Machine Learning

Prof. Sergei Gleyzer

Lecture

PH451, PH551

February 4, 2025

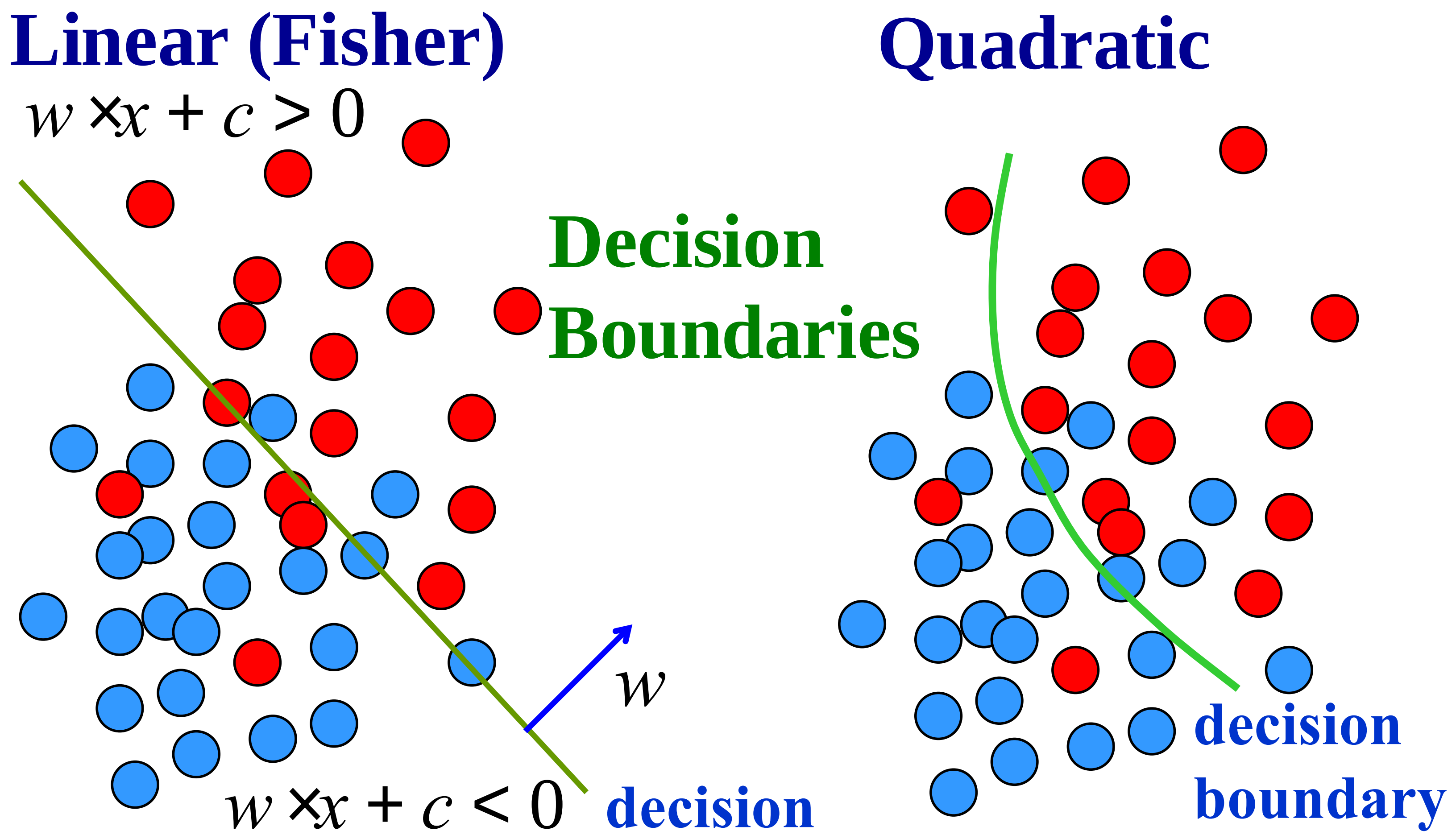
Announcements

- **Hands-on #3/Reading HW #2**
  - **due next Tue**
  - **Textbook: Chapter 6**

ML Methods (partial list)

- Fisher, Quadratic
- Naïve Bayes (Likelihood)
- Kernel Density Estimation
- Random Grid Search
- Rule ensembles
- Boosted decision trees
- Random forests
- Deep neural networks
- Support vector machines
- Genetic algorithms

