CS 446/ECE 449: Machine Learning

Shenlong Wang

University of Illinois at Urbana-Champaign, 2024

L07, Decision Trees

Goals of this lecture

- Getting to know Classification and Regression Trees
- Getting to know Random Forests
- Getting to know Ensembles
- Getting to know Cross-Validation

Reading material:

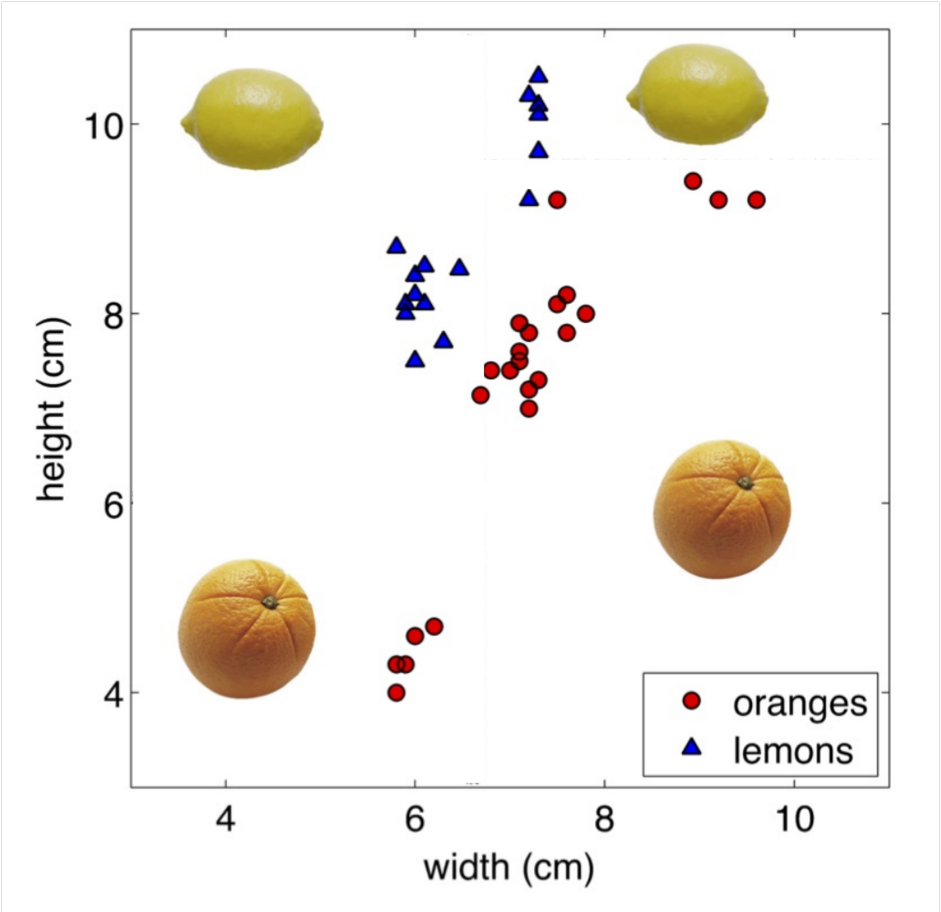
- Shai Shalev-Shwartz & Shai Ben-David, Understanding Machine Learning: From Theory to Algorithms, Chapter 4
- Kevin Murphy, Probabilistic Machine Learning An Introduction; Chapter 18.1, Chapter 18.2

Recap:

So far we have covered:

