6.867 Machine learning and neural networks

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

Lecture 10: complexity, model selection

Topics

- Complexity
  - shattering, VC-dimension
- Model selection
  - Basic idea
  - Structural risk minimization

Worst case analysis

- How complex a classifier can we estimate on the basis of only a small number of training examples?
  - first we need to *define* exactly what we mean by complexity
  - complexity is often but not always equal to the number of parameters (degrees of freedom) in the model.
- We will define *Vapnik-Chervonenkis dimension* or *VC-dimension* for a set of classifiers that we are considering

VC-dimension: preliminaries

- **A set of classifiers F:**

  For example, this could be the set of all possible linear separators, where $h \in F$ means that

  $$h(\mathbf{x}) = \text{sign}\left(w_0 + \mathbf{w}^T\mathbf{x}\right)$$

  for some values of the parameters $\mathbf{w}, w_0$.

- **Complexity:** how many different ways can we label n training points $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ with classifiers $h \in F$?

  In other words, how many distinct binary vectors

  $$[h(\mathbf{x}_1)\ h(\mathbf{x}_2)\ \ldots\ h(\mathbf{x}_n)]$$

  do we get by trying all $h \in F$?

  $$\begin{array}{cl} [\ \ -1 \quad 1 \quad \ldots \quad 1\ \ ] & h_1 \\ [\ \ \ \ 1 \quad -1 \quad \ldots \quad 1\ \ ] & h_2 \\ & \ldots \end{array}$$

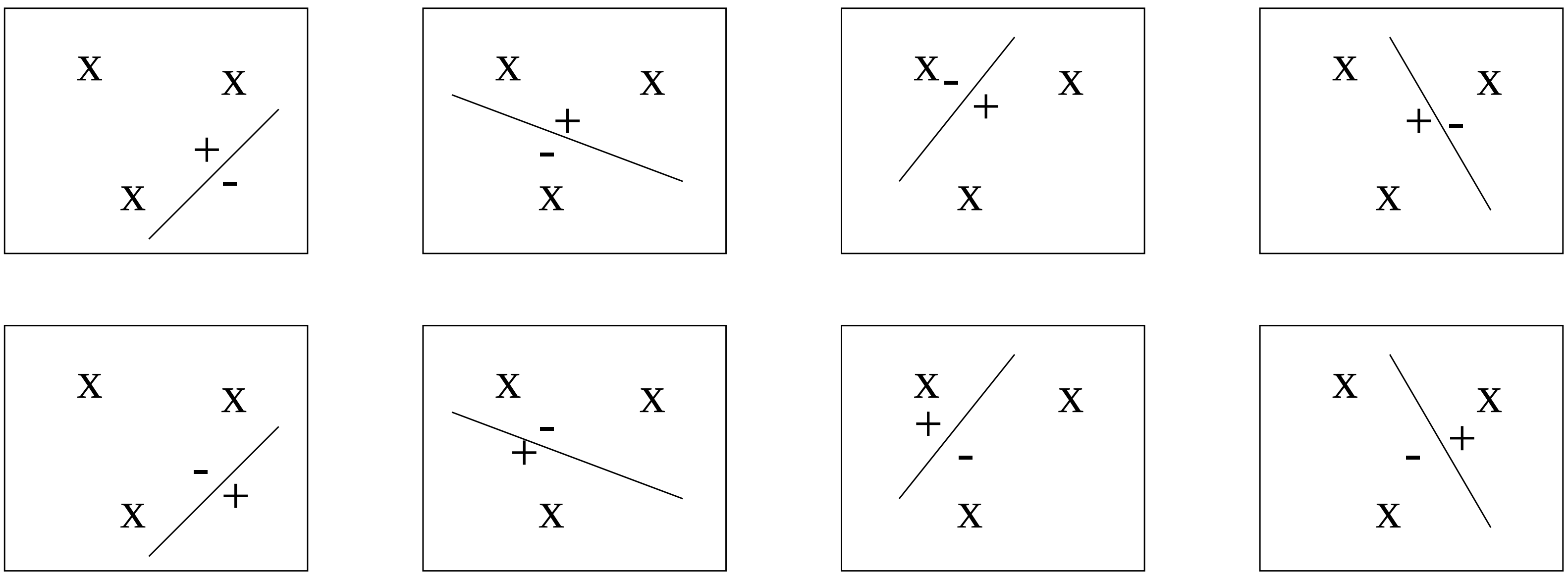
VC-dimension: shattering

- A set of classifiers F *shatters* n points $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ if

$$[h(\mathbf{x}_1)\ h(\mathbf{x}_2)\ \ldots\ h(\mathbf{x}_n)],\ \ h \in F$$

generates all 2^n distinct labelings.

