CS340 Machine learning
Lecture 5
Learning theory cont'd

Some slides are borrowed from Stuart Russell and Thorsten Joachims

Inductive learning

- Simplest form: learn a function from examples

f is the target function

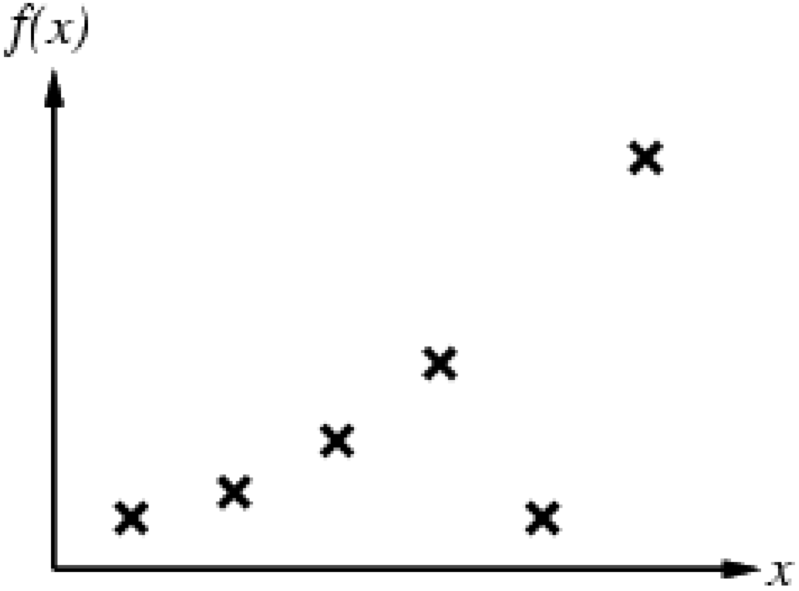
An example is a pair (x, $f(x)$)

Problem: find a hypothesis *h*
such that $h \approx f$
given a training set of examples

(This is a highly simplified model of real learning:
- Ignores prior knowledge
- Assumes examples are given)

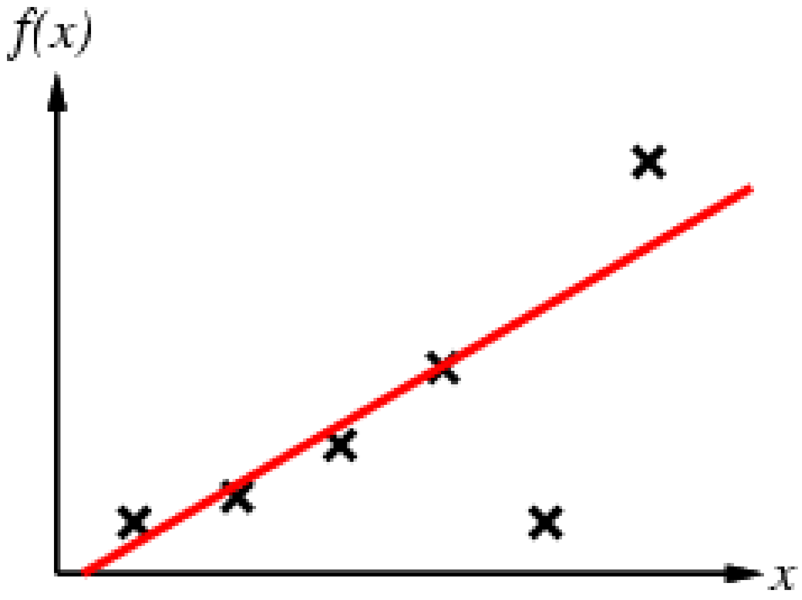
Inductive learning method

- Construct/adjust *h* to agree with *f* on training set
- (*h* is consistent if it agrees with *f* on all examples)
- E.g., curve fitting:



