6.867 Machine learning and neural networks

Tommi Jaakkola
MIT AI Lab
tommi@ai.mit.edu

Lecture 1: introduction

Course administrivia

- Instructor: Prof. Tommi Jaakkola (tommi@ai.mit.edu)
- TA: Nathan Srebro (nati@mit.edu)
- About the course:
  - lectures TR 2.30-4pm in 37-212
  - recitations (initially Wed 1pm, location tba)
  - midterm (15%), final (30%)
  - 5 problem sets (30%)
  - final project (25%)

  - satisfies the AI requirement for Area II graduate students

Machine learning

- Statistical machine learning
  - principles, methods, and algorithms of learning and inference
  - wide range of applications, e.g., speech recognition, medical diagnosis, image/text retrieval, commercial software, ...
- Topics covered in this course include:

  Classification and regression methods; overfitting, generalization, regularization; feature selection, feature induction, dimensionality reduction; combining multiple models, additive models, boosting; support vector machines, kernels, complexity, VC-dimension; model selection, structural risk minimization; density estimation, parametric, non-parametric clustering, agglomerative, mixture models, semi-supervised clustering; Markov/hidden Markov models, forward-backward and the EM algorithm; Bayesian networks, graphs, model selection, Markov random fields, dynamic Bayesian networks; active learning

Learning

- Some terminology: learning vs. adaptation
  - Adaptation is *temporary* and “resets” itself
    - e.g., brightness adaption of eyes
  - Learning involves a *persistent* change or memory
    - e.g., ?

Learning cont'd

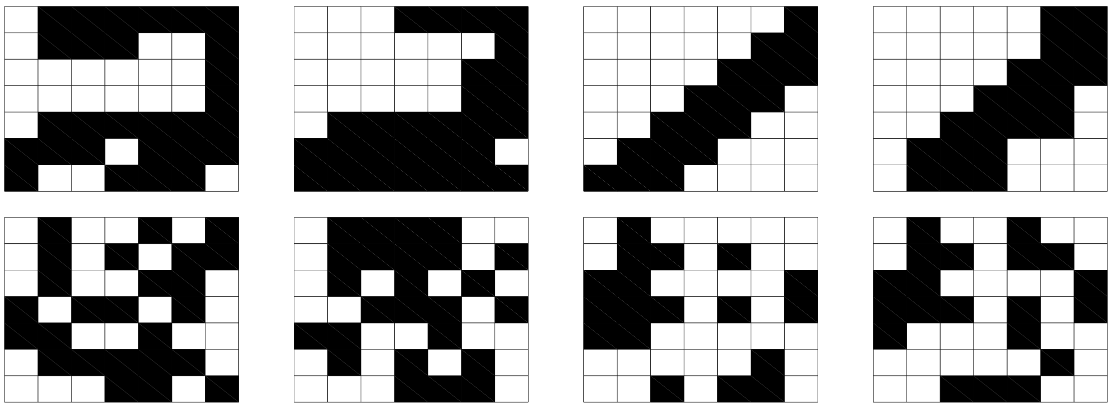
- Scientific progress
  - incremental, revolutions
- Human development/learning
  - learning facts, conceptual change
- Machine learning
  - ?

Learning cont'd

- Summary of concepts:
  1. Representation
     - we need to be able to represent the concept we wish to learn
     - the choice of representations may help or hurt performance
  2. Assumptions, biases
     - no learning takes place without assumptions
     - various types, from representations to more detailed preferences

Representation and assumptions

- Example: image labeling



- A change in the representation that nevertheless preserves the relevant information can preclude learning

Machine learning

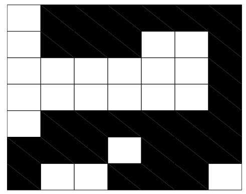
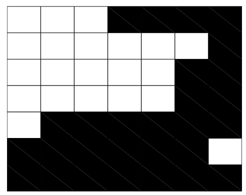
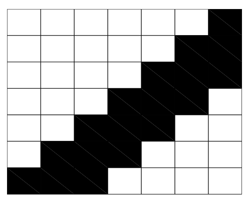
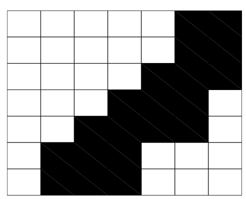
- Research in this area tries to understand why, how, and under what conditions learning takes place
- It is not just *deduction* but *induction* of rules and properties from examples to facilitate prediction and decision making
- Strong ties to statistical inference/estimation (Why?)

