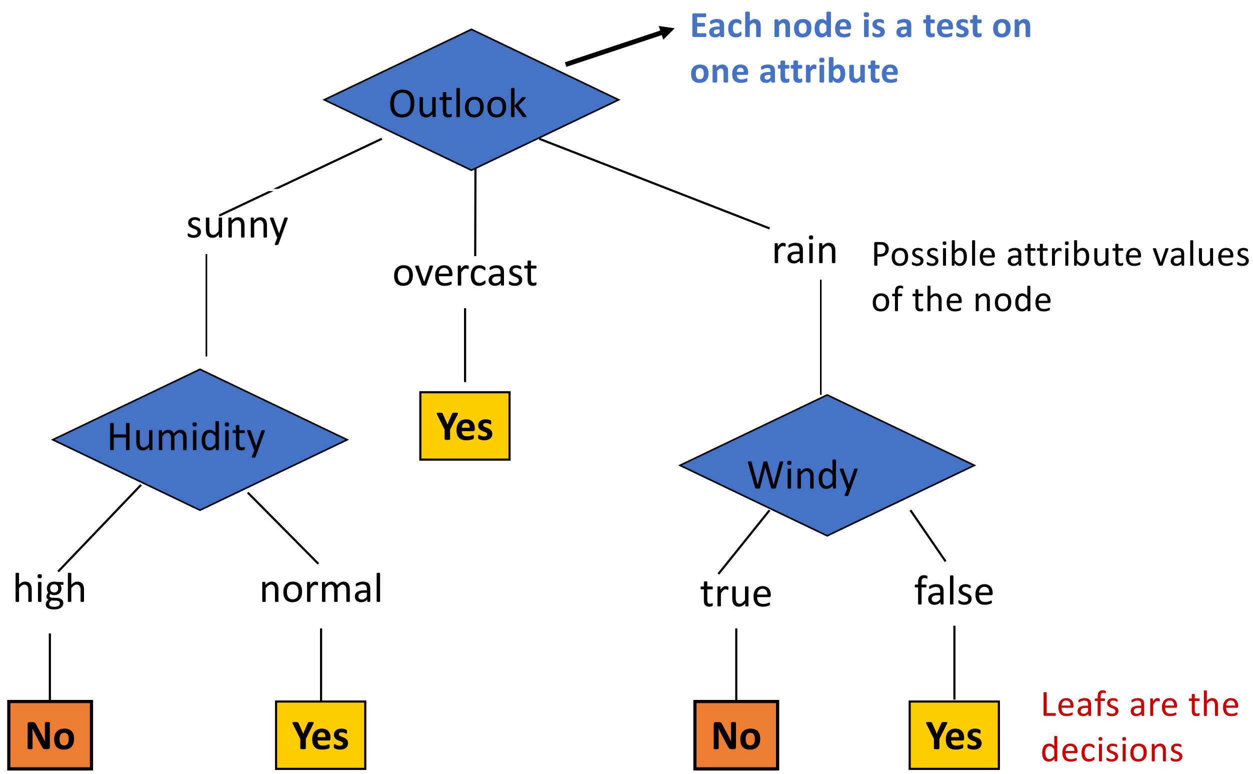
Outline

- Tree representation
- Brief information theory
- Learning decision trees
- Bagging
- Random forests

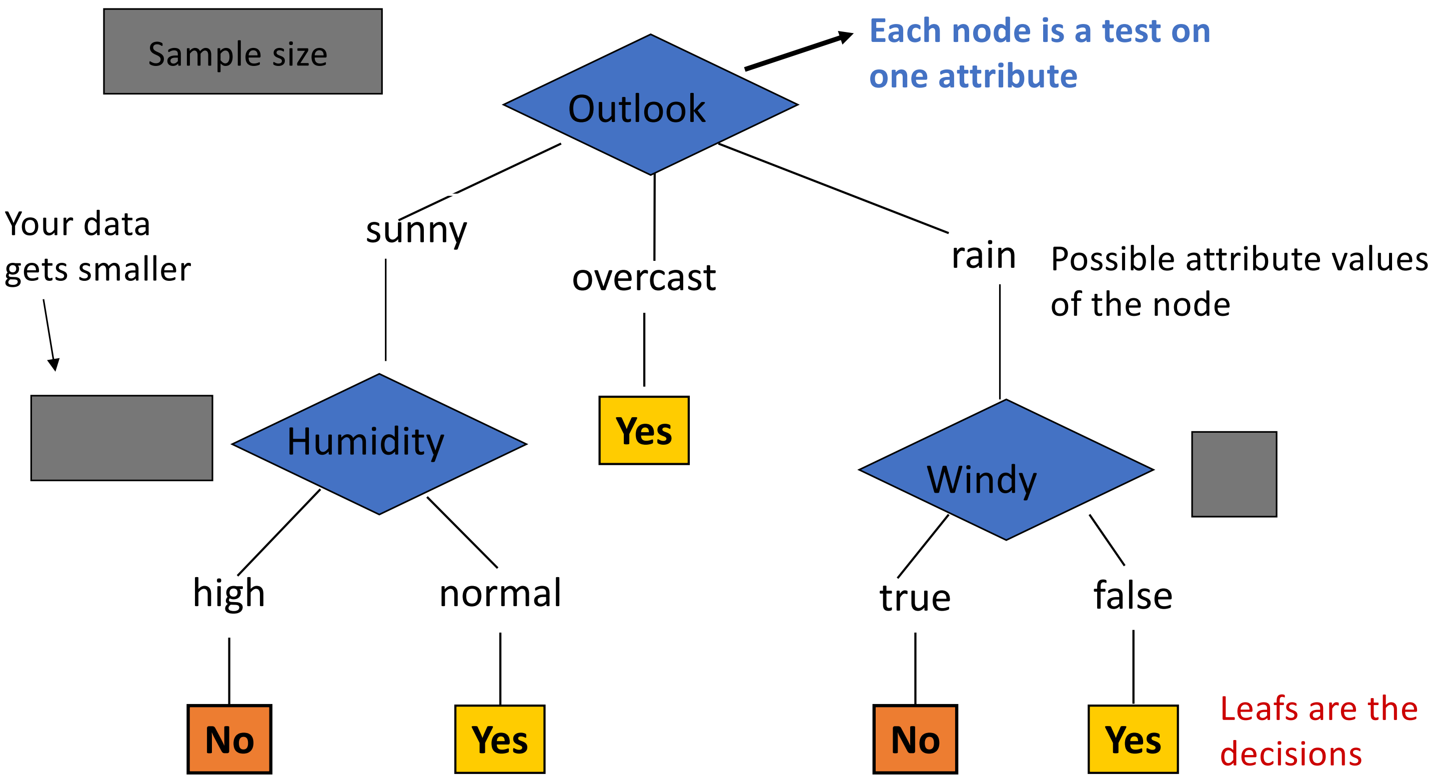
Decision trees

- Non-linear classifier & regressor
- Easy to use
  - Can handle both numerical and categorial variables
- Easy to interpret
- Non-parametric method
- Susceptible to overfitting but can be avoided