- Linear regression
- Logistic regression
- Linear SVM
- Kernel methods

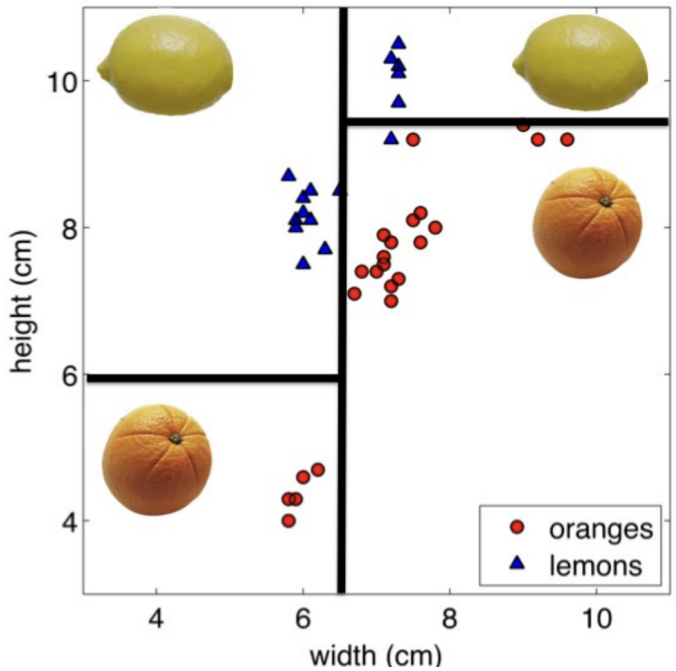
Decision Tree

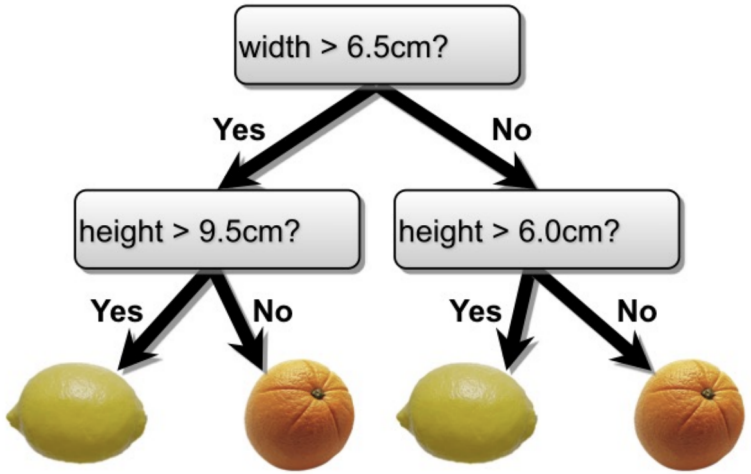


An alternative classifier idea:

- Pick an attribute, do a simple test
- Conditioned on a choice, pick another attribute, do another test
- In the leaves, assign a class with majority vote
- Do other branches as well

Decision Tree

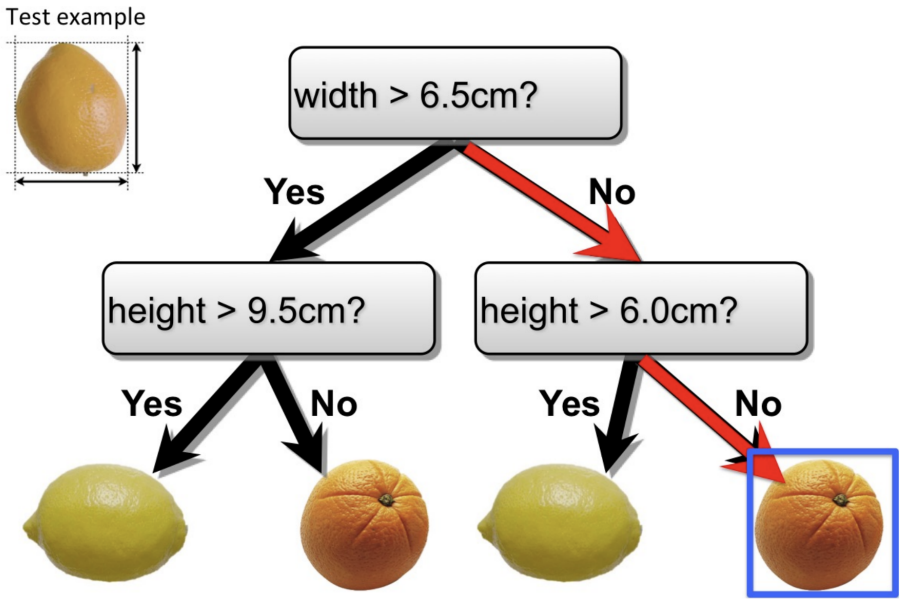




An alternative classifier idea:

- Pick an attribute, do a simple test
- Conditioned on a choice, pick another attribute, do another test
- In the leaves, assign a class with majority vote
- Do other branches as well

