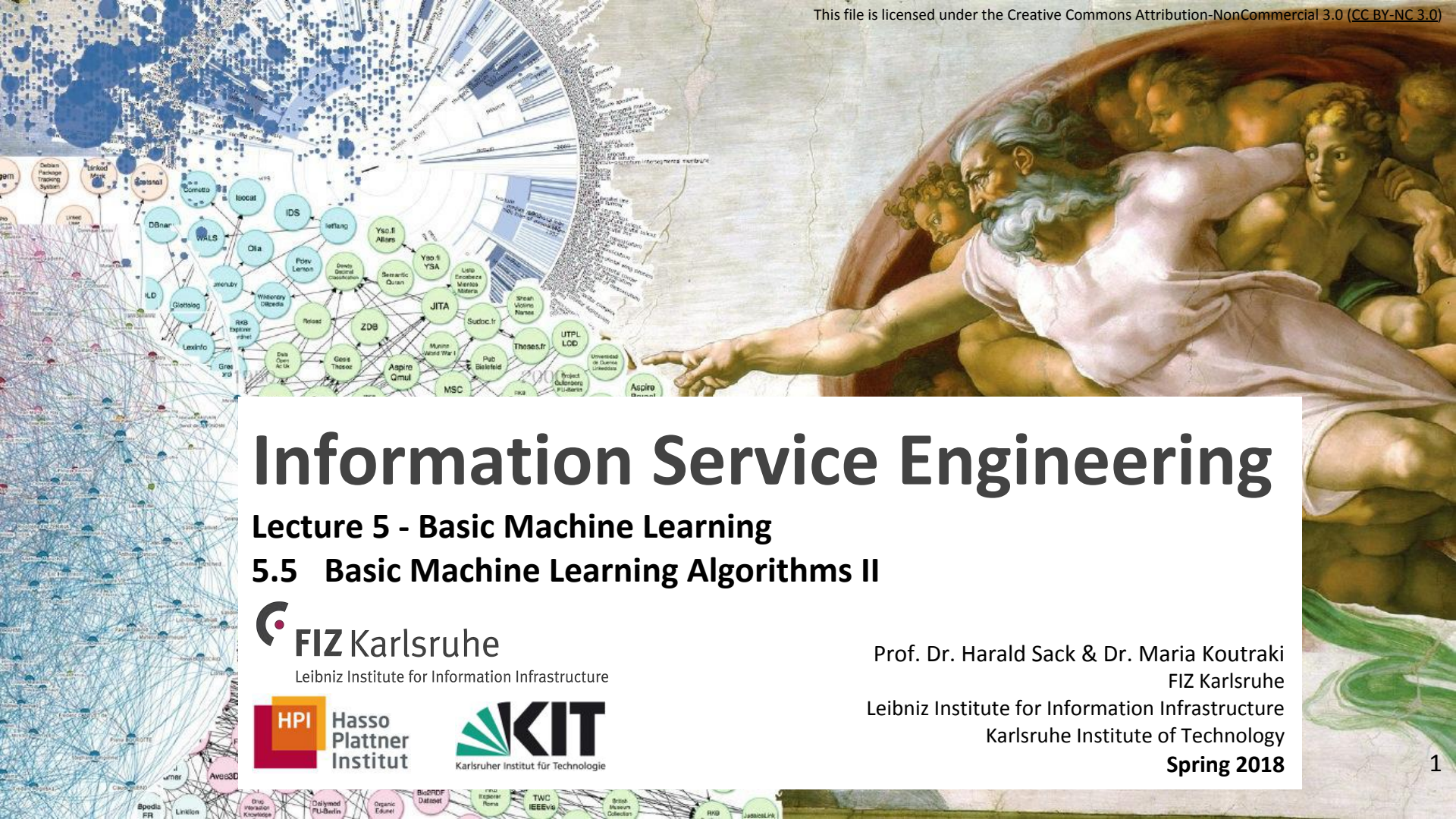




Information Service Engineering

Lecture 5 - Basic Machine Learning

5.5 Basic Machine Learning Algorithms II

 **FIZ Karlsruhe**

Leibniz Institute for Information Infrastructure

 **Hasso Plattner Institut**

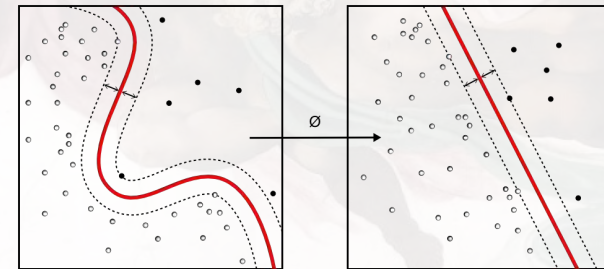
 **KIT**
Karlsruher Institut für Technologie

Prof. Dr. Harald Sack & Dr. Maria Koutraki
FIZ Karlsruhe
Leibniz Institute for Information Infrastructure
Karlsruhe Institute of Technology
Spring 2018

Information Service Engineering

Lecture 5: Basic Machine Learning

- 5.1 A Brief History of AI
- 5.2 Introduction to Machine Learning
- 5.3 Evaluation and Generalization Problems
- 5.4 Basic ML Algorithms