Linear and Quadratic



Outline

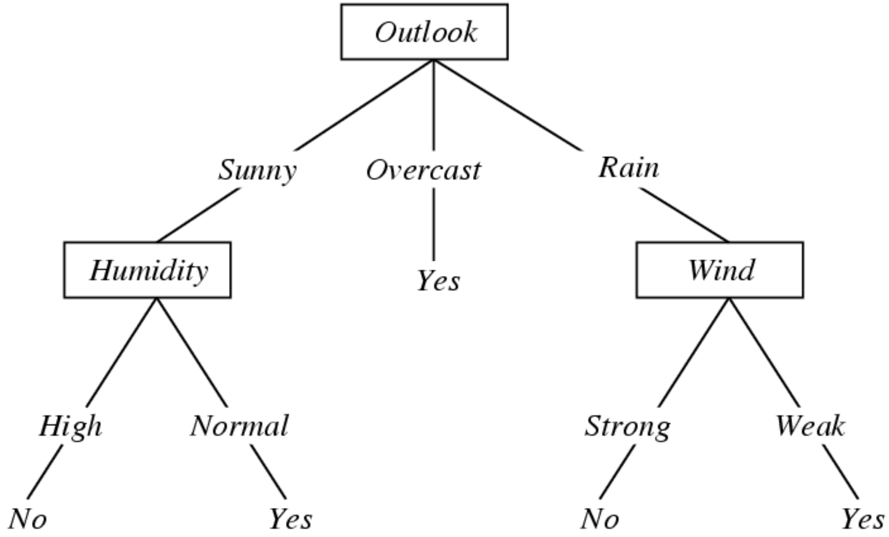
- **Decision Trees**

Binary Decision Trees

Decision Trees

Decision trees are recursively constructed **multidimensional histograms**

- Each leaf associated to the value (**class**) of f(x) to be approximated
- **Tennis-Playing** Decision Tree: f(outlook, humidity, wind, T)



Decision Trees

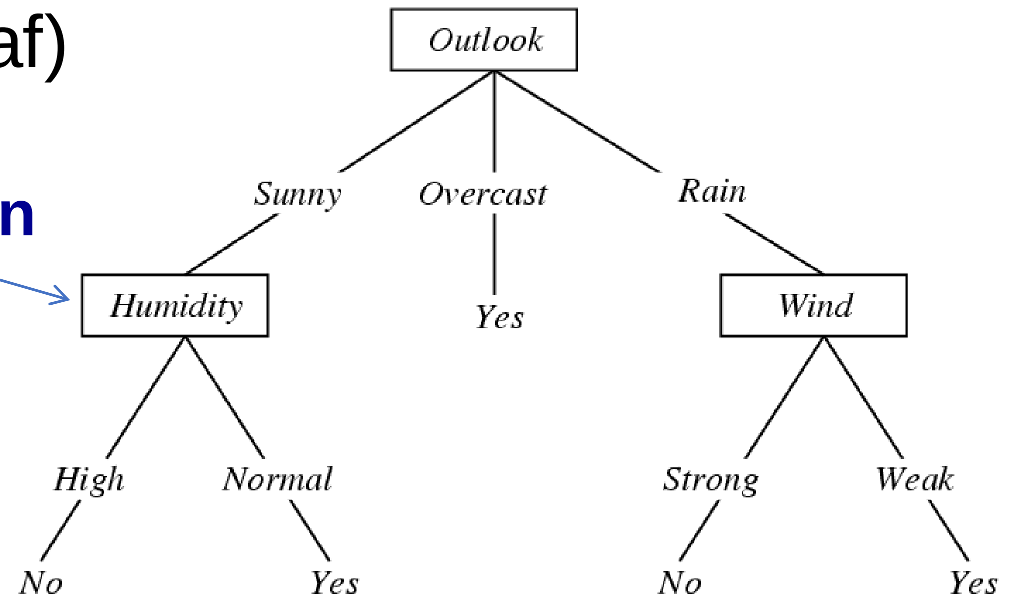
Each **internal** node: test one attribute X_i

Each **branch**: selects one value for X_i

Each **leaf** node: predict Y

- Or P(Y|X in leaf)

Decision Node



Decision Tree Learning

Unknown **target function** f: **X→Y**

- Set of possible instances **X**
  - each **instance** is a feature vector

    e.g.

  <Humidity = High, Wind = weak, Outlook = rain, Temp = hot>

Decision Tree Learning

Input:

- Training examples $\{<x^i,y^i>\}$

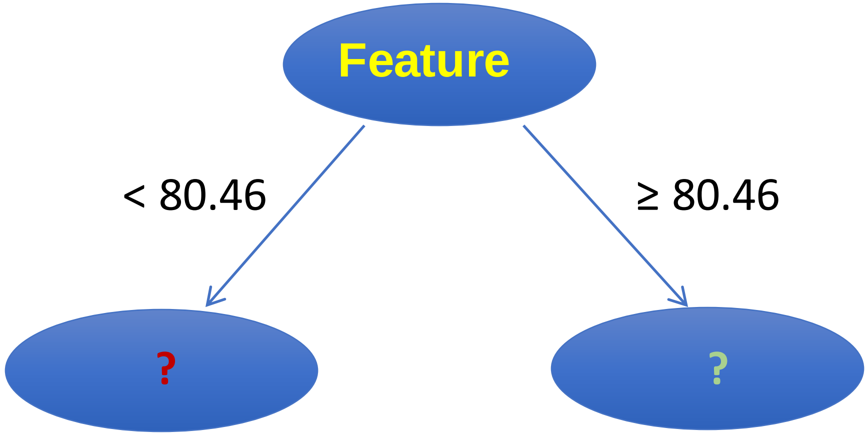
Output

- Hypothesis **h** $\in$ **H** that
  best approximates target function f
- Tree sorts **x** to **leaf**, which assigns **y**

