Introduction to Statistical Learning Theory

Lecture 1

What is learning?

"The activity or process of gaining knowledge or skill by studying, practicing, being taught, or experiencing something."

Merriam Webster dictionary

We will focus on *supervised learning*

The set-up:

- An input space $\mathcal{X}$. Examples: $\mathbb{R}^n$, images, texts, sound recordings, etc.
- An output space $\mathcal{Y}$. Examples: $\{\pm 1\}$, $\{1, ..., k\}$, $\mathbb{R}$.
- An **unknown** distribution $\mathcal{D}$ on $\mathcal{X} \times \mathcal{Y}$.
- A loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$. Examples: $0-1$ loss, square loss.
- A set of m **i.i.d** samples $(x_1, y_1), ..., (x_m, y_m)$ sampled from the distribution $\mathcal{D}$.

The goal: return a function (hypothesis) $h : \mathcal{X} \to \mathcal{Y}$ that minimizes the expect loss (risk) with respect to $\mathcal{D}$ i.e. find h that minimizes $L_{\mathcal{D}}(h) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell(h(x), y)]$

Goal of this course: Try to analyse what can we say about the expected risk $L_{\mathcal{D}}(h)$ of the unknown distribution given only a random sample.

We will mainly ignore computational issues, focus on statistical analysis.

This is a purely theoretical course - no programming involved.

Requires good understanding on basic probability.

Pass/fail grade, based only on homework.

- Computer vision: face recognition, face identification, pedestrian detection, pose estimation, ect.
- NLP: spam filtering, machine translation, sentiment analysis, etc.
- Speech recognition.
- Medical diagnostics.
- Fraud detection.
- Many more...

There are a few main paradigms in solving a learning problem:

- Generative approach - try to fit $P(x, y)$ by some parametric model, and use it to determine the optimal y given x.
- Discriminative approach - try to fit $P(y|x)$ directly by some parametric model.
- Agnostic approach - narrow yourself to some hypothesis space $\mathcal{H}$ and try to return the best hypothesis in $\mathcal{H}$.

We will focus on the agnostic approach.

The strength of the agnostic approach is that it doesn't assume anything on $\mathcal{D}$, but its weakness is that it depends on the quality of $\mathcal{H}$.

We want to find h^* that minimizes the risk (expected loss) -
$h^* = \arg\min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) = \arg\min_{h \in \mathcal{H}} \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell(h(x), y)]$.

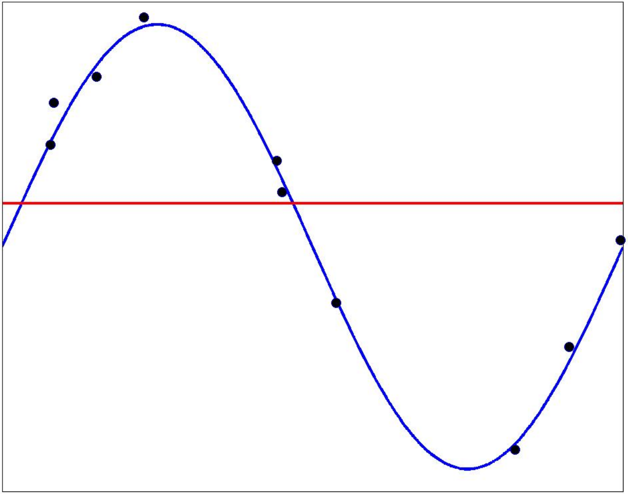
We will minimize the empirical risk -
$h_{ERM} = \arg\min_{h \in \mathcal{H}} L_S(h) = \arg\min_{h \in \mathcal{H}} \frac{1}{m} \sum_{i=1}^{m} \ell(h(x_i), y_i)$.

Consider the following scenario: $\mathcal{X} = [0, 2\pi]$ with uniform distribution, $\mathcal{Y} = \mathbb{R}$ and let ℓ be the square loss $\ell(y_1, y_2) = (y_1 - y_2)^2$. We define the probability on y (give x) as $y = sin(x) + \mathcal{N}(0, 0.05)$, and we are given $m = 10$ data points.

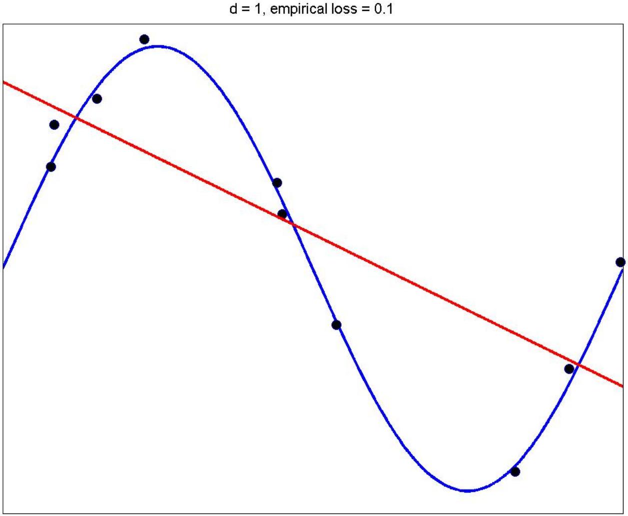
We will show how ERM preforms with $\mathcal{H}_d$ the set of polynomials of degree d.