- 5.5 Basic ML Algorithms II**
- 5.6 Unsupervised Learning
- 5.7 Deep Learning



Decision Trees

Information Service Engineering - Lecture 5 Basic Machine Learning - 5.5 Basic Machine Learning Algorithms II

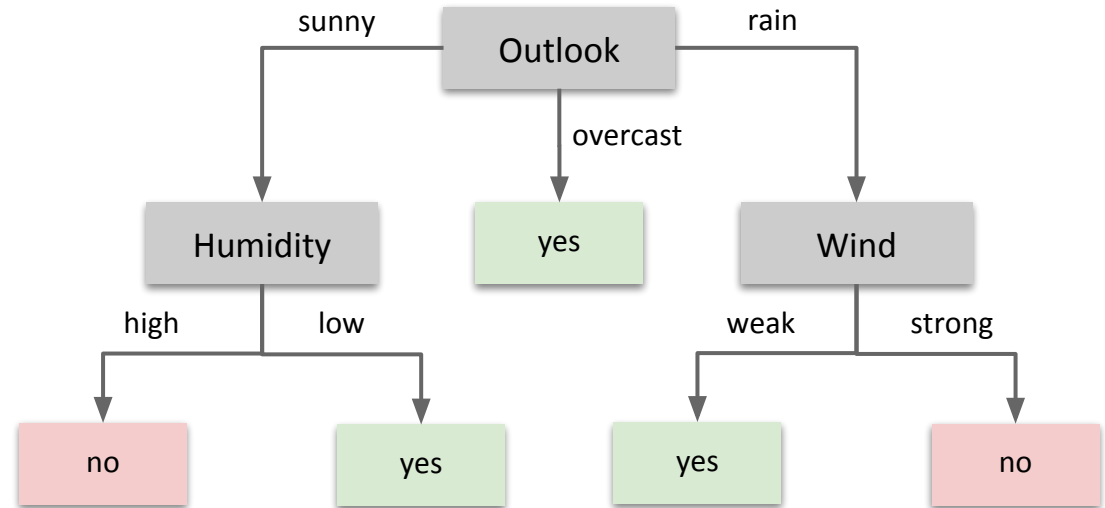
- In many classification problems as well as in Linear Regression, we expect **all attribute values present at the same time**
- Sometimes **new attributes occur only over time**
- Sometimes it is more efficient, **first to check the most promising attributes only**



Should we Play Tennis?

Information Service Engineering - Lecture 5 Basic Machine Learning - 5.5 Basic Machine Learning Algorithms II

- Depends on
 - Outlook
 - Wind
 - Temperature
 - Humidity
 - ...