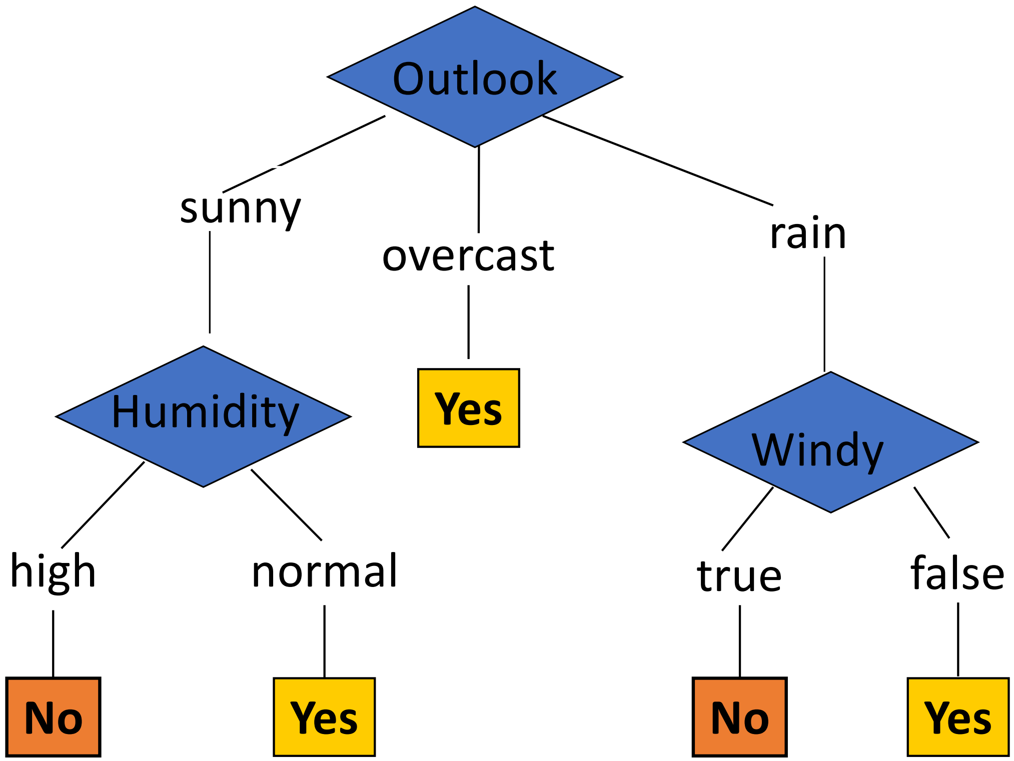
Anatomy of a decision tree



Anatomy of a decision tree



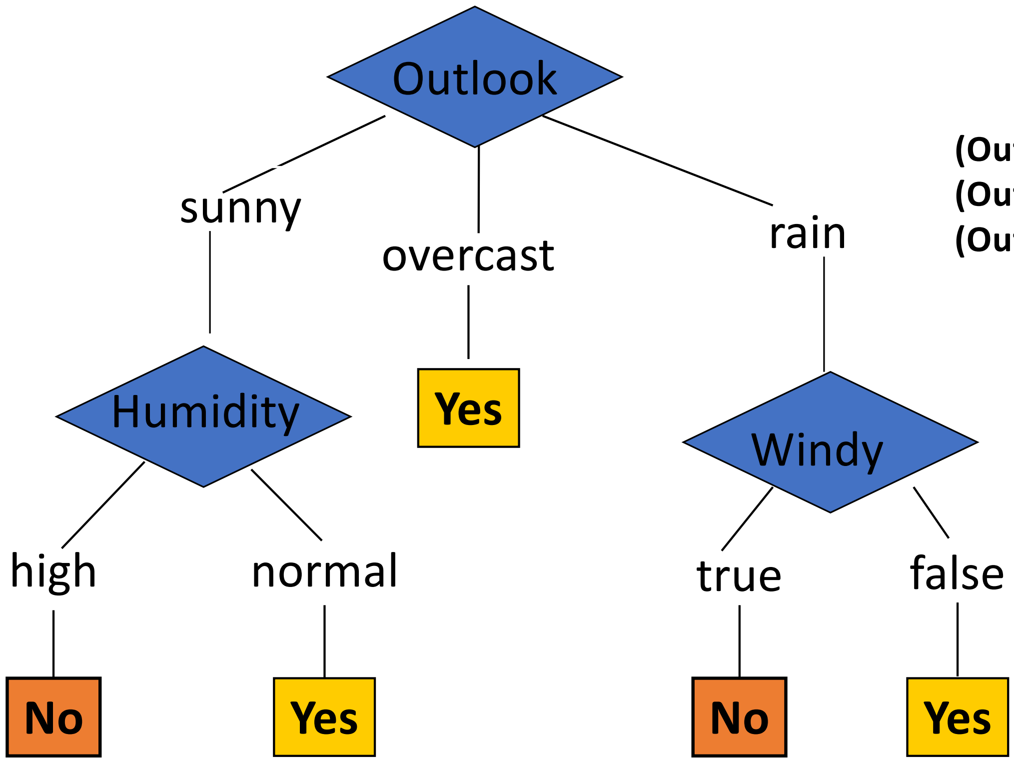
To ‘play tennis’ or not



A new test example:
(Outlook==rain) and
(Windy==false)

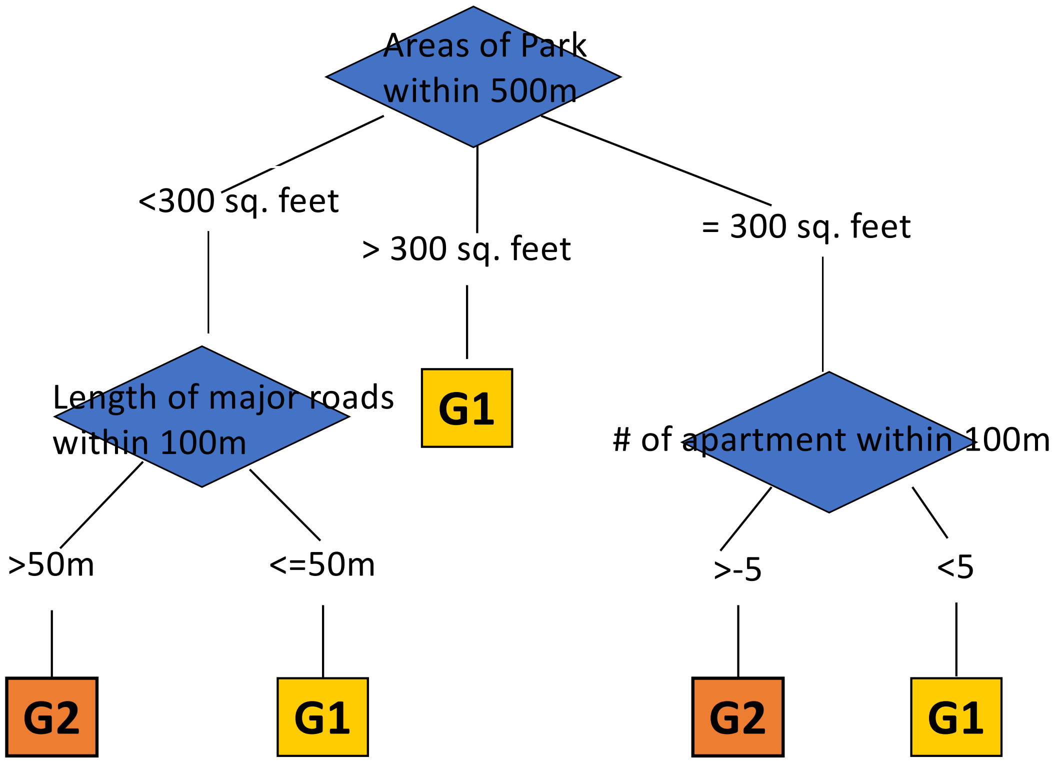
Pass it on the tree
-> Decision is yes.

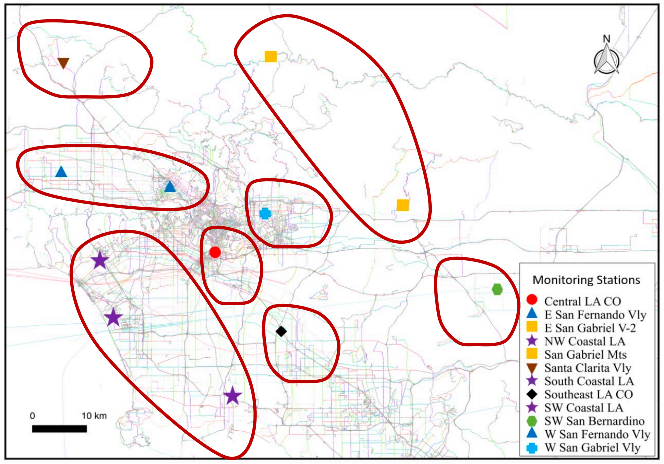
To ‘play tennis’ or not



(Outlook ==overcast) -> yes
(Outlook==rain) and (Windy==false) ->yes
(Outlook==sunny) and (Humidity=normal) ->yes