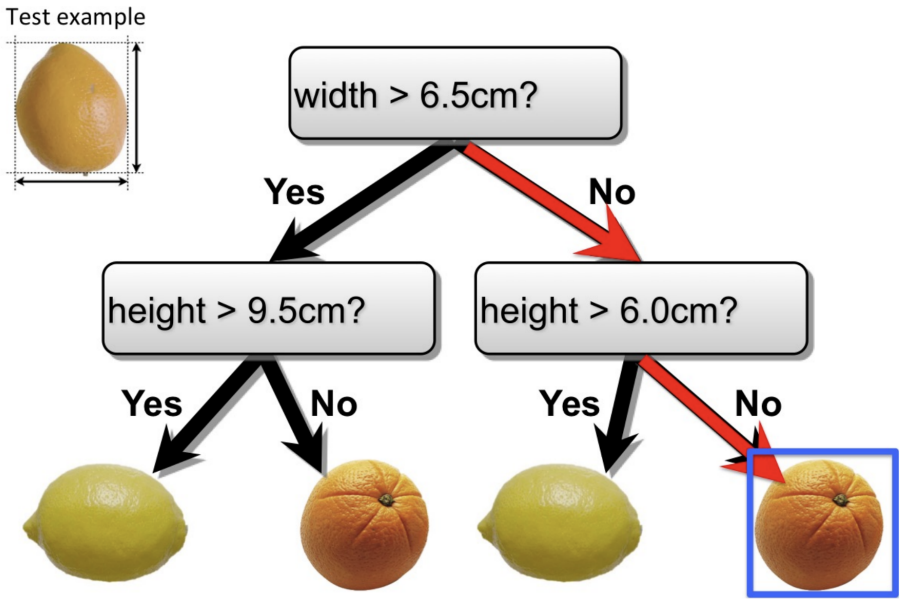
Decision Tree Testing



Inference:

- Start at the root and follow the decisions
- The leaf node reveals the result

Decision Tree Formulation



Formally: A decision tree is a function $f : \mathcal{X} \to \mathcal{Y}$, represented by a binary tree in which:

- Each tree node is associated with a splitting rule $g : X \to \{0, 1\}$.
- Each leaf node is associated with a label $y \in \mathcal{Y}$ (can be either classification or regression)
- Typically we consider axis-aligned splits $g(x) = \text{sign}(x_i - t)$

Question: *How to learn the decisions of such a tree?*

Decision Tree Learning

Learning the simplest (smallest) decision tree is an NP complete problem (if you are interested, check: Hyafil and Rivest 76)
Instead, we resort to a greedy heuristic:

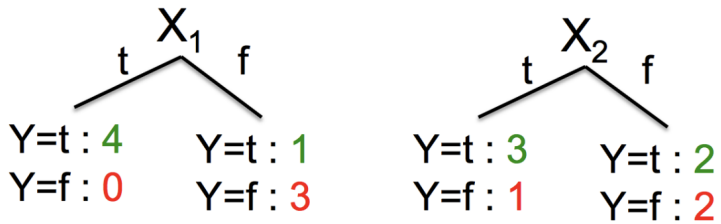
- Choose a variable that "best" splits the set of data items
- Split the data according to the chosen rule, append two nodes and let them process their data
- Stop once the number of datapoints in a node is reasonably small and compute the leaf node statistics

The leaf node statistics are the classification result.

What is best splitting rule? We use information theory to guide us

Decision Tree Learning

Which attribute is better to split on, $X1$ or $X2$?



X_1	X_2	Y
T	T	T
T	F	T
T	T	T
T	F	T
F	T	T
F	F	F
F	T	F
F	F	F

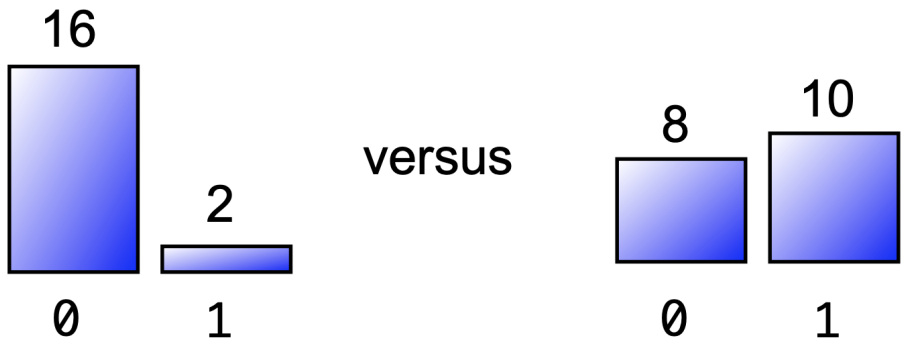
- Determinist: good (all are true or false; just one class in the leaf);
- Uniform: bad all classes in leaf equally probable

Idea: Use counts at leaves to define probability distributions, so we can choose the split that most decreases the uncertainty in prediction.
How to measure uncertainty?