Decision Trees

Building a tree:

- Scan along each variable and propose a **DECISION**
  - A cut on value that maximizes class separation (binary branching)

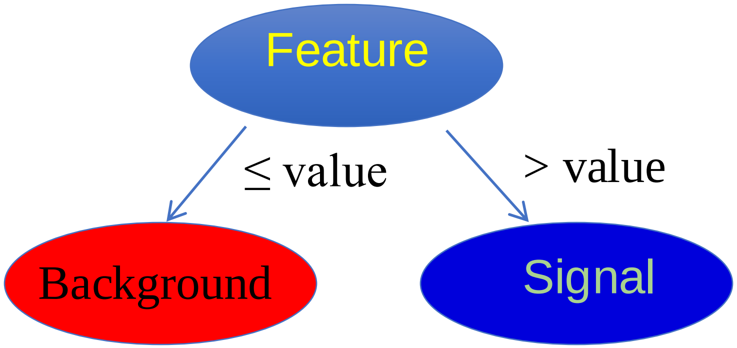


Decision Trees

Choose **decision** that leads to greatest separation among classes **signal**/**background**

- Based on the information gained from split
  - Build regions of **increasing purity**
  - **Stop** when no further improvement from additional branching
  - Reach **terminal** node (leaf) and assign purity-based class

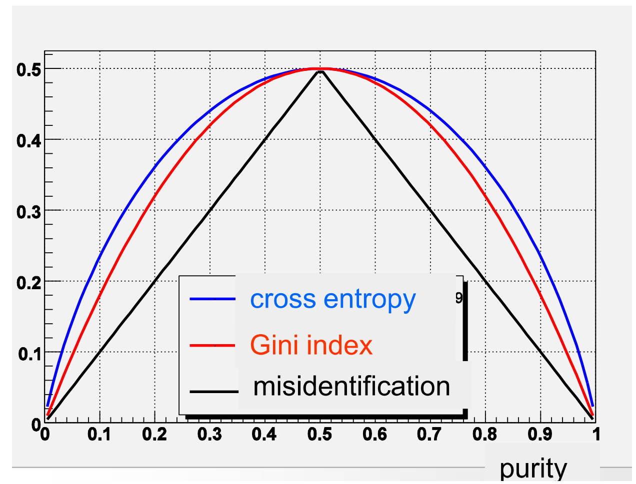
$$\frac{N_{signal}}{N_{signal} + N_{background}}$$

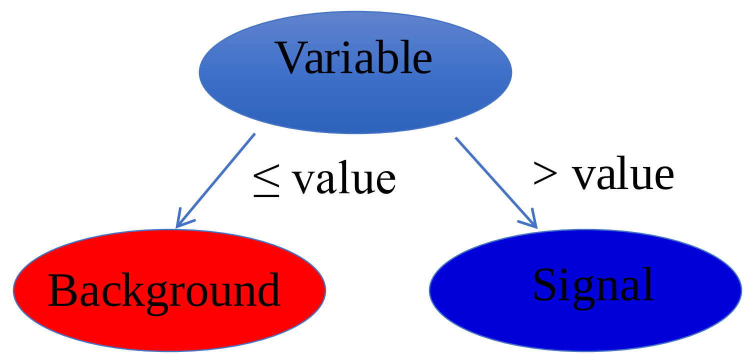


Separation Gain

Measures of Separation Gain

- Cross-Entropy
  - -p lnp + (1-p) ln(1-p)
- Gini Index
  - p (1 – p)
- Want to lower entropy due to split

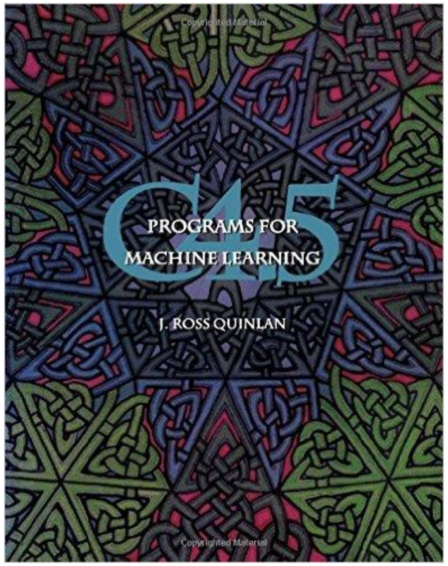




Decision Trees

Decision Trees

- Classic **ML tool** for
  - **decision trees**
  - **rules**
  - **boosted classifiers**
- Written by **J.R. Quinlan**
  - Name: ID3 → C4.5 → C5.0
    - Use c5.0 to familiarize with decision tree classifiers



Pruning

Decision trees can become large and complex and risk **over-fitting** the data

Pruning: remove parts of the tree that are less powerful or possibly **noisy**

- start from the leaves and work back up

Pruned trees smaller in size, easier to interpret