How environment affect air quality



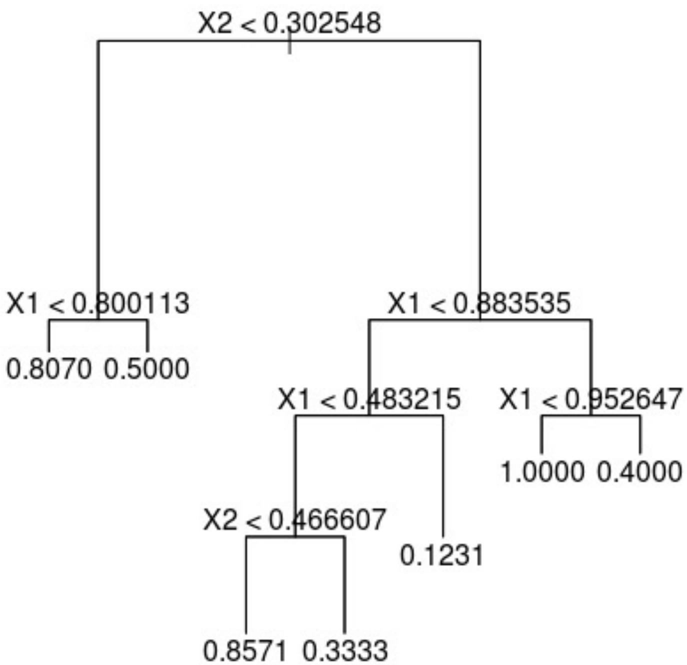


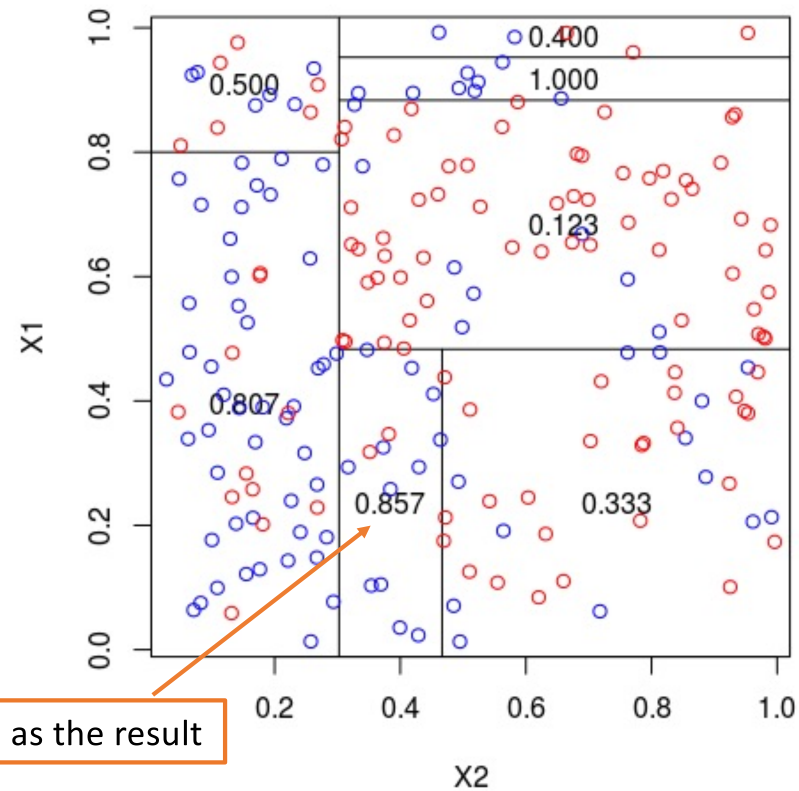
Decision trees

- Decision trees represent a disjunction of conjunctions of constraints on the attribute values of instances.

- (Outlook ==overcast)
- OR
- ((Outlook==rain) and (Windy==false))
- OR
- ((Outlook==sunny) and (Humidity=normal))
- => **yes** play tennis

Decision trees as a regressor

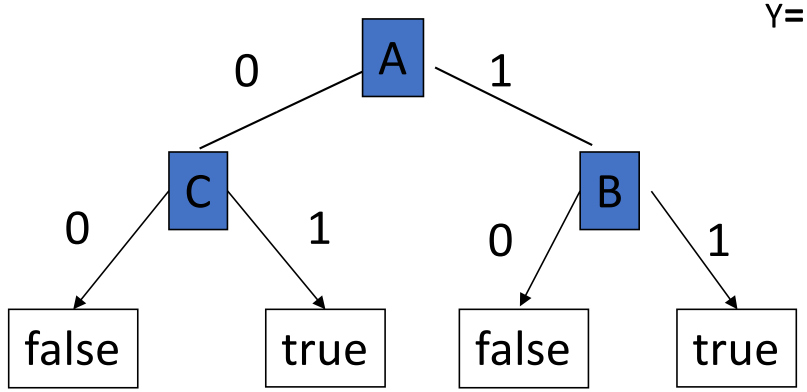




Use mean response in the region as the result

https://gdcoder.com/decision-tree-regressor-explained-in-depth/

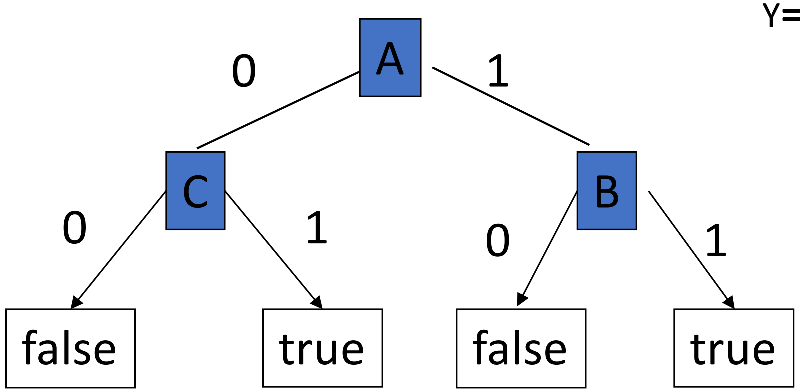
Tree Representation

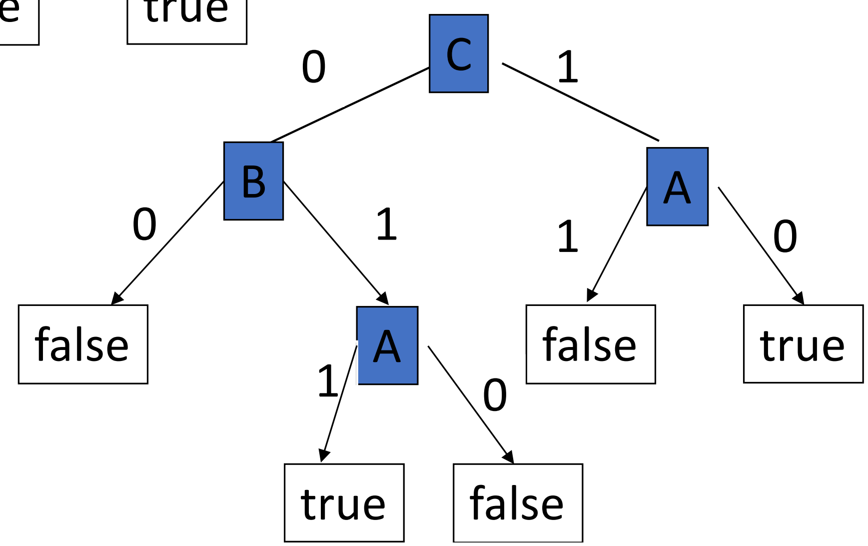


Y=((A and B) or ((not A) and C))

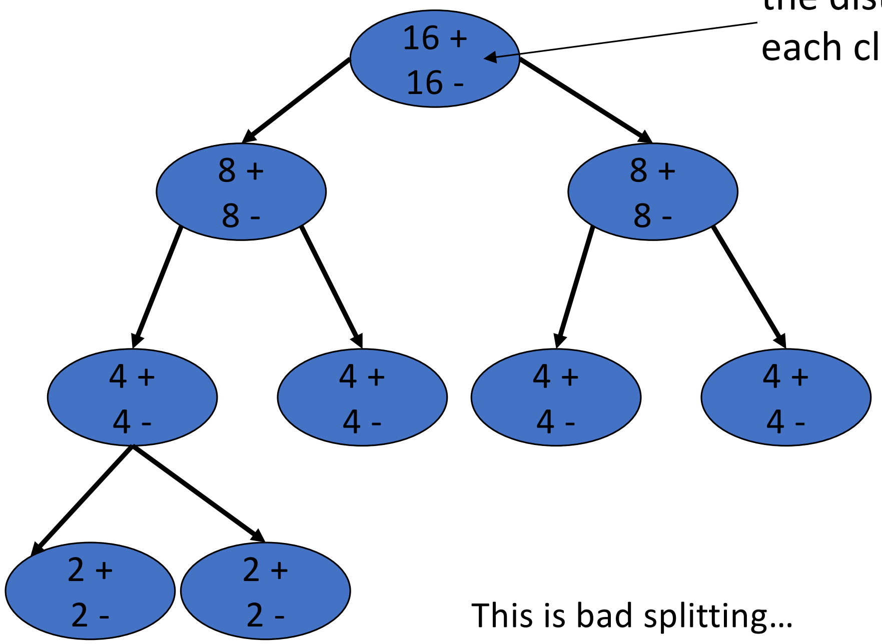
Same concept different representation

Y=((A and B) or ((not A) and C))





Which attribute to select for splitting?