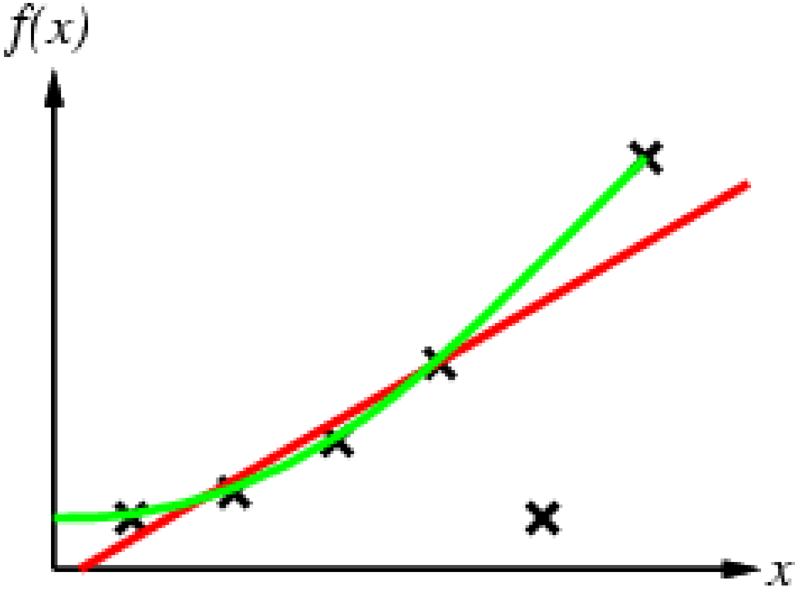
Inductive learning method

- Construct/adjust *h* to agree with *f* on training set
- (*h* is consistent if it agrees with *f* on all examples)
- E.g., curve fitting:



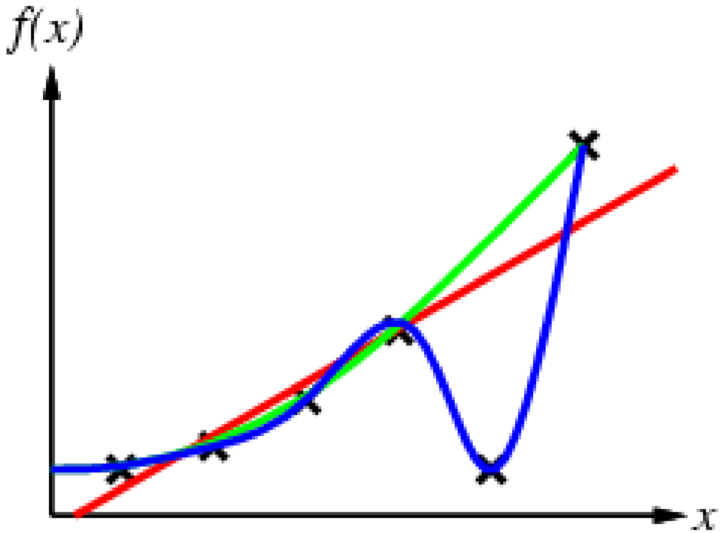
Inductive learning method

- Construct/adjust *h* to agree with *f* on training set
- (*h* is consistent if it agrees with *f* on all examples)
- E.g., curve fitting:



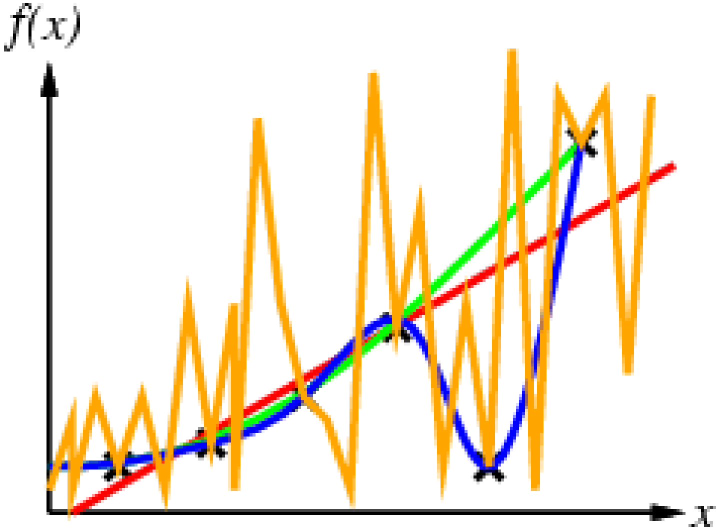
Inductive learning method

- Construct/adjust *h* to agree with *f* on training set
- (*h* is consistent if it agrees with *f* on all examples)
- E.g., curve fitting:



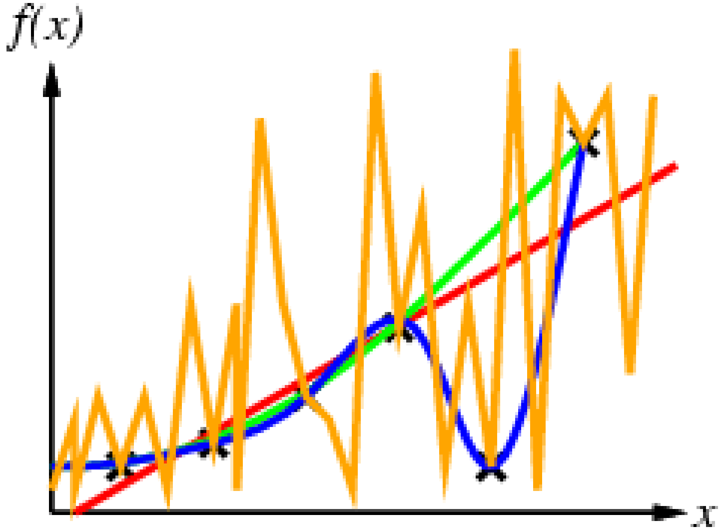
Inductive learning method

- Construct/adjust *h* to agree with *f* on training set
- (*h* is consistent if it agrees with *f* on all examples)
- E.g., curve fitting:



Ockham's razor

- Ockham’s razor: prefer the simplest hypothesis consistent with data
- Why?
- A simpler hypothesis is less likely to be correct "by chance" and is therefore more likely to generalize well



Ockham's razor: why?