Machine learning problems

- Types of learning problems:
  1. Supervised learning (with a teacher)
  2. Unsupervised learning (no teacher, objective)
  3. Reinforcement learning (interactive, temporally extended decision problems)
- This is not a complete nor "precise" list

Supervised learning: classification

Example: digit recognition (8x8 binary digits)

binary digit	target label
	“2”
	“2”
	“1”
	“1”
...	...

- We wish to learn the mapping from digits to labels

Supervised learning: classification

- We are given set of *input examples* $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, where

$$\mathbf{x}_i = (x_{i1} \ldots x_{id})^T, \;\; d = dim(\mathbf{x}_i),$$

  and corresponding *outputs* $\{y_1, \ldots, y_n\}$.

- Our task is to learn a *mapping or function* $f : \mathcal{X} \to \mathcal{Y}$ from inputs $\mathbf{x} \in \mathcal{X}$ to outputs $y \in \mathcal{Y}$ such that

$$y_i \approx f(\mathbf{x}_i), \;\; i = 1, \ldots, n$$

  where n is the number of *training examples*.

"*For any problem there is a simple solution that is wrong*"

Supervised learning

The input/output spaces may be discrete or continuous

- **Classification** (qualitative)

  Continous/discrete inputs, discrete outputs
  - (binary pixel values) $\rightarrow$ digit label
  - (gray level pixel values) $\rightarrow$ digit label

- **Regression** (quantitative)

  Continous/discrete inputs, continuous outputs
  - (interest rate) $\rightarrow$ stock price
  - (stock($t-1$), stock(t)) $\rightarrow$ stock($t+1$)

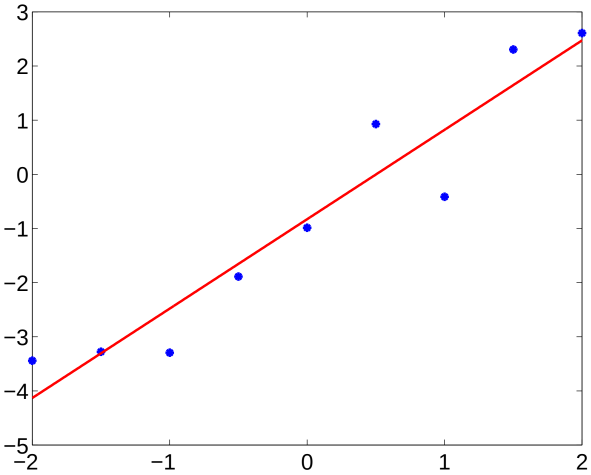
Supervised learning: regression

- Given a set of training examples $\{(\mathbf{x}_1, y_1) \ldots, (\mathbf{x}_n, y_n)\}$, we want to learn a mapping $f : \mathcal{X} \rightarrow \mathcal{Y}$ such that

$$y_i \approx f(\mathbf{x}_i), \;\; i = 1, \ldots, n$$

- We need to *parameterize* or otherwise restrict the set of functions (Why?)

  For example: $f(x) = ax + b$, where we learn the parameters a and b.



- BUT: how did we "fit" the function to the training examples?

Supervised learning: regression

- We define a *loss function* Loss$(y, f(x))$ that measures how much our predictions deviate from the given answers

$$y_i \approx f(\mathbf{x}_i), \;\; \text{Loss}(y_i, f(\mathbf{x}_i)), \;\; i = 1, \ldots, n$$

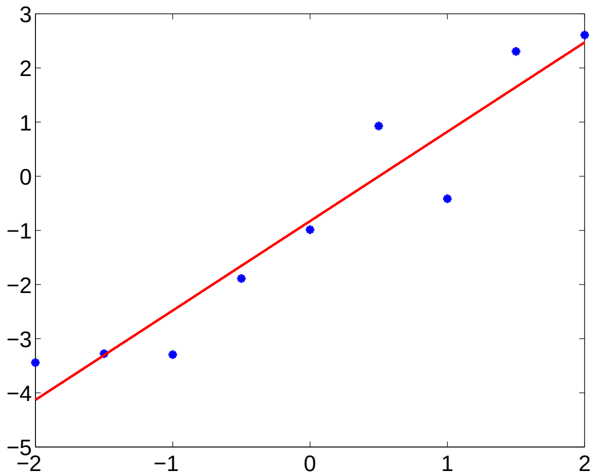
  For example: Loss$(y_i, f(\mathbf{x}_i)) = (y_i - f(\mathbf{x}_i))^2$.