- Internal node: test attribute X_i
- Branch: selects value for X_i
- Leaf node: predict Y
- Given: a set of possible instances **<Humidity=low, Wind=weak, Outlook=rain, Temp=high>**

Decision Tree Learning

Information Service Engineering - Lecture 5 Basic Machine Learning - 5.5 Basic Machine Learning Algorithms II

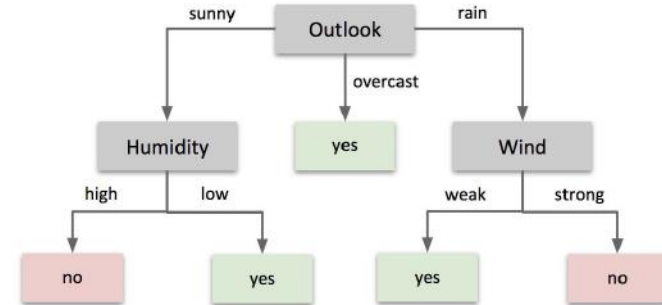
- **Problem Setting:**
 - Set of possible instances X
 - each instance $x \in X$ is a feature vector $x = \langle x_1, x_2, \dots, x_n \rangle$
- Unknown **target function** $f = X \rightarrow Y$
 - Y is discrete valued
- Set of **function hypotheses** $H = \{h \mid h : X \rightarrow Y\}$
 - each hypothesis h is a decision tree
- **Input:**
 - Training examples $\{\langle x^{(i)}, y^{(i)} \rangle\}$ of unknown target function f
- **Output:**
 - Hypothesis $h \in H$ that best approximates target function f

Decision Tree Learning

Information Service Engineering - Lecture 5 Basic Machine Learning - 5.5 Basic Machine Learning Algorithms II

- In general, decision trees represent a **disjunction of conjunctions** of constraints on the attribute values of instances
- The given decision tree corresponds to:

$$\begin{aligned} & (outlook = sunny \wedge humidity = low) \vee \\ & (outlook = overcast) \vee \\ & (outlook = rain \wedge wind = weak) \end{aligned}$$

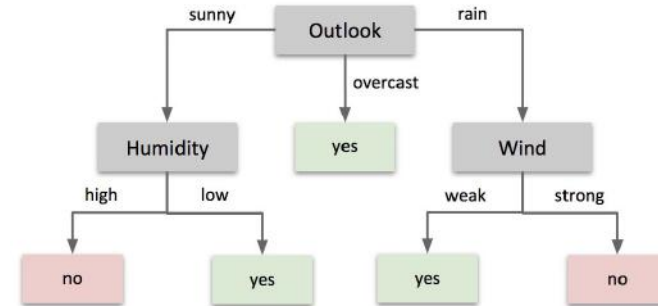


- Decision Trees can represent any function

Decision Tree Learning

Information Service Engineering - Lecture 5 Basic Machine Learning - 5.5 Basic Machine Learning Algorithms II

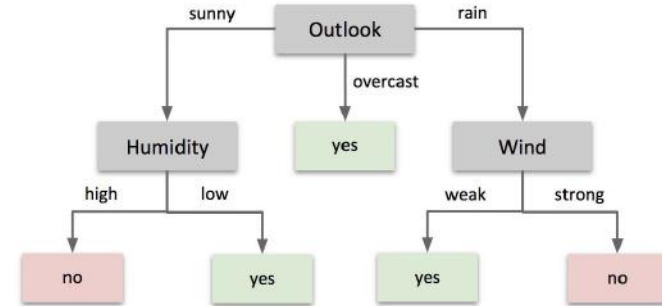
- Constructing a decision tree is easy
 - Given n features, there are $n!$ possible decision trees
- Unfortunately, a decision tree can grow very large
 - Given n features with m possible choices
 - Then a decision tree might grow up to size m^n
- The size of a decision tree depends on the order of its features
- Learning algorithm tries to determine a good ordering
- Learning the simplest (smallest) decision tree is an **NP complete problem** (if you are interested, check: Hyafil & Rivest'76)



Decision Tree Learning - ID3 Algorithm

Information Service Engineering - Lecture 5 Basic Machine Learning - 5.5 Basic Machine Learning Algorithms II

