*** Applied Machine Learning Fundamentals ***
Decision Trees and Ensembles

Daniel Wehner

SAP SE

October 21, 2019

Duale Hochschule Baden-Württemberg

Agenda October 21, 2019

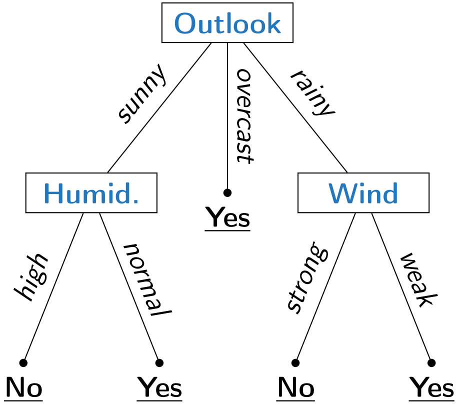
1. Introduction
   Lecture Overview
2. Wrap-Up
   Summary
   Self-Test Questions
   Recommended Literature and further Reading

Section:

Introduction

What we want...

Outlook	Temperature	Humidity	Wind	PlayGolf
sunny	hot	high	weak	**no**
sunny	hot	high	strong	**no**
overcast	hot	high	weak	**yes**
rainy	mild	high	weak	**yes**
rainy	cool	normal	weak	**yes**
rainy	cool	normal	strong	**no**
overcast	cool	normal	strong	**yes**
sunny	mild	high	weak	**no**
sunny	cool	normal	weak	**yes**
rainy	mild	normal	weak	**yes**
sunny	mild	normal	strong	**yes**
overcast	mild	high	strong	**yes**
overcast	hot	normal	weak	**yes**
rainy	mild	high	strong	**no**
rainy	mild	normal	strong	???



What are Decision Trees?

- Decision trees are induced in a **supervised** fashion
- Originally invented by *Ross Quinlan* (1986)
- Decision trees are grown **recursively** → *'divide-and-conquer'*
- A decision tree consists of:

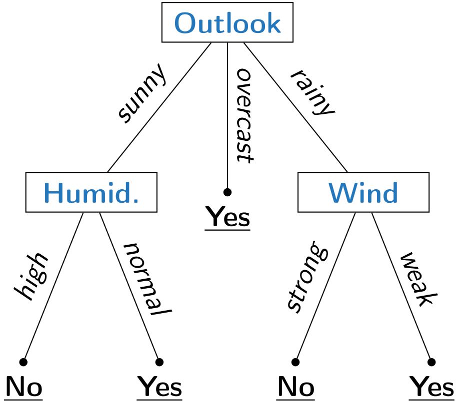
Nodes	Each node corresponds to an attribute test
Edges	One edge per possible test outcome
Leaves	Class label to predict

Classifying new Instances

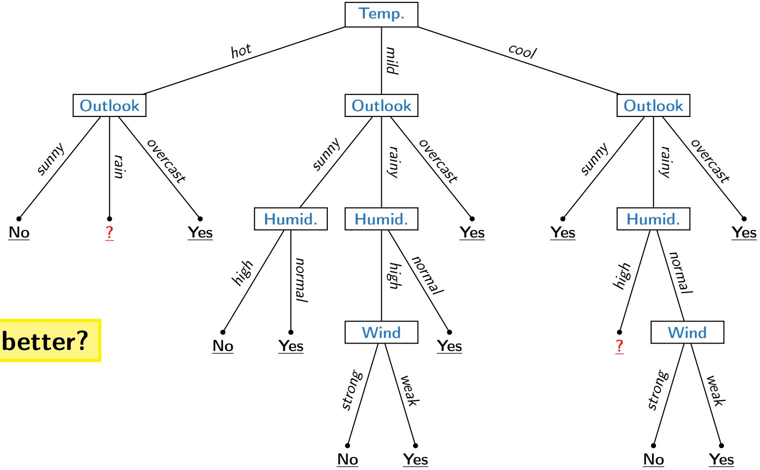
- Suppose we get a new instance:

`Outlook`	rainy
`Temperature`	mild
`Humidity`	normal
`Wind`	strong

- **What is its class?**
- Answer: **No**



Another Decision Tree...



Is this one better?

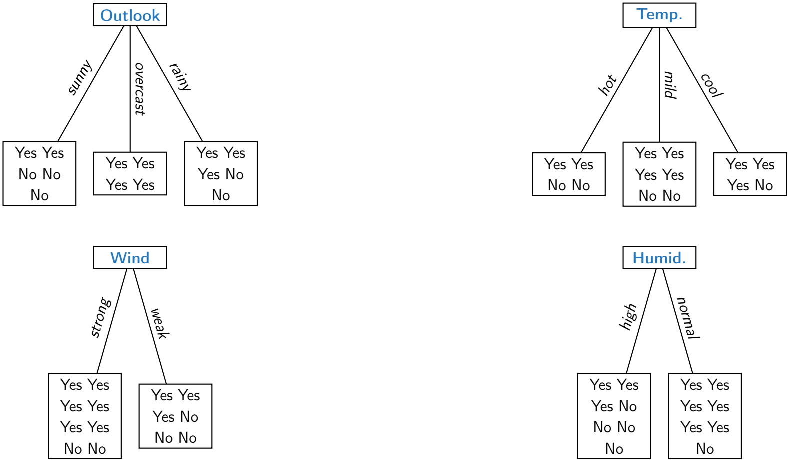
Inductive Bias of Decision Trees

- Complex models tend to **overfit** the data and do not generalize well
- Small decision trees are preferred