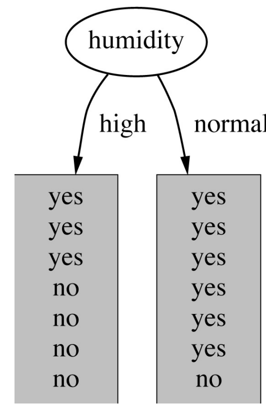
the distribution of each class (not attribute)

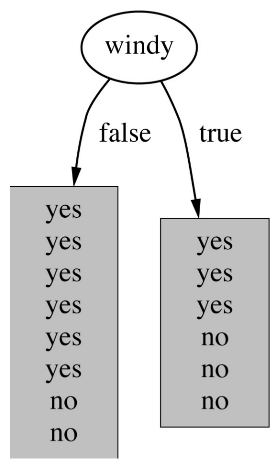
This is bad splitting…

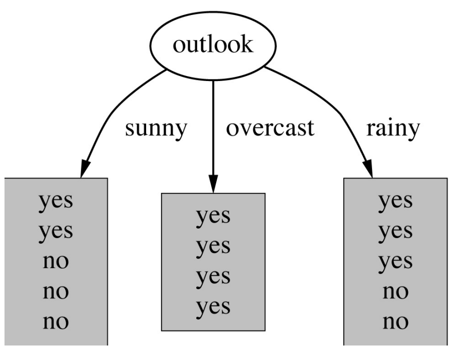
How do we choose the test ?

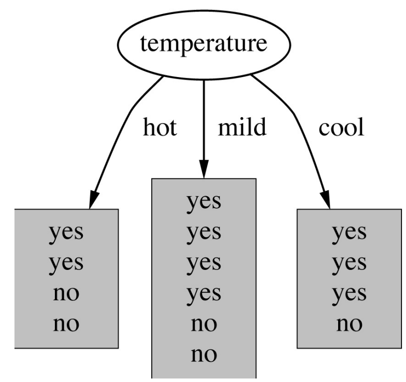
Which attribute should be used as the test?

Intuitively, you would prefer the attribute that *separates* the training examples as much as possible.









Information Gain

- Information gain is one criteria to decide on the split attribute.

Information

Imagine:

- 1. Someone is about to tell you your own name
- 2. You are about to observe the outcome of a dice roll
- 2. You are about to observe the outcome of a coin flip
- 3. You are about to observe the outcome of a biased coin flip

- Each situation have a different *amount of uncertainty* as to what outcome you will observe.

Information Theory

- Information:
- reduction in uncertainty (amount of surprise in the outcome)

$$I(E) = \log_2 \frac{1}{p(x)} = -\log_2 p(x)$$

If the probability of this event happening is small and it happens the information is large.

- Observing the outcome of a coin flip is head $\longrightarrow$ $I = -\log_2 1/2 = 1$
- Observe the outcome of a dice is 6 $\longrightarrow$ $I = -\log_2 1/6 = 2.58$

Watch this: https://www.youtube.com/watch?v=v68zYyaEmEA

Entropy of an information source

- The *expected amount of information* (in bits with log base 2) when observing the output of a random variable X

$$H(X) = E(I(X)) = \sum_i p(x_i) I(x_i) = -\sum_i p(x_i) \log_2 p(x_i)$$

If X can have 8 outcomes and all are equally likely

$$H(X) == -\sum_i 1/8 \log_2 1/8 = 3 \text{ bits}$$

If X can have 6 outcomes and all are equally likely

6 x 1/6 x $I = -\log_2 1/6 = 2.58$

Biased dice roll that shows only 1 or 2 50% of the chance

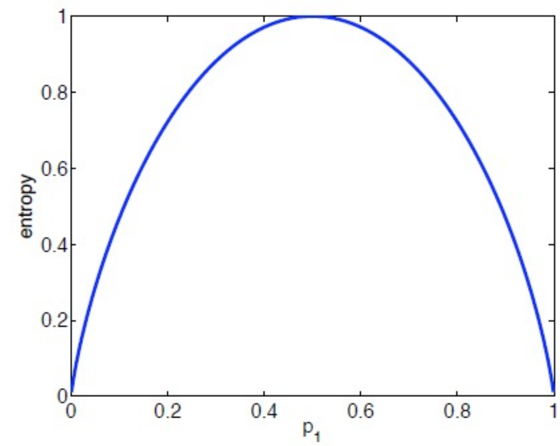
50% x 1 + 50% x1 + 4 x 0

$$I = -\log_2 1/2 = 1$$

Entropy

Equality holds when all outcomes are equally likely

The more the probability distribution deviates from uniformity the lower the entropy



e.g., unbiased coin toss – p = 0.5 has the highest entropy 1

Fair dice roll: 6 x 1/6 x $I = -\log_2 1/6 = 2.58$

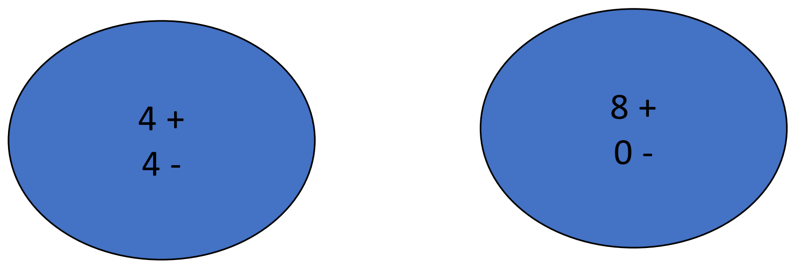
Biased dice roll that shows only 1 or 2 50% of the chance:

50% x 1 + 50% x1 + 4 x 0

$$I = -\log_2 1/2 = 1$$

Entropy, purity

Entropy measures the purity



The distribution is less uniform
Entropy is lower
The node is purer

Information Gain

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

Reduction in uncertainty by knowing Y

Information gain:
(information before split) – (information after split)

Conditional entropy

$$H(X) = -\sum_i p(x_i)\log_2 p(x_i)$$

$$H(X \mid Y) = -\sum_j p(y_j) H(X \mid Y = y_j)$$

$$= -\sum_j p(y_j) \sum_i p(x_i \mid y_j)\log_2 p(x_i \mid y_j)$$

Information Gain

Information gain:

- (information before split) – (information after split)

Example

Attributes Labels

X1	X2	Y	Count
T	T	+	2
T	F	+	2
F	T	-	5
F	F	+	1

Attributes Labels

X1	X2	Y	Count
T	T	+	2
T	F	+	2
F	T	-	5
F	F	+	1

Which one do we choose X1 or X2?

Information gain: (information before split) – (information after split)

IG(X1,Y) = H(Y) – H(Y|X1)

Y: +, -
X1: T, F
X2: T, F

H(Y) = - (5/10) log(5/10) -5/10log(5/10) = 1

H(Y|X1) = P(X1=T)H(Y|X1=T) + P(X1=F) H(Y|X1=F)

= 4/10 (1log 1 + 0 log 0) +6/10 (5/6log 5/6 +1/6log1/6)

= 0.39

Information gain (X1,Y)= 1-0.39=0.61

$$H(X) = -\sum_i p(x_i)\log_2 p(x_i)$$

$$H(X|Y) = -\sum_j p(y_j) H(X|Y=y_j)$$

$$= -\sum_j p(y_j)\sum_i p(x_i|y_j)\log_2 p(x_i|y_j)$$

J: loop through T, F

I: loop through +, -

Which one do we choose?

X1	X2	Y	Count
T	T	+	2
T	F	+	2
F	T	-	5
F	F	+	1

Information gain (X1,Y)= 0.61

Information gain (X2,Y)= 0.12

Pick the variable which provides
the most information gain about Y

Pick X1

Recurse on branches

X1	X2	Y	Count
T	T	+	2
T	F	+	2
F	T	-	5
F	F	+	1

One branch

The other branch

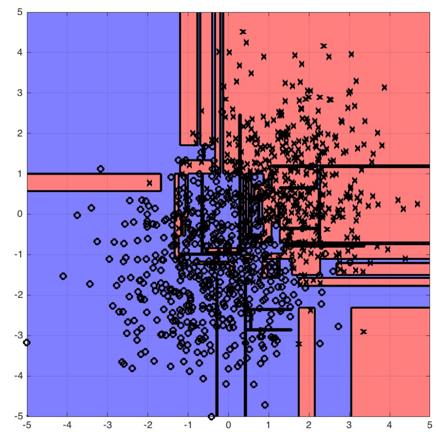
Purity (diversity) measures

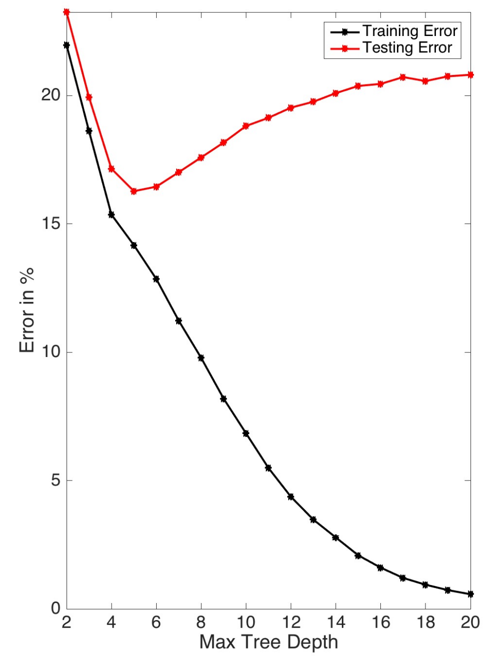
- Gini (population diversity)
- Information Gain
- Chi-square Test