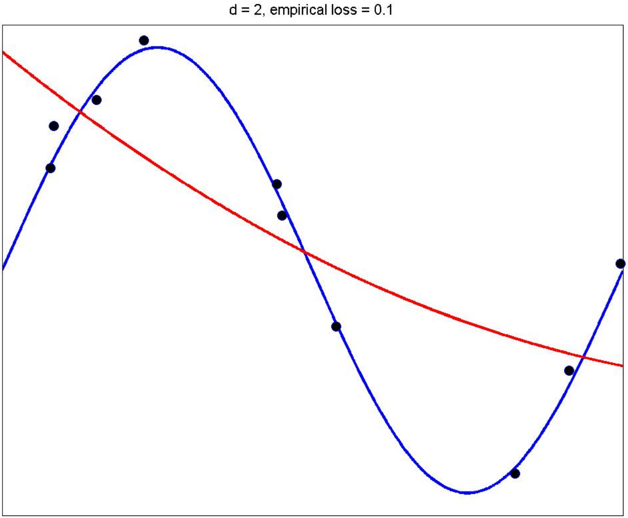
Enlightening example

d = 0, empirical loss = 0.3

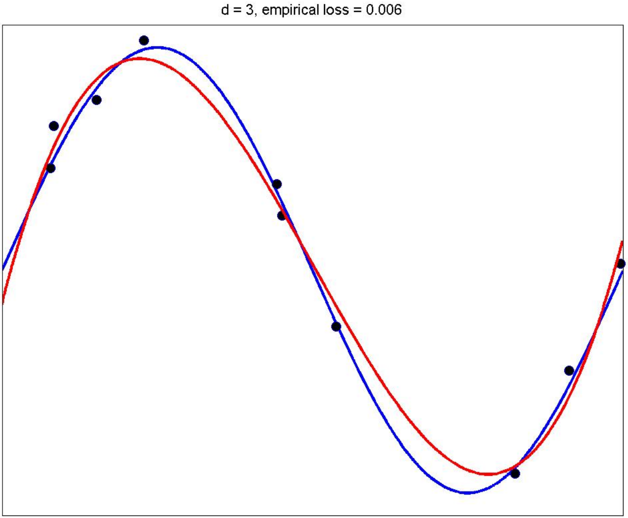


Enlightening example

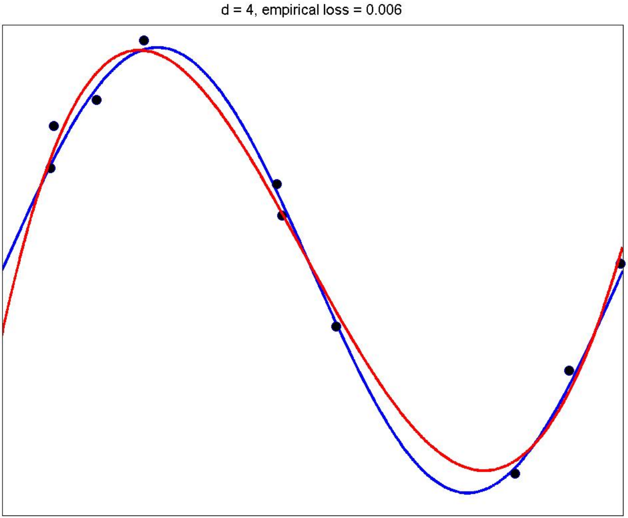


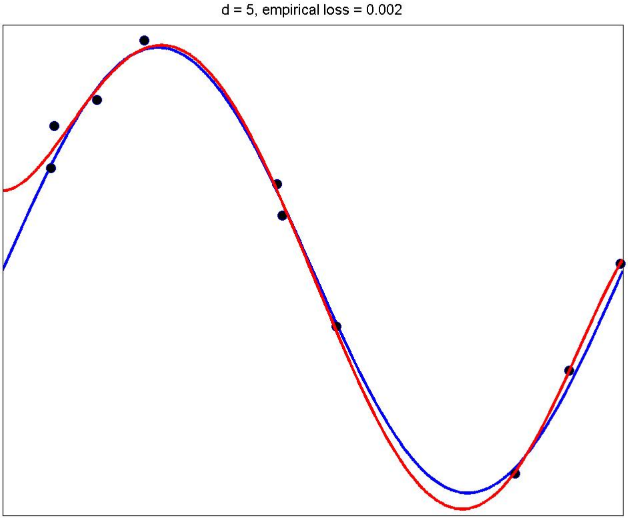

d = 2, empirical loss = 0.1

Enlightening example

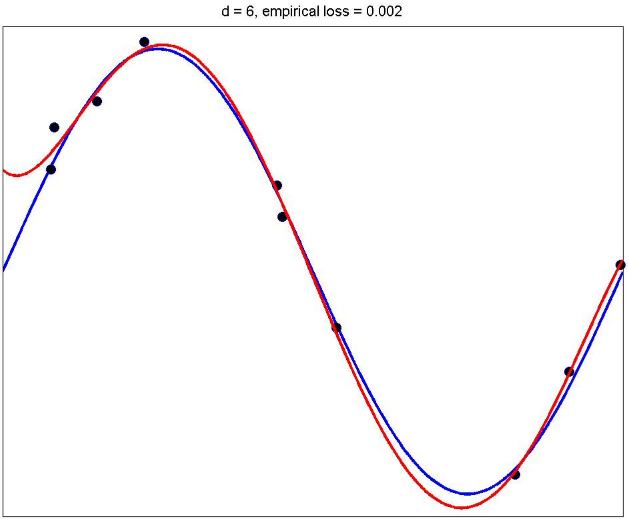


Enlightening example

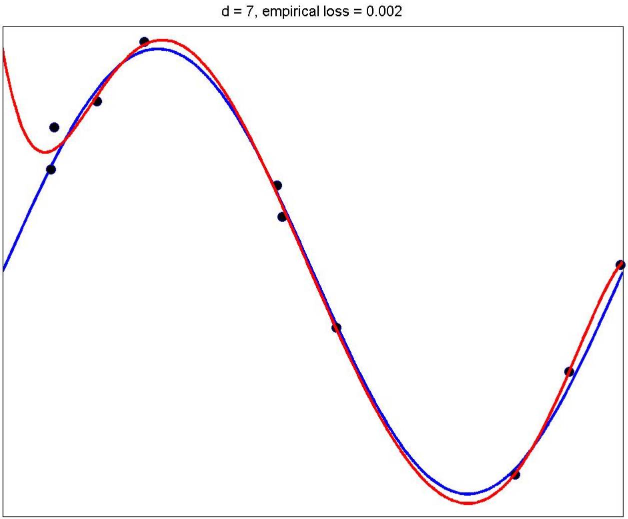



d = 5, empirical loss = 0.002

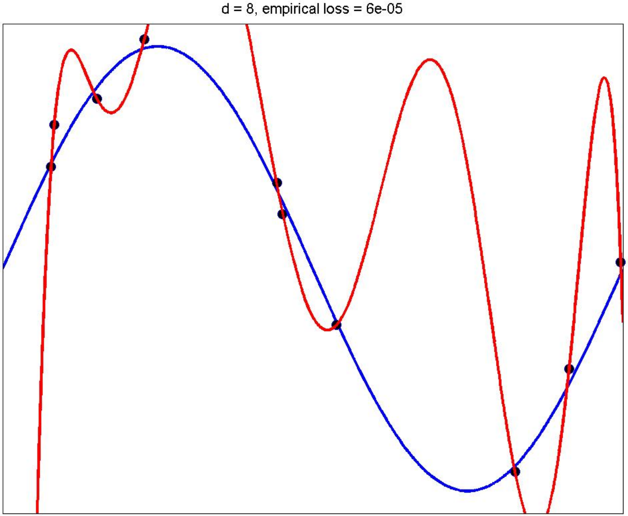
Enlightening example



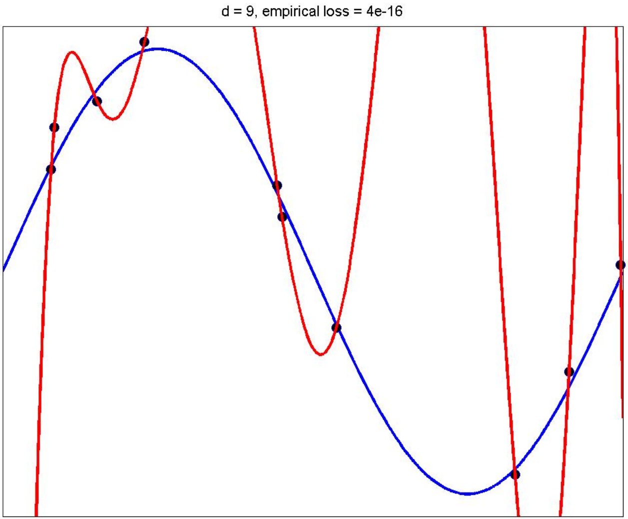
Enlightening example



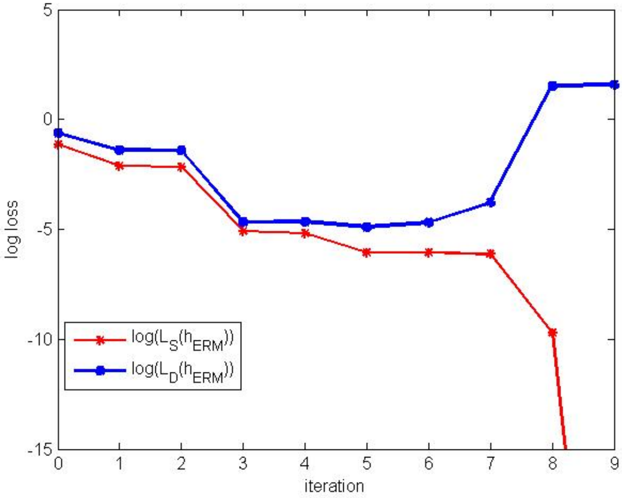
Enlightening example



Enlightening example



Enlightening example



Linear regression: $h_w(x) = \langle w, x \rangle + b$

Linear classifier: $h_w(x) = sign(\langle w, x \rangle + b)$

One can generalize using a transformation $\psi : \mathbb{R}^n \to \mathbb{R}^m$ and $h_w(x) = \langle w, \psi(x) \rangle + b$

The polynomials in the previous example are of that form - $\psi(x) = (x, x^2, ..., x^d)$, $\langle w, \psi(x) \rangle = b + w_1 x + w_2 x^2 + ... + w_d x^d$.