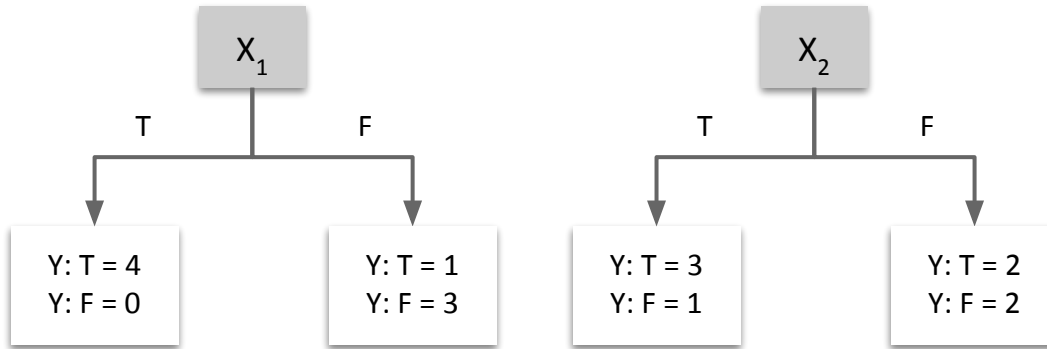
- Resort to a **greedy heuristic**:
 - Start from an empty decision tree
 - Split on next “best” attribute
 - Recurse
- What is best attribute?
- We use **information theory** to guide us



Decision Tree Learning - ID3 Algorithm

Information Service Engineering - Lecture 5 Basic Machine Learning - 5.5 Basic Machine Learning Algorithms II

- Which attribute is better to split on, X_1 or X_2 ?



- Idea:** Use counts at leaves to define probability distributions, so we can measure uncertainty

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

Decision Tree Learning - ID3 Algorithm

Information Service Engineering - Lecture 5 Basic Machine Learning - 5.5 Basic Machine Learning Algorithms II

Entropy

- The **Information Content (Entropy)** H of a feature variable x with n possible feature values is based on it's relative frequency of occurrence (probability):

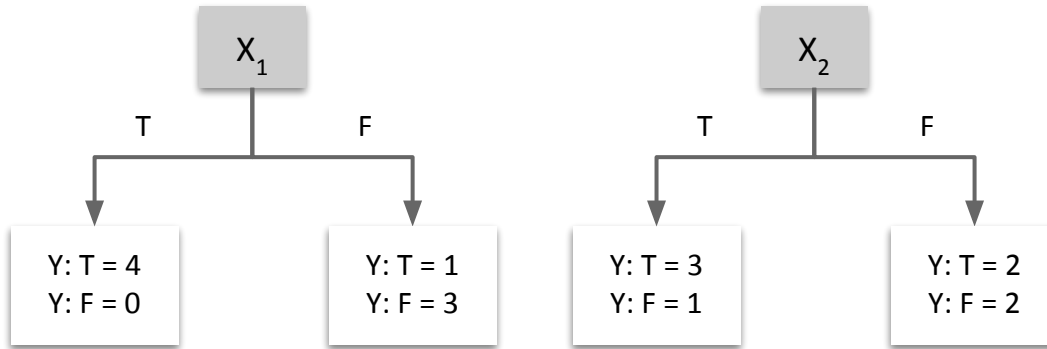
$$H(x) = - \sum_{i=1}^n p_i \cdot \log_2(p_i)$$

- The unit to measure information content is **bit** (*binary digit, basic indissoluble information unit*)

Decision Tree Learning - ID3 Algorithm

Information Service Engineering - Lecture 5 Basic Machine Learning - 5.5 Basic Machine Learning Algorithms II

- Which attribute is better to split on, X_1 or X_2 ?