Overfitting

- You can perfectly fit to any training data

- Two approaches:
  - Stop growing the tree when further splitting the data does not yield an improvement
  - Grow a full tree, then prune the tree, by eliminating nodes





https://www.cs.cornell.edu/courses/cs4780/2018fa/lectures/lecturenote17.html

Bagging

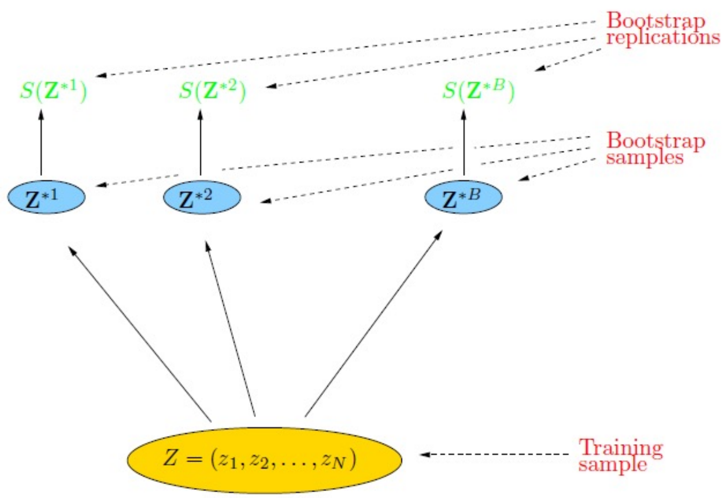
- Bagging or *bootstrap aggregation* a technique for reducing the variance of an estimated prediction function.

- For classification, a *committee* of trees each cast a vote for the predicted class.

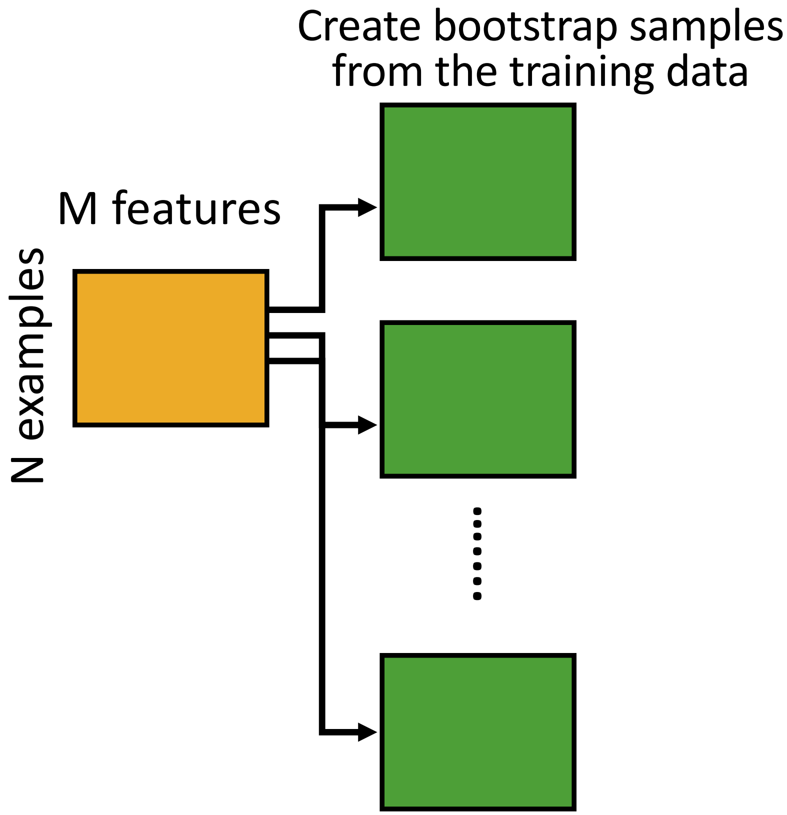
Bootstrap

The basic idea:

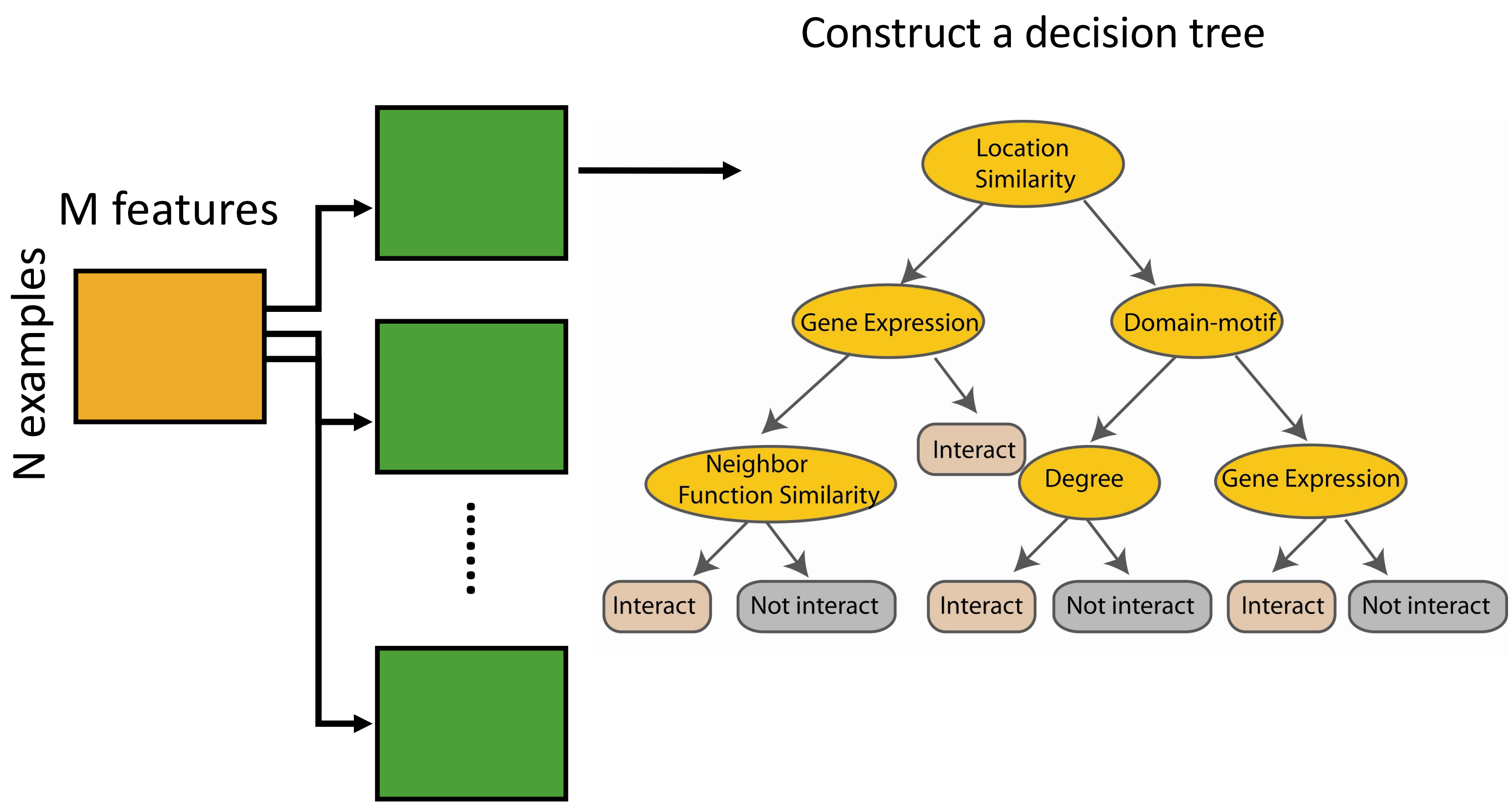
randomly draw datasets *with replacement* from the training data, each sample *the same size as the original training set*