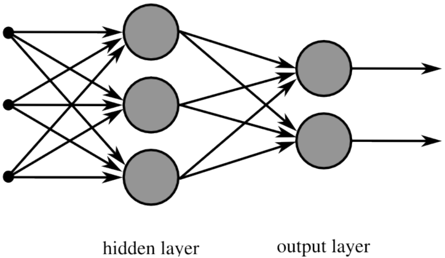
Advantages: Fast to train and to predict, simple "workhorse", tends not to overfit.
Disadvantages: Can be limited, especially in lower dimensions.

Consider a tree (binary most often) where each internal node corresponds to a split of the data, and each leaf corresponds to a prediction.

Advantages: Very flexible, works well with various data types, fast to predict.

Disadvantages: ERM is NP hard, tends to overfit.

Each "neuron" computes a simple function on the sum of its inputs from other neurons, and neurons are connected by some structure.



Advantages: Recently became state of the art in many fields.

Disadvantages: Not as simple and fast as previous methods to train.

If we fix some $h \in \mathcal{H}$, then $\ell(h(x_i), y_i)$ are i.i.d random variables with mean $L_{\mathcal{D}}(h)$.

The law of large numbers shows that
$L_S(h) = \frac{1}{m}\sum_{i=1}^{m} \ell(h(x_i), y_i) \xrightarrow{m\to\infty} L_{\mathcal{D}}(h)$ with probability 1.