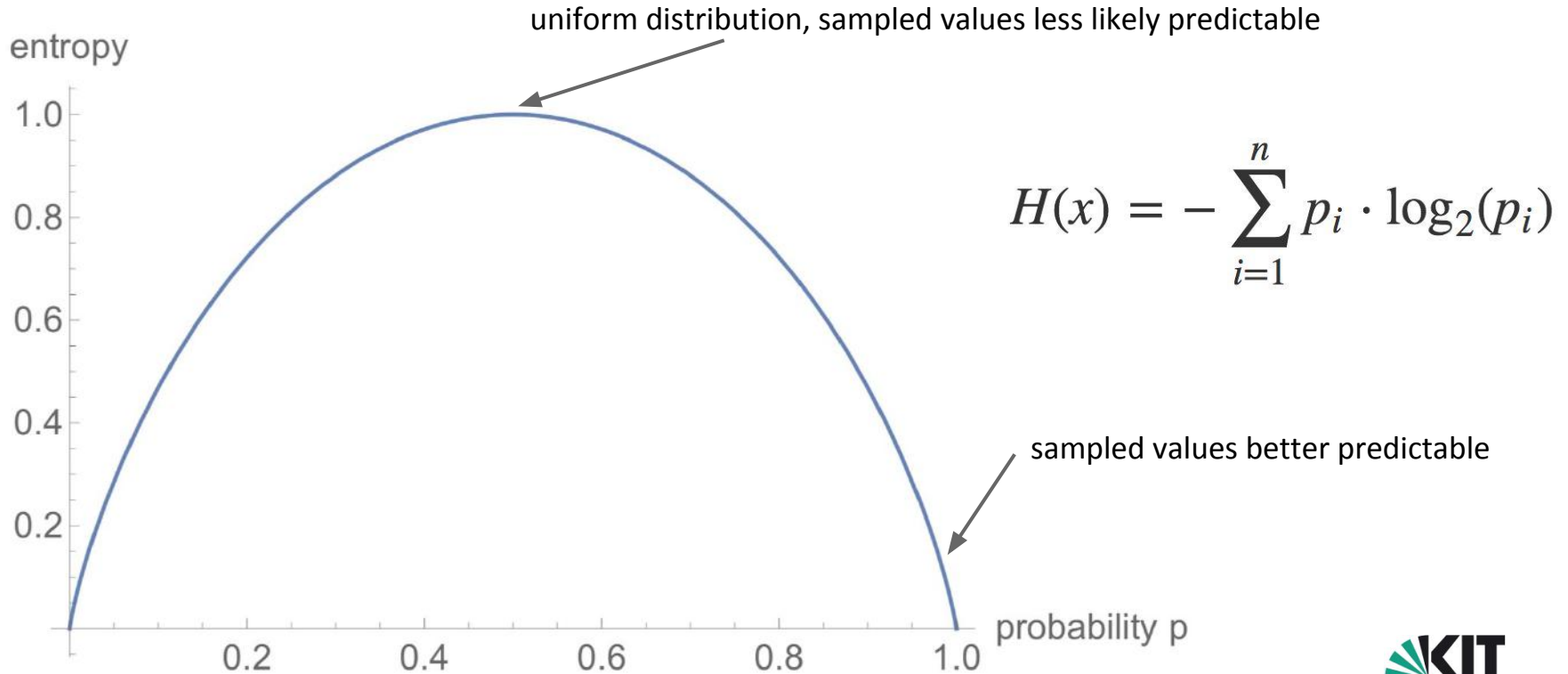


- Entropy of Y :** $[T=5, F=3]$ $H(Y) = -\frac{5}{8} \log_2(\frac{5}{8}) - \frac{3}{8} \log_2(\frac{3}{8}) = 0.95$ bit

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

Decision Tree Learning - ID3 Algorithm

Information Service Engineering - Lecture 5 Basic Machine Learning - 5.5 Basic Machine Learning Algorithms II



Decision Tree Learning - ID3 Algorithm

Information Service Engineering - Lecture 5 Basic Machine Learning - 5.5 Basic Machine Learning Algorithms II

Information Gain

- The **Information Gain** of a feature **Y** due to a feature **X** is a measure of the effectiveness of a feature in classifying the training data.
- It is simply the **expected reduction in entropy** caused by partitioning the examples according to this feature.