> **Occam's razor:**
> **'More things should not be used than are necessary.'**

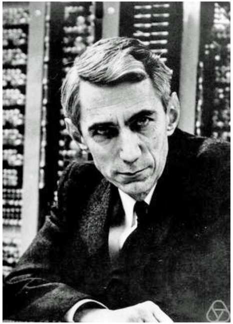
- **Prefer the simplest hypothesis that fits the data!**

The Root of all Evil... Which Attribute to choose?



Finding a proper Attribute

- Simple and small trees are preferred
  - Data in successor node should be **as pure as possible**
  - I. e. nodes containing one class only are preferable
- **Question:** How can we express this thought as a mathematical formula?
- **Answer:**
  - Entropy (*Claude E. Shannon*)
  - Originates in the field of **information theory**



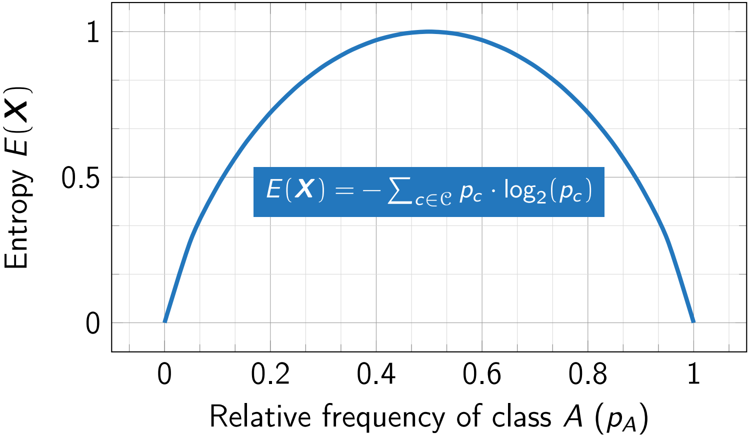
Measure of Impurity: Entropy

- Entropy is a measure of chaos in the data (measured in bits)
- **Example:** Consider two classes A and B ($\mathcal{C} = \{A, B\}$)

$$
\begin{array}{lll}
E(\{\boldsymbol{A}, \boldsymbol{A}, \boldsymbol{A}, \boldsymbol{A}, \boldsymbol{A}, \boldsymbol{A}, \boldsymbol{A}, \boldsymbol{A}\}) & \rightarrow 0 & Bits \\
E(\{\boldsymbol{A}, \boldsymbol{A}, \boldsymbol{A}, \boldsymbol{A}, \boldsymbol{A}, \boldsymbol{A}, B, B\}) & \rightarrow 0.81 & Bits \\
E(\{\boldsymbol{A}, \boldsymbol{A}, \boldsymbol{A}, \boldsymbol{A}, B, B, B, B\}) & \rightarrow 1 & Bit \\
E(\{\boldsymbol{A}, \boldsymbol{A}, B, B, B, B, B, B\}) & \rightarrow 0.81 & Bits \\
E(\{B, B, B, B, B, B, B, B\}) & \rightarrow 0 & Bits
\end{array}
$$

If both classes are equally distributed, the entropy function E reaches its maximum. Pure data sets have minimal entropy.

Measure of Impurity: Entropy (Ctd.)



Measure of Impurity: Entropy (Ctd.)

Entropy formula:

$$E(\mathcal{D}) = -\sum_{c \in \mathcal{C}} p_c \cdot \log_2 p_c \tag{1}$$

- Where p_c denotes the relative frequency of class $c \in \mathcal{C}$
- **Weather data:**

$$\mathcal{C} = \{yes, no\} \qquad \text{i.e.} \qquad p_{yes} = {}^{9}/_{14} \quad \text{and} \quad p_{no} = {}^{5}/_{14}$$

$$E(\mathcal{D}) = -\sum_{c \in \mathcal{C}} p_c \cdot \log_2(p_c) = -({}^{9}/_{14} \cdot \log_2({}^{9}/_{14}) + {}^{5}/_{14} \cdot \log_2({}^{5}/_{14})) = \mathbf{0.9403}$$

Section:

Wrap-Up

Summary

Lecture Overview

Unit I: Machine Learning Introduction

Self-Test Questions

Recommended Literature and further Reading

Thank you very much for the attention!

Topic: *** Applied Machine Learning Fundamentals *** Decision Trees and Ensembles
Date: October 21, 2019

Contact:
Daniel Wehner (D062271)
SAP SE
daniel.wehner@sap.com

Do you have any questions?