- Example: linear decision boundaries shatter (any) 3 points in 2D

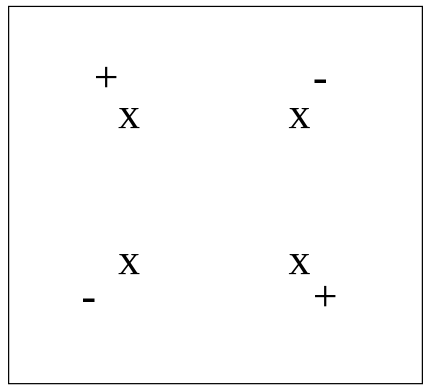


but not any 4 points...

VC-dimension: shattering cont'd

- We cannot shatter 4 points in 2D with linear separators

  For example, the following labeling



  cannot be produced with any linear separator

- More generally: the set of all d-dimensional linear separators can shatter exactly $d+1$ points

VC-dimension

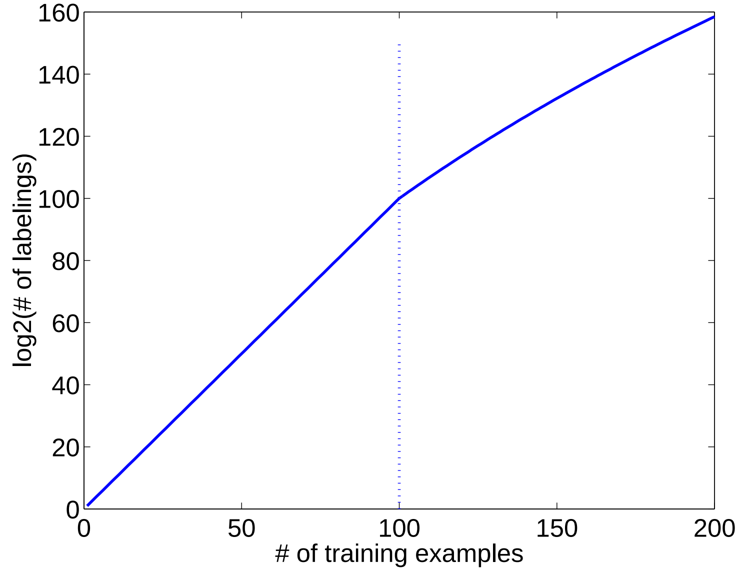
- *The VC-dimension d_{VC} of a set of classifiers F is the largest number of points that F can shatter*
- This is a combinatorial concept and doesn't depend on what type of classifiers we use, only how "flexible" the set of classifiers is

  Example: Let F be a set of classifiers defined in terms of linear combinations of m **fixed** basis functions

  $$h(\mathbf{x}) = \mathrm{sign}\,(\,w_0 + w_1\phi_1(\mathbf{x}) + \ldots + w_m\phi_m(\mathbf{x})\,)$$

  d_{VC} is at most $m + 1$ regardless of the form of the fixed basis functions.

Learning and VC-dimension



- The number of labelings that the set of classifiers can generate over n points increases sub-exponentially after $n > d_{VC}$ (in this case $d_{VC} = 100$)

Learning and VC-dimension

- *Finite VC-dimension is necessary and sufficient for (exponentially) fast convergence of a learning method*

  By convergence we mean here:

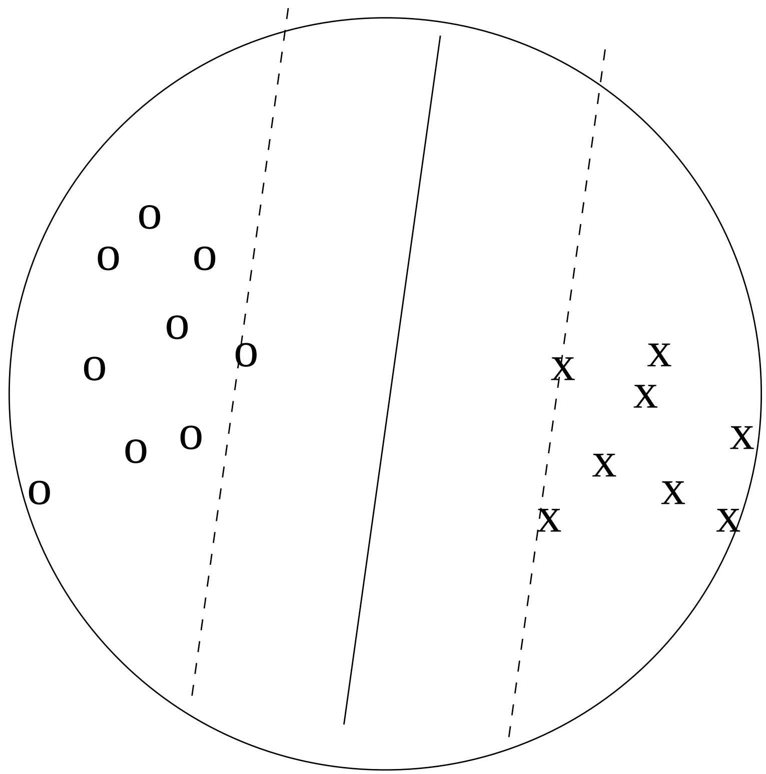
$$\overbrace{\frac{1}{n}\sum_{i=1}^{n} \text{Loss}(y_i, h(\mathbf{x}_i))}^{\text{Empirical loss}} - \overbrace{E\{\,\text{Loss}(y, h(\mathbf{x}))\,\}}^{\text{Expected loss}} \rightarrow 0$$

  uniformly for all $h \in F$. Here $\text{Loss}(y, h(\mathbf{x})) = 1$ if $y \neq h(\mathbf{x})$ and zero otherwise (so called zero-one loss)

- This result holds for **any** underlying probability distribution from which the examples and the labels are generated

Extensions: complexity and margin

- The number of possible labelings of points with large margin can be dramatically less than the (basic) VC-dimension



- The set of separating hyperplaces which attain margin γ or better for examples within a sphere of radius R has VC-dimension bounded by $d_{VC}(\gamma) \leq R^2/\gamma^2$

Topics

- Model selection
  - Basic idea
  - Structural risk minimization

Model selection

- Model selection concerns with trying to balance the complexity of the model with the fit to the training data
- We need to have a (preferably) nested sequence of models of increasing complexity

  Model 1 d_1
  Model 2 d_2
  Model 3 d_3

  where $d_1 \leq d_2 \leq d_3 \leq \ldots$

- Basic formulation: we derive a model selection criterion:

$$\text{Criterion} = \text{(empirical) score} + \text{Complexity penalty}$$

Model selection cont'd

- We aim to balance the trade-off between the model complexity and the fit to the training data

$$\text{Criterion} = \text{(empirical) score} + \text{Complexity penalty}$$

- There are a number of (related) model selection criteria

1. Statistical hypothesis test
2. Minimum description/message length (MDL/MML)
3. Structural risk minimization

   etc.

Structural risk minimization

- We have a nested sequence of models of increasing complexity; complexity measured in terms of VC-dimension (or refinements)

  Model 1 $d_{VC} = d_1$
  Model 2 $d_{VC} = d_2$
  Model 3 $d_{VC} = d_3$

  where $d_1 \leq d_2 \leq d_3 \leq \ldots$

- Basic formulation: we derive an upper *bound* on the expected loss

  $$\text{Expected loss} \leq \text{Empirical loss} + \text{Complexity penalty}$$

  and select the model that gives the lowest *bound*.

Example

- Models of increasing complexity

Model 1	$K(\mathbf{x}_1, \mathbf{x}_2) = (1 + (\mathbf{x}_1^T \mathbf{x}_2))$
Model 2	$K(\mathbf{x}_1, \mathbf{x}_2) = (1 + (\mathbf{x}_1^T \mathbf{x}_2))^2$
Model 3	$K(\mathbf{x}_1, \mathbf{x}_2) = (1 + (\mathbf{x}_1^T \mathbf{x}_2))^3$
...	...

- These are nested, i.e.,

$$F_1 \subseteq F_2 \subseteq F_3 \subseteq \ldots$$

  where F_k refers to the set of possible decision boundaries that the model k can represent.

- Still need to derive the criterion...