$$\begin{aligned} IG(Y|X) &= H(Y) - H(Y|X) \\ &= H(Y) - \sum_{x \in X} \frac{|Y_x|}{|Y|} H(Y_x) \end{aligned}$$

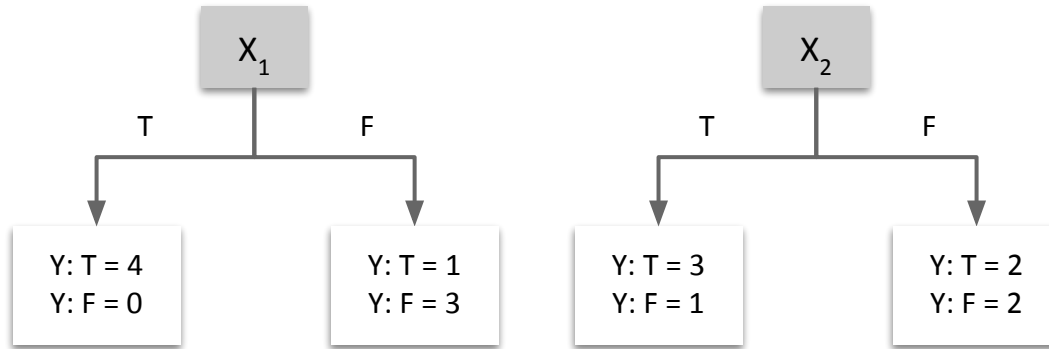
- where x are all possible values of feature X ,
and Y_x the subset of all instances in Y with $X=x$

Decision Tree Learning - ID3 Algorithm

Information Service Engineering - Lecture 5 Basic Machine Learning - 5.5 Basic Machine Learning Algorithms II

- Which attribute is better to split on, X_1 or X_2 ?

$$IG(Y|X) = H(Y) - \sum_{x \in X} \frac{|Y_x|}{|Y|} H(Y_x)$$



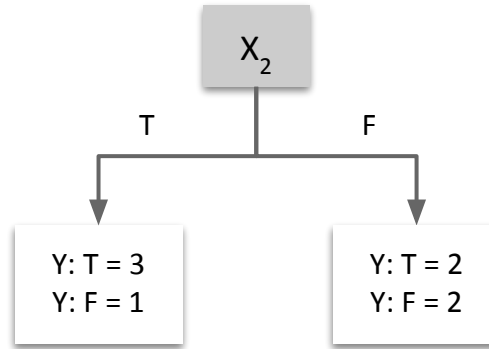
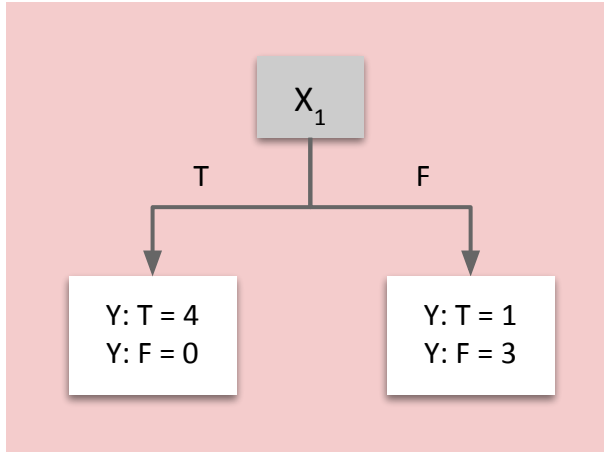
- $IG(Y|X_1) = 0.95 - (4/8 H(X_1=\text{True}) + 4/8 H(X_1=\text{False})) = 0.95 - (0 + 0.40) = \mathbf{0.55}$
 $IG(Y|X_2) = 0.95 - (4/8 H(X_2=\text{True}) + 4/8 H(X_2=\text{False})) = 0.95 - (0.40 + 0.5) = \mathbf{0.05}$

Decision Tree Learning - ID3 Algorithm

Information Service Engineering - Lecture 5 Basic Machine Learning - 5.5 Basic Machine Learning Algorithms II