Which one is more uncertain?

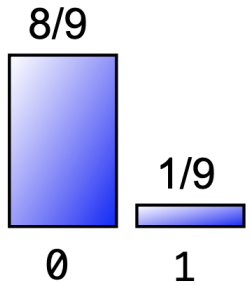
Sequence 1:
0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1 0 0 ... ?

Sequence 2:
0 1 0 1 0 1 1 1 0 1 0 0 1 1 0 1 0 1 ... ?

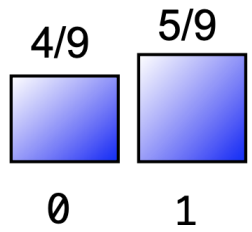


Entropy

$$I(\mathcal{D}) = -\sum_{c=1}^{C} p(c|\mathcal{D}) \log p(c|\mathcal{D})$$



$$-\frac{8}{9}\log_2\frac{8}{9} - \frac{1}{9}\log_2\frac{1}{9} \approx \frac{1}{2}$$



$$-\frac{4}{9}\log_2\frac{4}{9} - \frac{5}{9}\log_2\frac{5}{9} \approx 0.99$$

The logarithm base here can be 2 (bits), e (nats) or 10 (dits) (we use bits for the rest of this lecture)

- How surprised we are by a new value
- How much information this sequence conveys

Information Gain Metric for measuring "best" split: Information Gain (N child nodes, $f \in \mathcal{F}$ split function, $\mathcal{D}$ data at parent node, $\mathcal{D}_j$ data at j-th child)

$$IG(\mathcal{D}, f) = I(\mathcal{D}) - I(\mathcal{D}|f) = I(\mathcal{D}) - \sum_{j=1}^{N} \frac{|\mathcal{D}_j|}{|\mathcal{D}|} I(\mathcal{D}_j)$$

how much uncertainty reduction we achieve through the split f

Information Gain

$$IG(\mathcal{D}, f) = I(\mathcal{D}) - \sum_{j=1}^{N} \frac{|\mathcal{D}_j|}{|\mathcal{D}|} I(\mathcal{D}_j)$$

There are multiple choices for measurement of uncertainty:

- Entropy

$$I(\mathcal{D}) = -\sum_{c=1}^{C} p(c|\mathcal{D}) \log p(c|\mathcal{D})$$

- Gini impurity

$$I(\mathcal{D}) = 1 - \sum_{c=1}^{C} p(c|\mathcal{D})^2$$

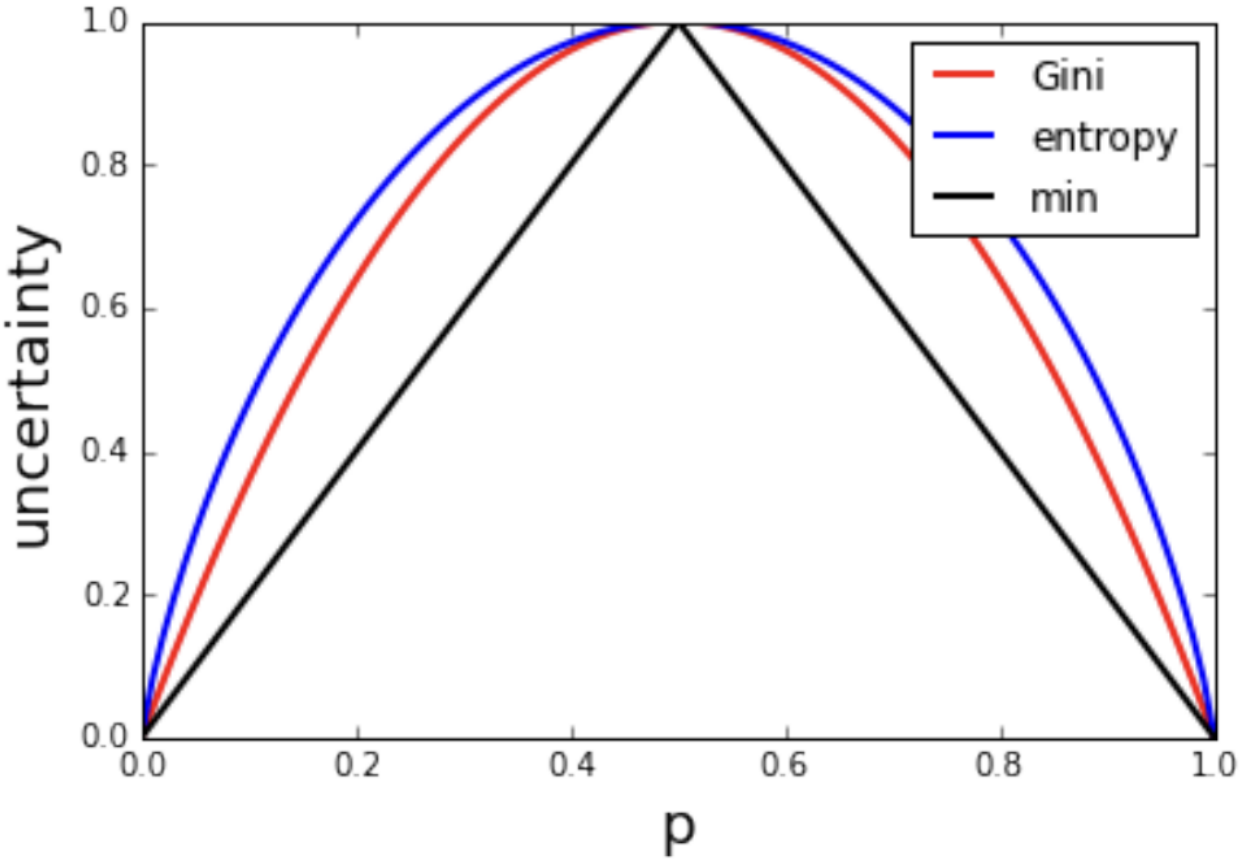
- Classification error

$$I(\mathcal{D}) = 1 - \max_{c \in \{1,\ldots,C\}} p(c|\mathcal{D})$$

Information Gain

$$IG(\mathcal{D}, f) = I(\mathcal{D}) - I(\mathcal{D}|f) = I(\mathcal{D}) - \sum_{j=1}^{N} \frac{|\mathcal{D}_j|}{|\mathcal{D}|} I(\mathcal{D}_j)$$

There are multiple choices for measurement of uncertainty:



Information Gain

$$IG(\mathcal{D}, f) = I(\mathcal{D}) - \sum_{j=1}^{N} \frac{|\mathcal{D}_j|}{|\mathcal{D}|} I(\mathcal{D}_j)$$

Example A:

- Gini impurity:

$$I(\mathcal{D}) = 1 - \sum_{c=1}^{C} p(c|\mathcal{D})^2$$

- $\mathcal{D}$: 10 examples of class 0 and 10 examples of class 1
- $\mathcal{D}_1$: 10 examples of class 0 and 0 examples of class 1
- $\mathcal{D}_2$: 0 examples of class 0 and 10 examples of class 1

$$1 - 0.5^2 - 0.5^2 - \frac{1}{2}(1 - 0^2 - 1^2) - \frac{1}{2}(1 - 1^2 - 0^2)$$

Information Gain

$$IG(\mathcal{D}, f) = I(\mathcal{D}) - \sum_{j=1}^{N} \frac{|\mathcal{D}_j|}{|\mathcal{D}|} I(\mathcal{D}_j)$$

Example B:

- Gini impurity:

$$I(\mathcal{D}) = 1 - \sum_{c=1}^{C} p(c|\mathcal{D})^2$$

- $\mathcal{D}$: 10 examples of class 0 and 10 examples of class 1
- $\mathcal{D}_1$: 5 examples of class 0 and 5 examples of class 1
- $\mathcal{D}_2$: 5 examples of class 0 and 5 examples of class 1

$$1 - 0.5^2 - 0.5^2 - \frac{1}{2}(1 - 0.5^2 - 0.5^2) - \frac{1}{2}(1 - 0.5^2 - 0.5^2)$$

