

## Outline

- Introduction
- Bayesian Probability Theory
- Sequence Prediction and Data Compression
- Decision Trees and Ensemble Learning
- Bayesian Networks

## Review of Lecture 1

- Introduction
  - §18.1, 18.2, 18.5 of textbook
  - supervised learning
  - learning is not always possible
  - when learning is possible, empirical risk minimisation is good
  - justification: laws of large numbers
  - small finite function classes are PAC learnable
  - Problems: may not have enough examples, minimising empirical risk is computationally intractable

## Review of Lecture 2

- Bayesian Probability Theory
  - chapter 13 and §20.1 of textbook
  - Bayes rule and its many uses
  - Bayes rule applied to learning
  - Information-theoretic interpretation of MAP estimator
  - Bayesian optimal predictor
  - Problems: prescribed solution often intractable to compute, prior may be hard to construct
- Lectures 1 and 2 study machine learning principles and theoretically optimal solutions without regard to computational issues.

## Review of Lecture 3

- Sequence Prediction and Data Compression
  - Based on F.M.J. Willems et al, The Context Tree Weighting Method: Basic Properties  
[http://www.sps.ele.tue.nl/members/F.M.J.Willems/RESEARCH\\_files/CTW/ResearchCTW.htm](http://www.sps.ele.tue.nl/members/F.M.J.Willems/RESEARCH_files/CTW/ResearchCTW.htm)
  - A relatively rare case where the Bayesian optimal predictor can be efficiently computed
  - Good prediction performance implies good compression performance

## Review of Lecture 4

- Decision Trees and Ensemble Learning
  - §18.3 and 18.4 of textbook
  - an example of what we do in practice when we can't compute the theoretically optimal solutions
  - decision trees are popular because they are comprehensible (unlike, e.g., neural networks)
  - a greedy algorithm that produces smallish decision trees
  - overfitting, underfitting, and cross validations
  - ensemble learning can improve the performance of basic algorithms

## Review of Lecture 5

- Bayesian networks
  - §14.1 and 14.5 of textbook
  - a graphical notation for compactly representing large distributions
  - asserts and exploits conditional independence between variables
  - variable elimination algorithm
  - sampling-based inference algorithms

## Questions?