- If we have 2 hypotheses with equally small training error, how can we pick the right one?
- If we pick the wrong one, with enough data, we will eventually find out.
- The amount of data we need (to be sure we pick the right hypothesis) depends on the complexity of the hypothesis class.
- There are more ways to "accidently" fit the training data if we have a very flexible hypothesis class.
- If we want to avoid the possibility of overfitting, we should restrict the complexity of the hypothesis class, or use a larger training set.
- (Of course, an overly simple hypothesis class may underfit.)

Empirical risk minimization

- The ERM principle says to pick the hypothesis with the lowest error on the training set.
- When does low training error guarantee low generalization error?
- Equivalently, given an observed error rate on a finite sample, can we bound the expected error rate on future data? (Empirical process theory)
- Bound will depend on size of the hypothesis class.
- A very complex hypothesis class can fit anything, so if it has low training error, this does not mean it will have low generalization error.

Statistical learning theory

- SLT is concerned with establishing conditions under which we can say

$$p(|error_{train}(h) - error_{true}(h)| \leq \epsilon) \geq 1 - \delta$$

- We say that h is **probably approximately correct**
- This statement holds if the hypothesis class is sufficiently constrained and/or the training set size is sufficiently large

Hoeffding/ Chernoff bound

- Imagine estimating the probability of heads from m iid coin tosses $x_1, ..., x_m$, $x_i \cancel{2} \in \{0,1\}$
- The probability of making an error of size ε is bounded by

$$p\left(\left|\frac{1}{m}\sum_{i=1}^{m} x_i - \theta\right| > \epsilon\right) \leq 2e^{-2m\epsilon^2}$$

Training vs generalization error

- Let S = training set, h(A(S)) be the hypothesis learned by algorithm A on S
- Let $err_S(h)$ be error of h on sample S, and $err_P(h)$ be the true expected error on distribution P
- We want to be sure low training error will give rise to low generalization error

$$p(|err_S(h(A(S)) - err_P(h(A(S))| \geq \epsilon \quad \leq \quad \delta$$

- We use the union and Chernoff bounds

$$\begin{aligned} p(\max_i |err_S(h_i) - err_P(h_i)| \geq \epsilon) &\leq |H| \, p(\exists i. \; |err_S(h_i) - err_P(h_i)| \geq \epsilon) \\ &\leq 2|H|e^{-2m\epsilon^2} \end{aligned}$$

Bounds on err_{train} - err_{true}

- Hence wp ¸ 1-δ,

$$err_{true} < err_{train} + \sqrt{\frac{\log|H| + \log\frac{1}{\delta}}{2m}}$$

- The 2nd term on RHS is the growth function

$$\phi(m, |H|, \delta) = \sqrt{\frac{\log|H| + \log\frac{1}{\delta}}{2m}}$$

Sample complexity

- To ensure

$$p(|err_S(h(A(S)) - err_P(h(A(S))| \geq \epsilon) \quad \leq \quad \delta$$

we need this many samples

$$m \geq \frac{1}{2\epsilon^2} \left(\log|H| + \log\frac{1}{\delta} \right)$$

Finite H, zero training error

- Suppose H is finite, and there exists and h with zero training error ("truth is in the hypothesis space")
- We showed last time that prob. exists h 2 H with high true error rate, but zero training error (i.e., h is consistent), is bounded by

$$p(\exists h \in H. err_S(h) = 0,\ err_P(h) > \epsilon) \leq |H|(1-\epsilon)^m \leq |H|e^{-\epsilon m}$$

  which is tighter than

$$p(\exists h \in H. |err_S(h) - err_P(h)| > \epsilon) \leq 2|H|e^{-2m\epsilon^2}$$

- Tighter bounds means lower sample complexity.

PAC bounds for finite H, zero training error

- Partition H into H_ε, an ε "ball" around f^{true}, and $H_{bad} = H \setminus H_\varepsilon$
- What is the prob. that a "seriously wrong" hypothesis $h_b \cancel{2} \in H_{bad}$ is consistent with m examples (so we are fooled)? We can use a union bound

$$\begin{aligned} error(h_b) &> \epsilon \\ p(h_b \text{ agrees with 1 example }) &\leq 1-\epsilon \\ p(h_b \text{ agrees with m examples }) &\leq (1-\epsilon)^m \end{aligned}$$

The prob of finding such an h_b is bounded by

$$\begin{aligned} p(H_{bad} \text{ contains a consistent hypothesis}) &\leq |H_{bad}|(1-\epsilon)^m \\ &\leq |H|(1-\epsilon)^n \end{aligned}$$

Infinite H

- What if H is infinite?
- Union bound no longer works.
- Also, many hypotheses may be very similar (eg rectangles of slightly different size).
- Roughly speaking, we replace log |H| with VC(H).