Information Gain

$$IG(\mathcal{D}, f) = I(\mathcal{D}) - \sum_{j=1}^{N} \frac{|\mathcal{D}_j|}{|\mathcal{D}|} I(\mathcal{D}_j)$$

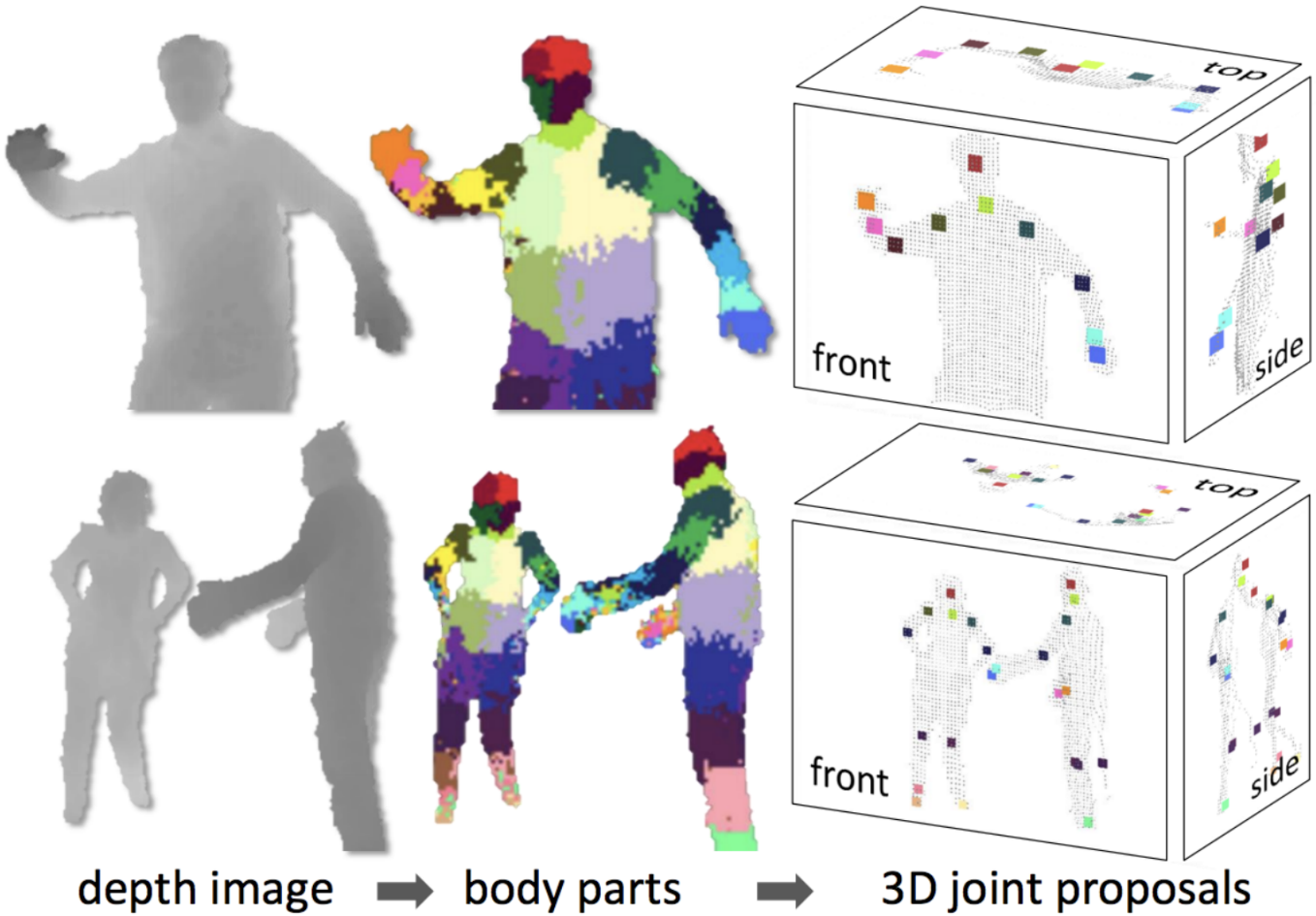
Several things to know:

- $I(\mathcal{D})$ is always positive. $I(\mathcal{D}) = 0$: no uncertainty
- $IG(\mathcal{D}, f) = 0$: f and $\mathcal{D}$ are independent, f doesn't tell anything about $\mathcal{D}$
- $IG(\mathcal{D}, f) = I(\mathcal{D})$, f tells everything about $\mathcal{D}$, perfect splits.