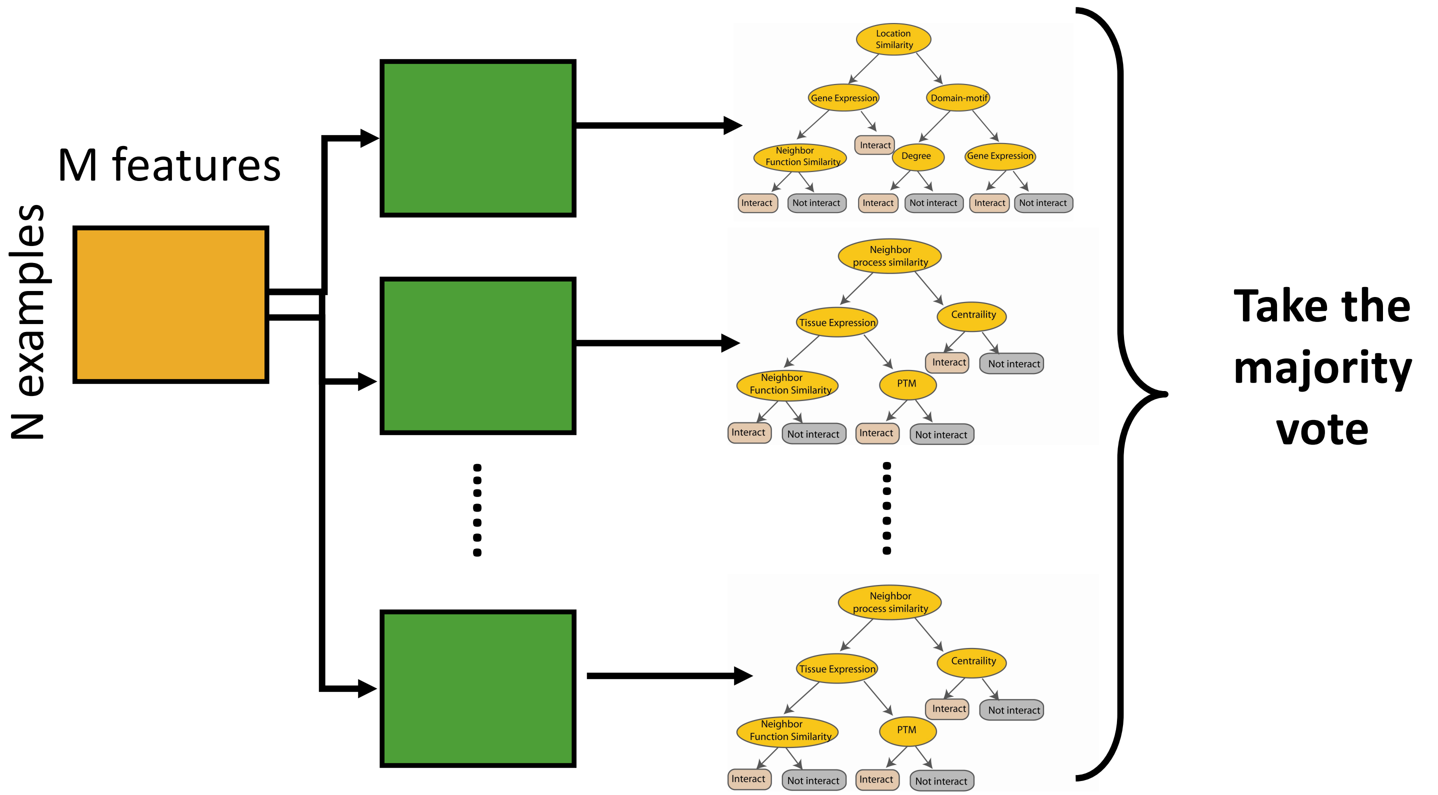
Bagging


Create bootstrap samples
from the training data
M features
N examples

Random Forest Classifier



Random Forest Classifier



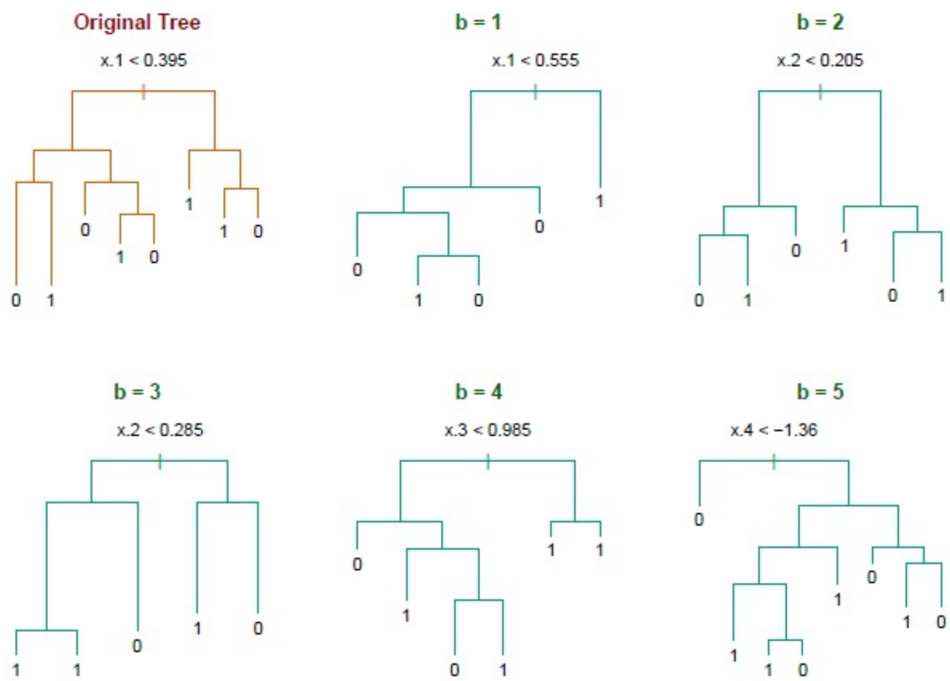
Bagging : a simulated example

- Generated a sample of size $N = 30$,
  - two classes and $p = 5$ features, each having a standard Gaussian distribution with pairwise correlation 0.95.

- The response Y was generated according to
  - $\Pr(Y = 1 / x1 \leq 0.5) = 0.2$,
  - $\Pr(Y = 0 / x1 > 0.5) = 0.8$.

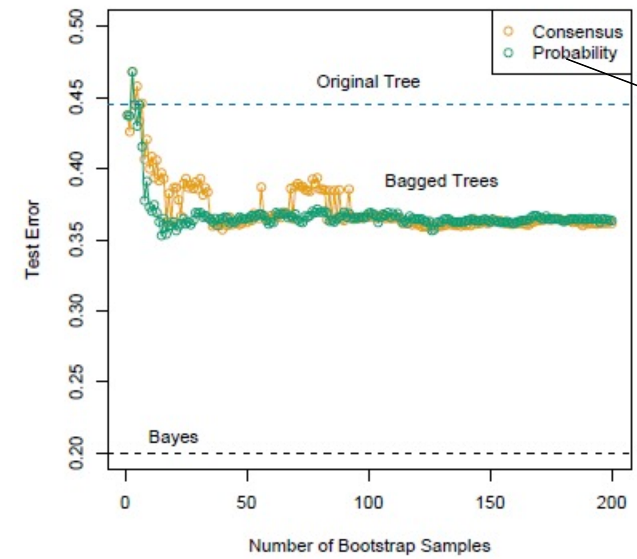
Bagging

Notice the bootstrap trees are different than the original tree



$$\Pr(Y = 1 / x1 \leq 0.5) = 0.2,$$
$$\Pr(Y = 0 / x1 > 0.5) = 0.8.$$

Bagging



Treat the voting
Proportions as
probabilities

FIGURE 8.10. *Error curves for the bagging example of Figure 8.9. Shown is the test error of the original tree and bagged trees as a function of the number of bootstrap samples. The orange points correspond to the consensus vote, while the green points average the probabilities.*

bagging helps under squared-error loss, in short because averaging reduces

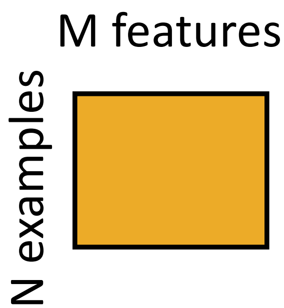
http://www-stat.stanford.edu/~hastie/Papers/ESLII.pdf Example 8.7.1