- Which attribute is better to split on, X_1 or X_2 ?

$$IG(Y|X) = H(Y) - \sum_{x \in X} \frac{|Y_x|}{|Y|} H(Y_x)$$

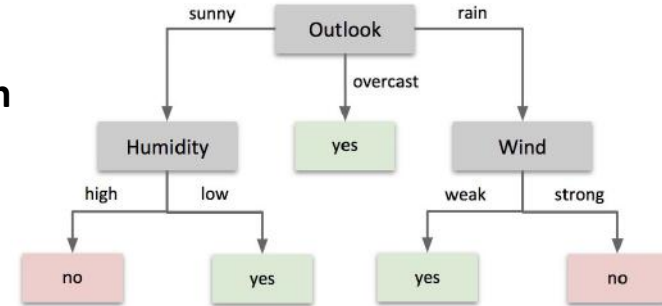


- $IG(Y|X_1) = 0.95 - (4/8 H(X_1=\text{True}) + 4/8 H(X_1=\text{False})) = 0.95 - (0 + 0.40) = \mathbf{0.55}$
 $IG(Y|X_2) = 0.95 - (4/8 H(X_2=\text{True}) + 4/8 H(X_2=\text{False})) = 0.95 - (0.40 + 0.5) = \mathbf{0.05}$

Decision Tree Learning - ID3 Algorithm

Information Service Engineering - Lecture 5 Basic Machine Learning - 5.5 Basic Machine Learning Algorithms II

- ID3 greedy heuristic:
 1. Start from an empty decision tree
 2. Split on next “best” attribute, i.e. the attribute with the **next highest information gain**
 3. Recurse



ID3 Algorithm - Example

Information Service Engineering - Lecture 5 Basic Machine Learning - 5.5 Basic Machine Learning Algorithms II

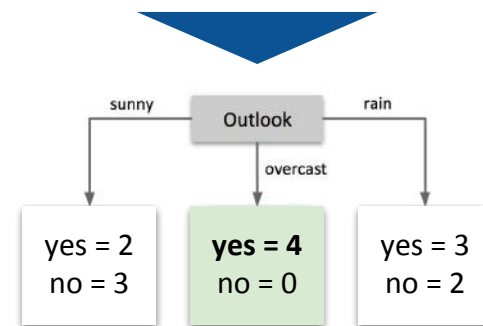
Outlook	Temperature	Humidity	Wind	PlayTennis
sunny	hot	high	weak	no
sunny	hot	high	strong	no
overcast	hot	high	weak	yes
rain	mild	high	weak	yes
rain	cool	low	weak	yes
rain	cool	low	strong	no
overcast	cool	low	strong	yes
sunny	mild	high	weak	no
sunny	cool	low	weak	yes
rain	mild	low	weak	yes
sunny	mild	low	strong	yes
overcast	mild	high	strong	yes
overcast	hot	low	weak	yes
rain	mild	high	strong	no

$IG(\text{PlayTennis}, \text{Outlook}) = 0.246$

$IG(\text{PlayTennis}, \text{Humidity}) = 0.151$

$IG(\text{PlayTennis}, \text{Wind}) = 0.048$

$IG(\text{PlayTennis}, \text{Temperature}) = 0.029$

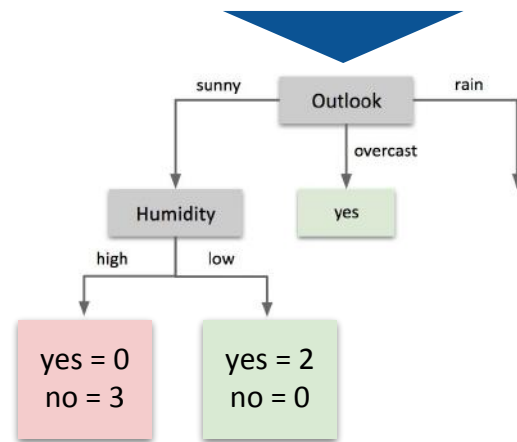


ID3 Algorithm - Example

Information Service Engineering - Lecture 5 Basic Machine Learning - 5.5 Basic Machine Learning Algorithms II