VC dimension

- Consider a sample S of size m.
- The set of all possible binary labelings realizable by hypothesis class H on S is

$$\Pi_H(S) = \{(h(x_1), \ldots, h(x_m)) : h \in H\}$$

- H shatters S if H can produce all possible labelings

$$|\Pi_H(S)| = 2^m$$

- The VC dimensions of H is equal to the maximal number d of examples that can be shattered
- Intuitively, VC = number free parameters.

VC bounds on err_{train} - err_{true}

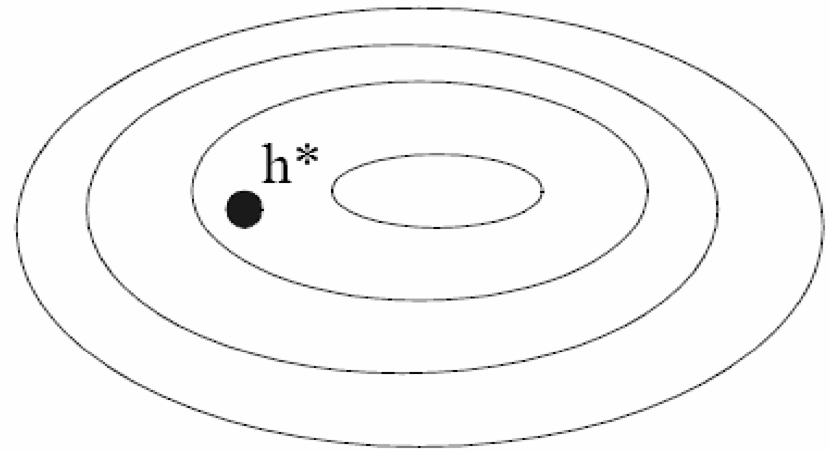
- Thm: wp ¸ 1-δ, with d=VCD(H)

$$
\begin{aligned}
err_{true} &\leq err_{train} + \Phi(m, d, \delta) \\
\Phi(m, d, \delta) &= \sqrt{\frac{d\left(\log\frac{2m}{d} + 1\right) + \log\frac{4}{\delta}}{m}}
\end{aligned}
$$

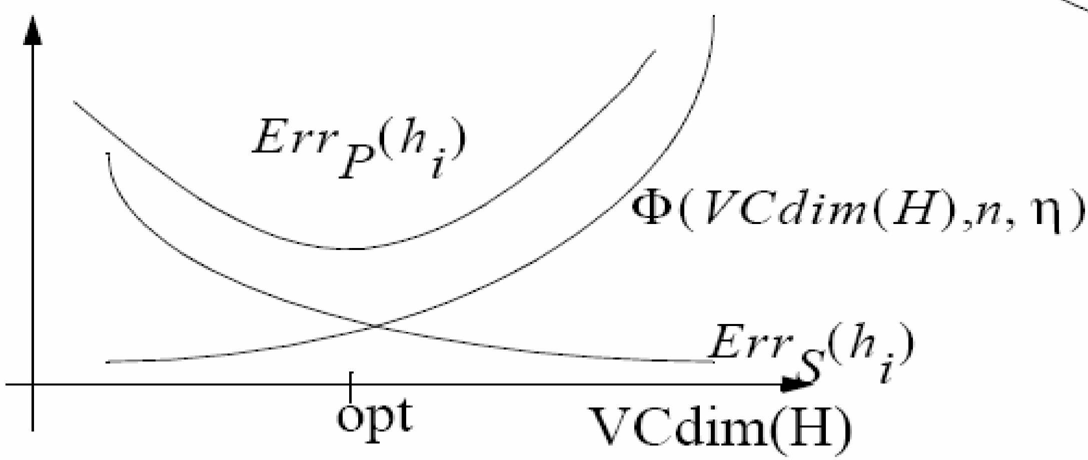
Structural risk minimization

$$Err_P(h_i) \leq Err_S(h_i) + \Phi(VCdim(H), n, \eta)$$

Idea: Structure on hypothesis space.



Goal: Minimize upper bound on true error rate.



Or use cross validation!

Data dependent bounds

- The bound $\Phi(m,d,\delta)$ is independent of the observed data set, and is therefore very loose.
- More complex bounds can be derived.
- These depend on the **margin,** i.e., the degree of overlap between positive and negative examples