Random forest classifier

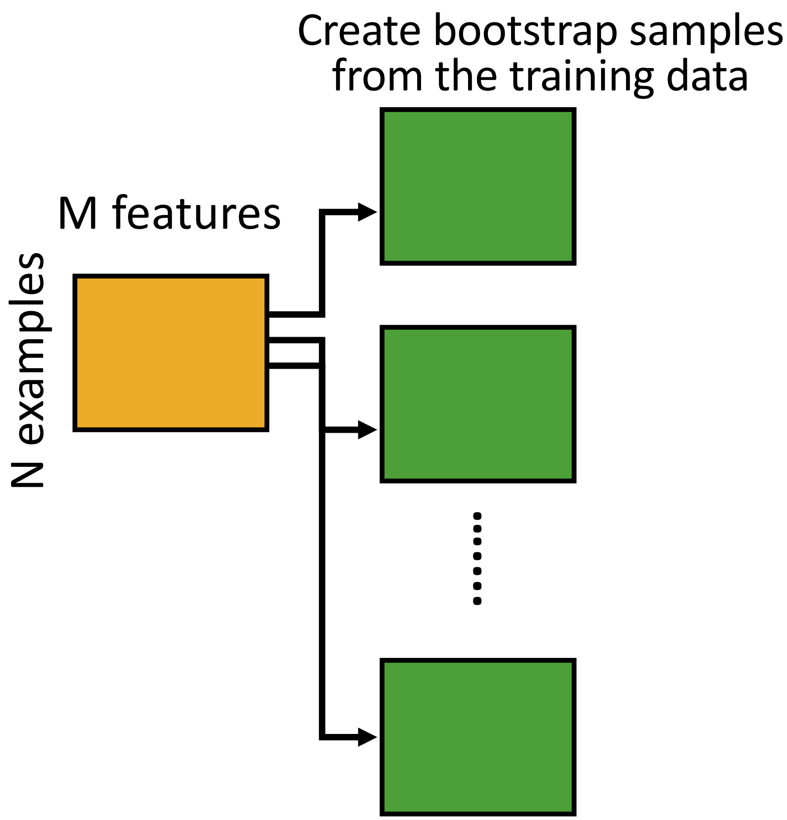
- Random forest classifier, an extension to bagging which uses *de-correlated* trees.