Outlook	Temperature	Humidity	Wind	PlayTennis
sunny	hot	high	weak	no
sunny	hot	high	strong	no
overcast	hot	high	weak	yes
rain	mild	high	weak	yes
rain	cool	low	weak	yes
rain	cool	low	strong	no
overcast	cool	low	strong	yes
sunny	mild	high	weak	no
sunny	cool	low	weak	yes
rain	mild	low	weak	yes
sunny	mild	low	strong	yes
overcast	mild	high	strong	yes
overcast	hot	low	weak	yes
rain	mild	high	strong	no

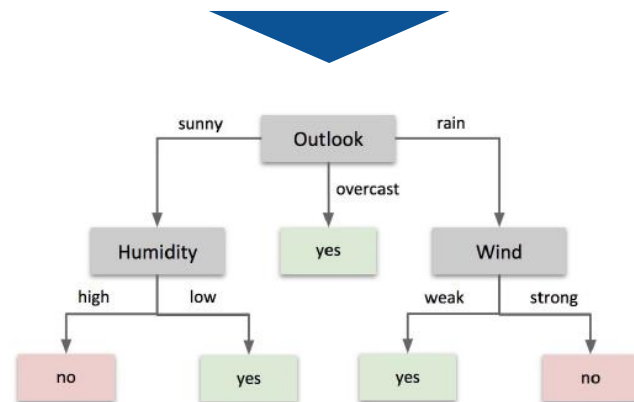
$IG(\text{Outlook}, \text{Humidity}) = 0.97$
 $IG(\text{Outlook}, \text{Temperature}) = 0.57$
 $IG(\text{Outlook}, \text{wind}) = 0.019$



ID3 Algorithm - Example

Information Service Engineering - Lecture 5 Basic Machine Learning - 5.5 Basic Machine Learning Algorithms II

Outlook	Temperature	Humidity	Wind	PlayTennis
sunny	hot	high	weak	no
sunny	hot	high	strong	no
overcast	hot	high	weak	yes
rain	mild	high	weak	yes
rain	cool	low	weak	yes
rain	cool	low	strong	no
overcast	cool	low	strong	yes
sunny	mild	high	weak	no
sunny	cool	low	weak	yes
rain	mild	low	weak	yes
sunny	mild	low	strong	yes
overcast	mild	high	strong	yes
overcast	hot	low	weak	yes
rain	mild	high	strong	no



Decision Trees - Pros and Cons

Information Service Engineering - Lecture 5 Basic Machine Learning - 5.5 Basic Machine Learning Algorithms II

- **Simple to understand** and interpret.
 - Able to handle both **numerical and categorical data**.
 - Requires **little data preparation**.
 - Uses a **white box model**.
 - Possible to **validate a model using statistical tests**.
 - Performs well with **large datasets**.
 - **Mirrors human decision making** more closely than other approaches
 - **BEWARE: Decision Tree models easily overfit**
- **DT Learning Algorithms:**
 - **ID3** (Iterative Dichotomiser 3)
 - **C4.5** (successor of ID3)
 - **CART** (Classification And Regression Tree)
 - **CHAID** (CHi-squared Automatic Interaction Detector). Performs multi-level splits when computing classification trees.
 - **MARS**: extends decision trees to handle numerical data better.



Unsupervised Learning

Picture References:

- P.4 A physician giving a medicine to a sick man in bed, and a surgeon, supervised by a physician, amputating the leg of seated patient, representing pharmacy and surgery respectively. Engraving, 1646.
https://commons.wikimedia.org/wiki/File:A_physician_giving_a_medicine_to_a_sick_man_in_bed,_and_a_surgeon_supervised_by_a_physician,_amputating_the_leg_of_seated_patient,_representing_pharmacy_and_surgery_respectively._Engraving,_1646._Wellcome_V0016715.jpg
- P.22: Hiroshige: Kanbara (1833), https://commons.wikimedia.org/wiki/File:Hiroshige16_kanbara.jpg