Summary

(part II, Atsuto)

Classification

- We would like to enable a computer to learn from data to answer a question - "What is it?"

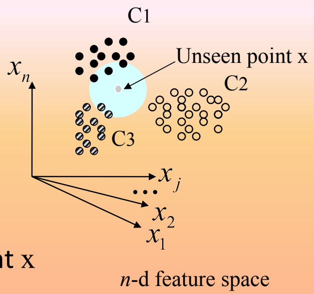
  You are given sample data (for finding patterns).

The framework of classification

- Training phase: to give the concept of classes to a machine using labeled data
- Testing phase: to determine the class of new unseen (unlabeled) data

Nearest Neigbour methods

- Compute the distances to all the samples from new data x
- Pick *k* neighbours that are nearest to *x*



→ Majority vote to classify point x

(Nearest Neighbour is 1-NN)

- How does *k*-NN compare to 1-NN ?

Entropy

How to measure information gain?

The Shannon information content of an outcome is:

$$\log_2 \frac{1}{p_i}$$

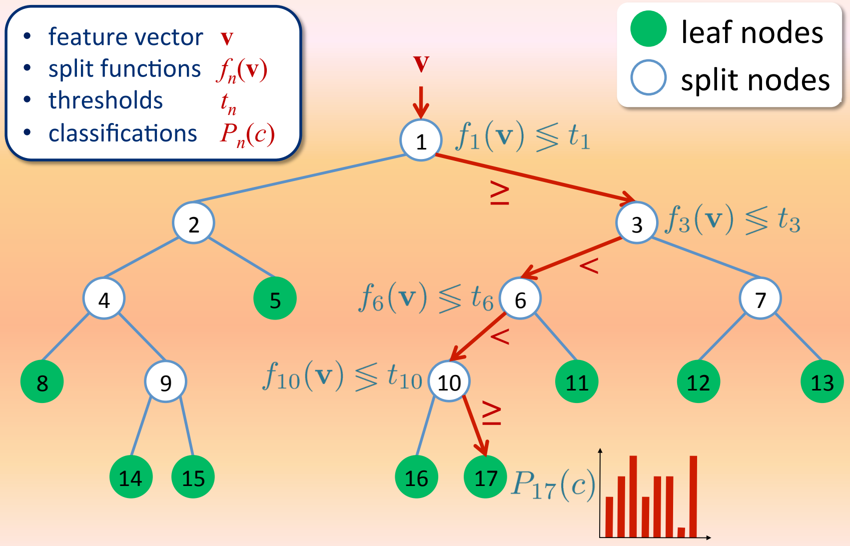
(p_i : probability for event i)

The *Entropy* — measure of uncertainty (unpredictability)

$$\text{Entropy} = \sum_i -p_i \log_2 p_i$$

is a sensible measure of expected information content.

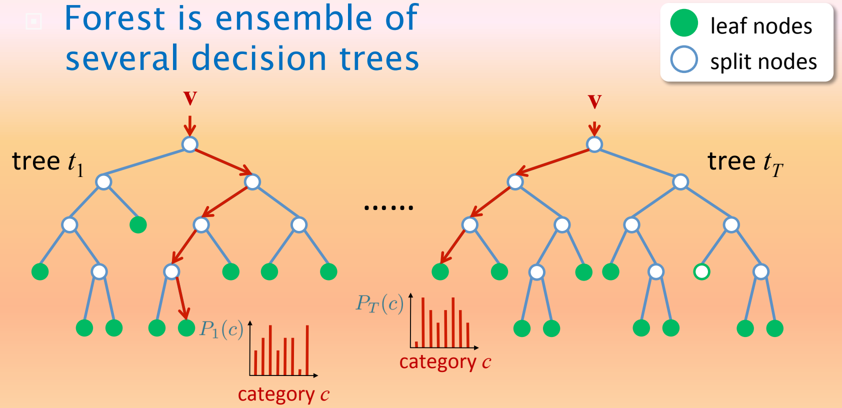
The Basics: Binary Decision Trees



(figure courtesy to Jamie Shotton)

A Forest of Trees

- Forest is ensemble of several decision trees



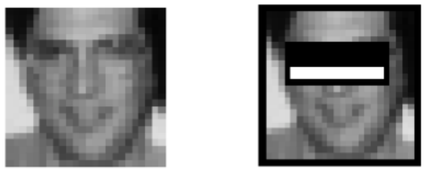
- classification is $P(c|\mathbf{v}) = \sum_{t=1}^{T} P_t(c|\mathbf{v})$

[Amit & Geman 97]
[Breiman 01]
[Lepetit *et al.* 06]

(figure courtesy to Jamie Shotton)

Boosting

Adaboost



FACE or NON-FACE

Input: $\mathbf{x}$ **Apply filter:** $f^j(\mathbf{x})$ **Output:** $h(\mathbf{x}) = (f^j(\mathbf{x}) > \theta)$

Probability Based Learning

- **Maximum A Posteriori (*MAP*) Estimate:**
  Hypothesis with highest probability given observed data

$$y_{\text{MAP}} = \arg\max_{y \in \mathcal{Y}} P(y \mid \mathbf{x}) = \arg\max_{y \in \mathcal{Y}} \frac{P(\mathbf{x} \mid y)\, P(y)}{P(\mathbf{x})} = \arg\max_{y \in \mathcal{Y}} P(\mathbf{x} \mid y)\, P(y)$$

- **Maximum Likelihood Estimate (*MLE*):**
  Hypothesis with highest likelihood of generating observed data.

$$y_{\text{MLE}} = \arg\max_{y \in \mathcal{Y}} P(\mathbf{x} \mid y)$$

Useful if we do not know prior distribution or if it is uniform.

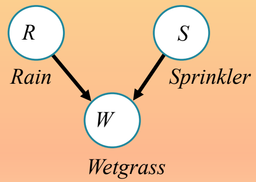
Probability Based Learning

Models for the (joint) distributions

- Naïve Bayes
  - Assumes all features to be independent
- Bayesian networks
  - Models dependencies between features
- Mixtures of Gaussians
  - Expectation-Maximization (EM) algorithms (related to K-means clustering)

Bayesian networks

- A probabilistic method for modeling dependencies between random variables through a **graph structure:**
  - random variables by nodes
  - conditional dependencies by edges (directed acyclic graph)

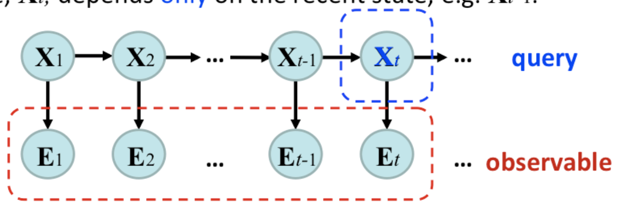


→ Probabilistic inference system $P(R,S,W) = P(W|R,S)P(S)P(R)$

Synonyms:
= probabilistic network / causal network / knowledge map ...

Inference from sequential data

- The series of models are linked by a markov assumption: the state, $\mathbf{X}_t$, depends only on the recent state, e.g. $\mathbf{X}_{t-1}$.



- As we deal with time-series variables, **the question is**:

What is the posterior probability for a set of query variables, $\mathbf{X}_t$, given past and present evidence variables, $\mathbf{E}_{1:t}$?

$$P(\mathbf{X}_t \mid \mathbf{E}_{1:t}) = ?$$