Random Forest Classifier

Training Data

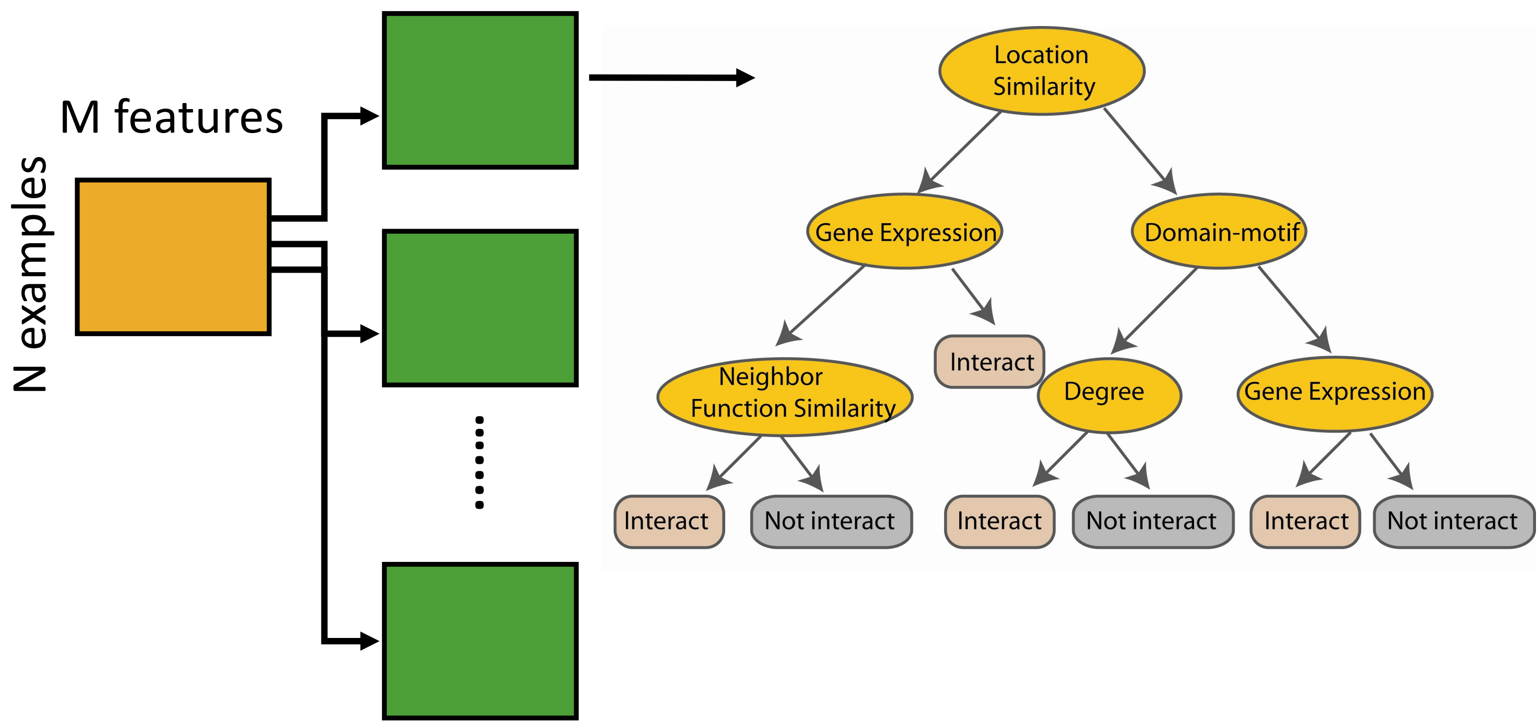


Random Forest Classifier



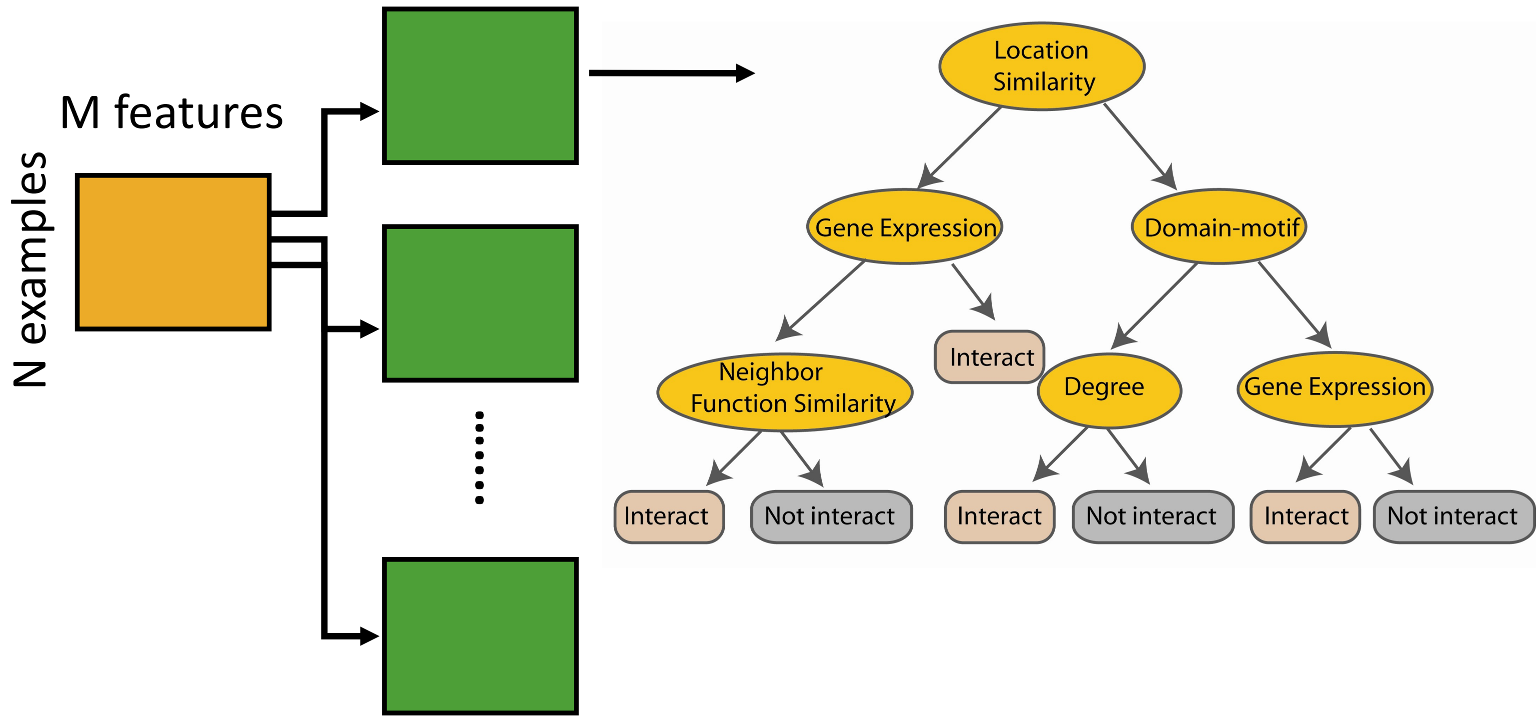
Random Forest Classifier

Construct a decision tree

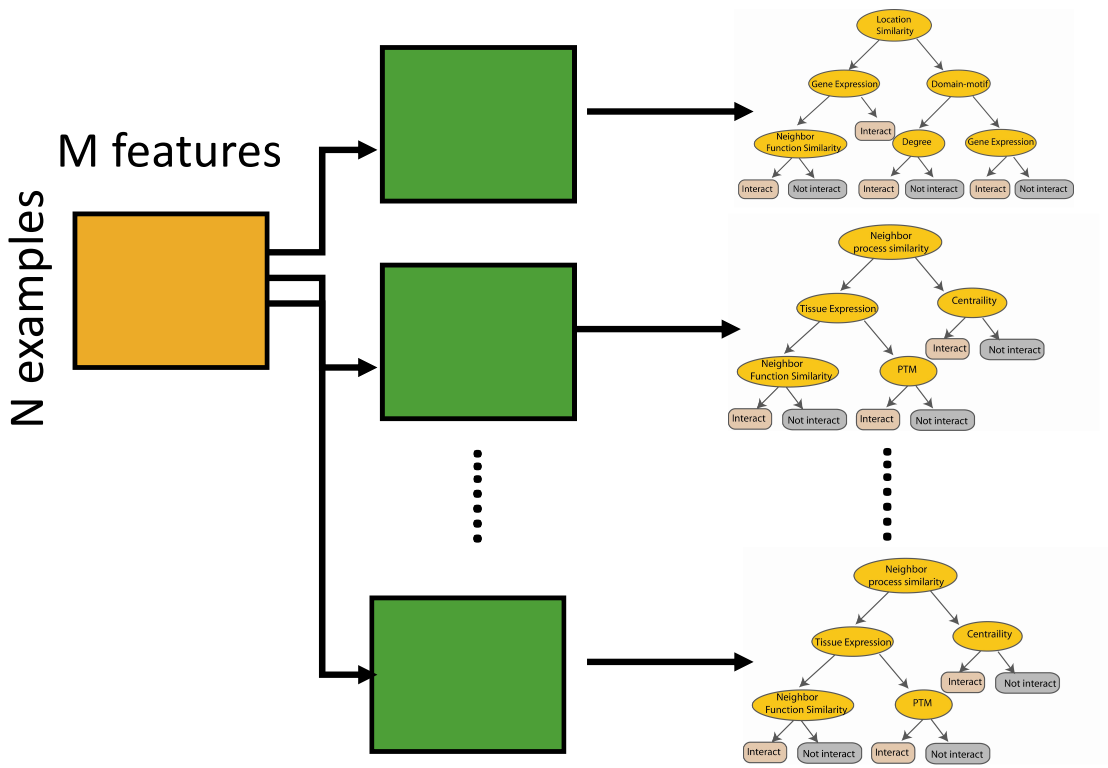


Random Forest Classifier

At each node in choosing the split feature
choose only among $m<M$ features

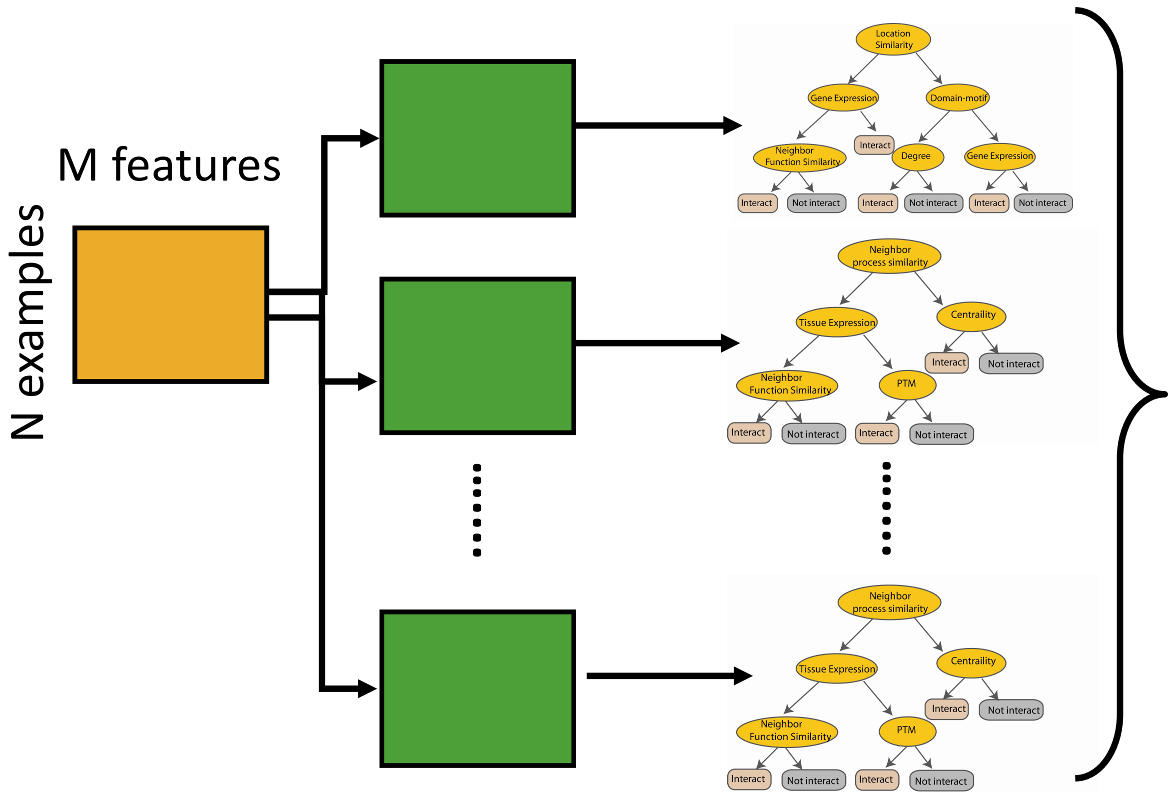


Random Forest Classifier



Create decision tree from each bootstrap sample

Random Forest Classifier



Take he majority vote

Importance Score (Categorical RF)

Random Forest: income

Correct predictions (based on out-of-bag sample): 82% (<=50K: 90.13%; >50K: 57.03%)

	<=50K	>50K	MeanDecreaseAccuracy	Importance (MeanDecreaseGini)
occupation	0.022	0.063	0.032	115.53
age	-0.001	0.065	0.016	108.30
education_num	0.023	0.077	0.036	96.92
relationship	0.022	0.079	0.036	80.74
hrs_per_week	0.002	0.039	0.011	67.80
marital	0.024	0.066	0.034	61.21
workclass	0.007	-0.006	0.004	41.75
country	0.000	-0.006		

n = 2000 cases used in estimation;

How much the accuracy decreases when the variable is excluded

The decrease of Gini impurity when a variable is chosen to split a node

https://www.displayr.com/how-is-variable-importance-calculated-for-a-random-forest/

Importance Scores (Categorical RF)

- Gini Importance (mean decrease impurity)
  - On average, how the selected feature at a node decreases the impurity of the split
  - Measured for every three
  - Derived from the RF structure
  - Often prefer numerical features (or categorical features with high cardinality)
  - Ignore important but not the most important features at a node
- Mean Decrease Accuracy
  - Set a feature with random values (so that it has no predictive power)
  - Calculate how the accuracy number decreases

Random forest Resouces

- Available package:
- https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html
- https://spark.apache.org/docs/latest/api/python/reference/api/pyspark.mllib.tree.RandomForest.html
- To read more:
- http://www-stat.stanford.edu/~hastie/Papers/ESLII.pdf

Acknowledgements

- Gil, Yolanda (Ed.) Introduction to Computational Thinking and Data Science. Available from http://www.datascience4all.org
- Oznur Tastan, Geoffrey J. Gordon, Machine Learning 10601, Recitation 8, Oct 21, 2009 (http://people.sabanciuniv.edu/otastan/, https://www.cs.cmu.edu/~ggordon/10601/recitations/rec08/Rec08_Oct21.ppt)