Demo

Application:

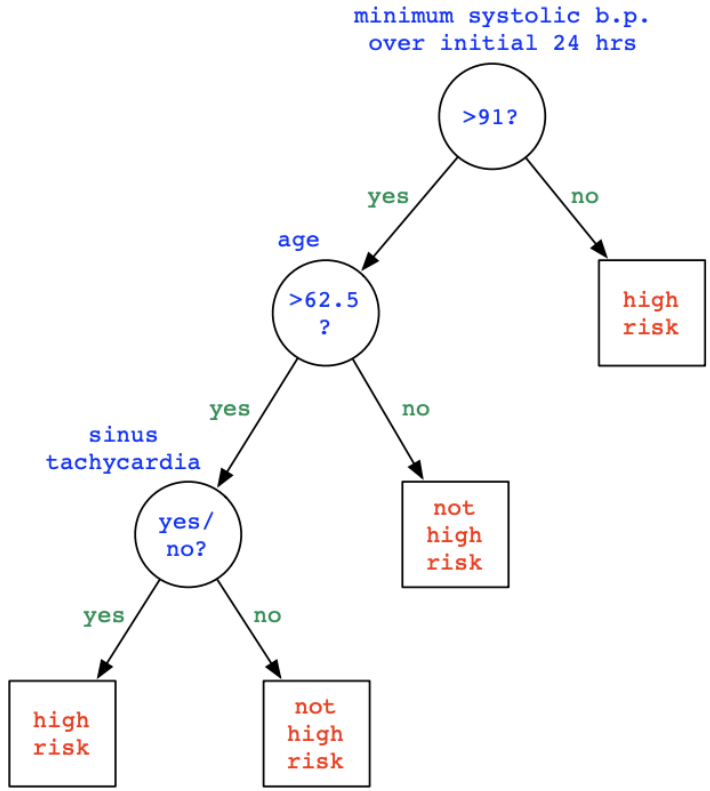
Application:



Application:

UCSD Medical Center (1970s):
identify patients at risk of dying within 30 days after heart attack.

Data set:
215 patients.
37 (=20%) died.
19 features.



Remarks (Decision trees).

- Decision trees are very flexible classifiers (like NN).
- Certain greedy strategies for training decision trees are work well in theory with large data.
- Decision trees are also very prone to overfitting without additional regularization e.g. pruning.
- NP-hard to find smallest decision tree consistent with data (ie. combinatorial search)
- Sequential tree growing is a practical heuristic.

Classification Ensembles

- Train a variety of classification algorithms (same or different type)
- Average their result

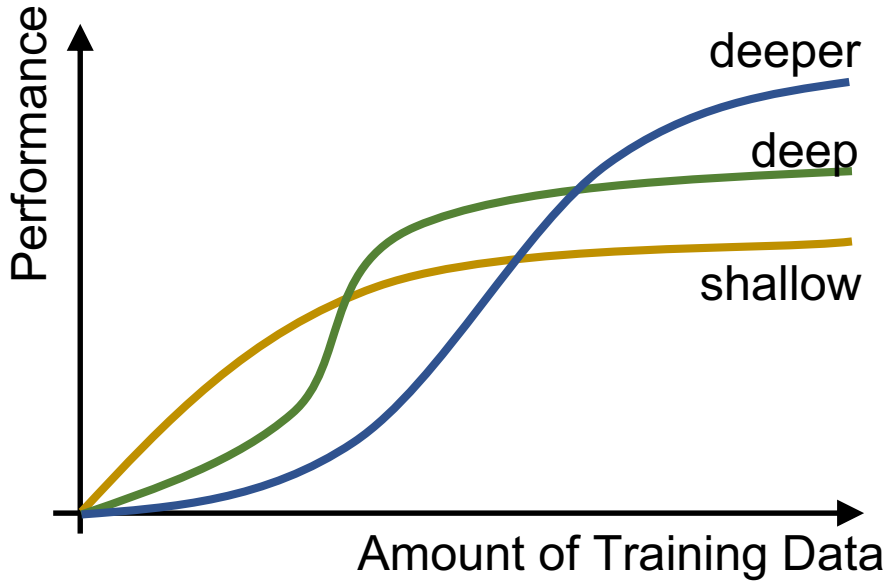
Example:

- Random Forest (parallel ensemble)
- Boosting (sequential ensemble)

Bagging (bootstrap aggregation):

$$F(x) = \frac{1}{M} \sum_{m=1}^{M} F_T^{(m)}(x)$$

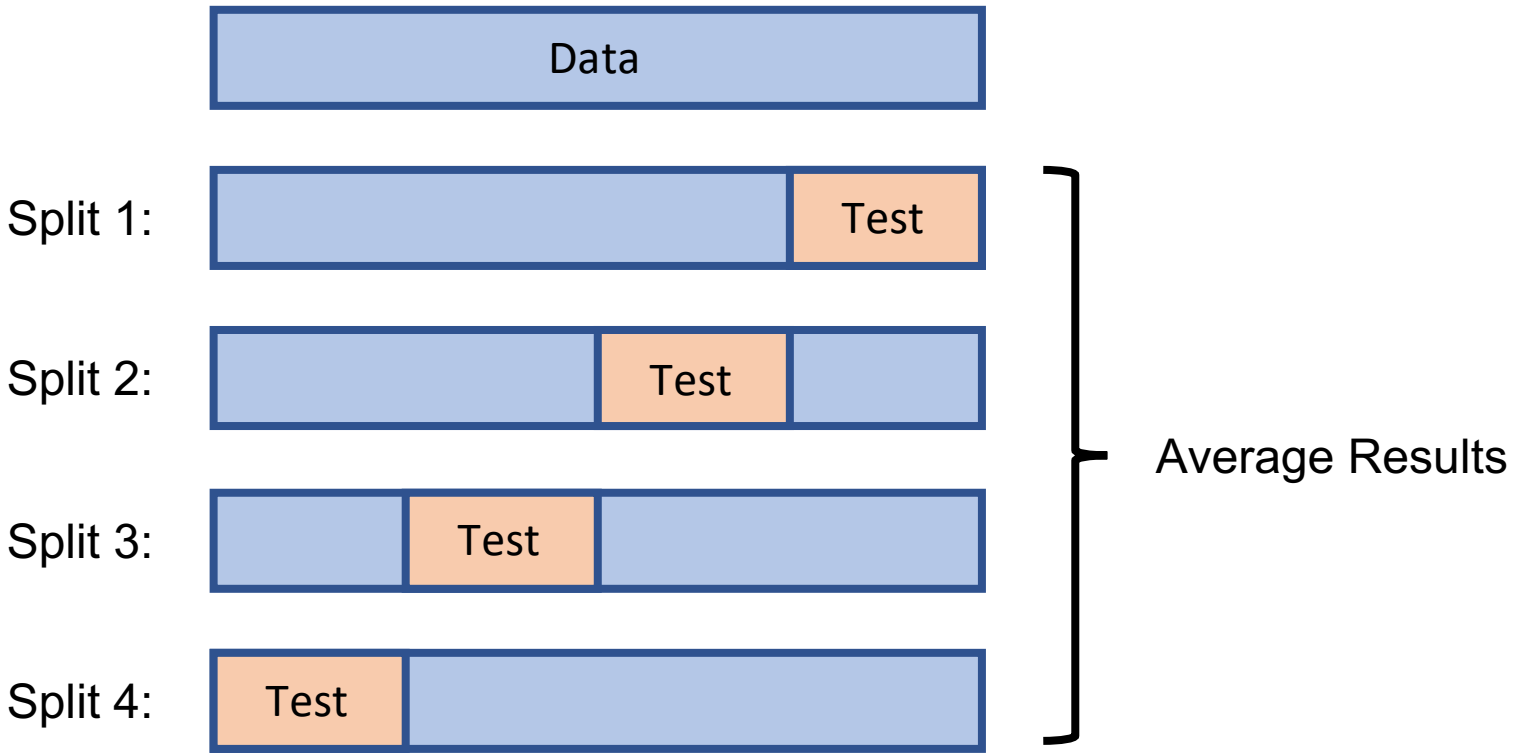
When to choose which method?



How to select hyper-parameters of machine learning models?

- Split your data into train/val/test set
- Choose parameters based on val set
- Report results on test set

4-fold Cross validation:



Important topics of this lecture

- Decision tree
- Ensembles

Up next:

- Boosting