- To "fit" the parameters such as the linear coefficients a and b in $f(x) = ax + b$, we can minimize the *training error/loss*:

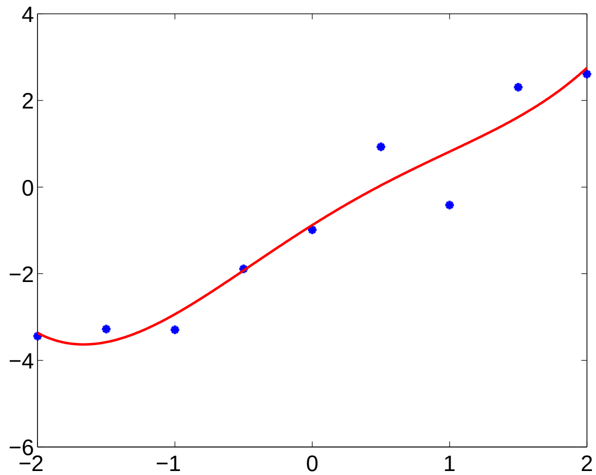
$$\sum_{i=1}^{n} \text{Loss}(y_i, f(\mathbf{x}_i))$$

Supervised learning: regression

- Of course, the more parameters we have in the function, the smaller training error that we can achieve



f = linear

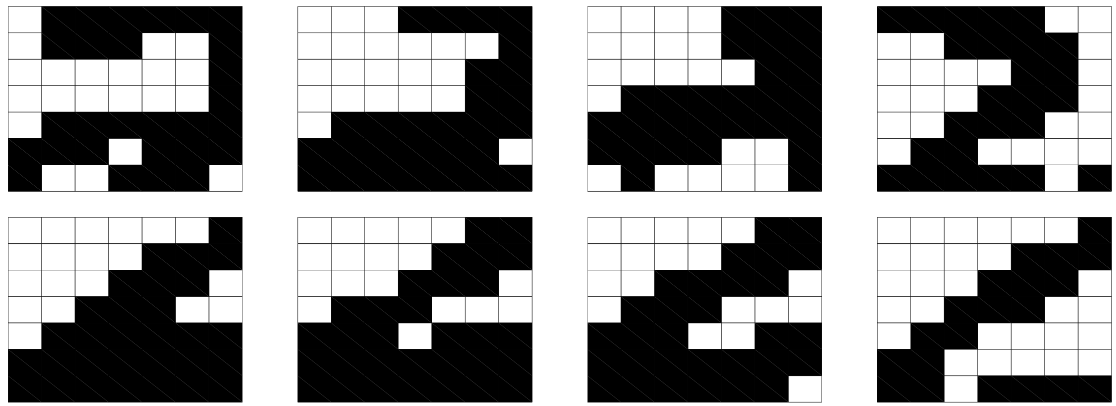


$f = 4^{th}$ order polynomial

- Overfitting
- Generalization
- Model selection

Unsupervised learning: data organization

The digits again...

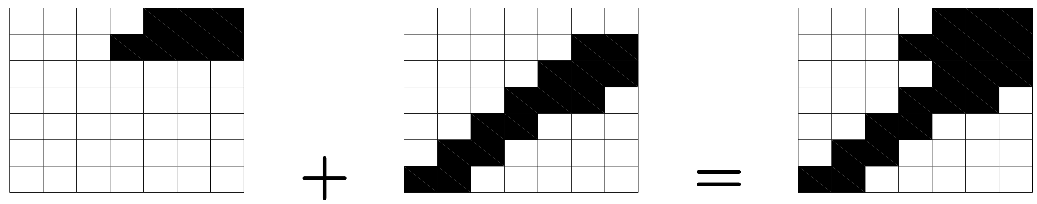


- **Clustering**: group together "similar" digits

Unsupervised learning cont'd

- Density estimation:

  We construct a *generative* probability distribution over the examples

- A closely related problem is how to best *compress* a given set of examples.

Reinforcement learning

- There's no teacher (explicit target labels) but we receive some, possibly noisy, feed-back about how good our actions are

  Take action a $\rightarrow$ receive $Cost(a, \omega)$

  (ω is the state of the "environment")

  Often we do not know the cost function $Cost(a, \omega)$ but only observe its values after taking actions.

- The goal is to select the action that minimizes the cumulative cost

$$\sum_{i=1}^{n} Cost(a, \omega_i)$$

- This becomes more interesting when our actions actually *affect* the state of the environment... games, planning, etc.

Summary

Some of the key concepts we have discussed:

- Assumptions & representation in learning
- Overfitting, generalization, model selection
- Clustering, generative density estimation
- Interactive decision problems