This is will not enough for our purposes, we need to say something for a specific finite m. We will prove upper bounds on $P(|\frac{1}{m}\sum x_i - \mu| > \epsilon)$ for i.i.d random variables x_i with mean μ.

Theorem (Markov's inequality)

Let X be a ***nonnegative*** *random variable with expected value $\mathbb{E}[X]$, then $P(X \geq a) \leq \frac{\mathbb{E}[X]}{a}$ for all $a > 0$.*

Proof.

Define $A = \{\omega : X(\omega) \geq a\}$ then $\mathbb{E}[X] = \mathbb{E}[X \cdot \mathbb{1}_A + X \cdot \mathbb{1}_{A^C}]$ when $\mathbb{1}_A$ is the indicator function and A^C is A's complement. Because X is nonnegative this implies that $\mathbb{E}[X] \geq \mathbb{E}[X \cdot \mathbb{1}_A] \geq \mathbb{E}[a \cdot \mathbb{1}_A] = a \cdot P(X \geq a)$ □

Theorem (Chebyshev's inequality)

Let X be a random variable with mean and variance μ and σ^2 respectively then $P(|X-\mu| \geq k\sigma) \leq \frac{1}{k^2}$ for all $k > 0$.

Proof.

$P(|X-\mu| \geq k\sigma) = P\left((X-\mu)^2 \geq k^2\sigma^2\right) \overset{Markov}{\leq} \frac{\mathbb{E}[(X-\mu)^2]}{k^2\sigma^2} = \frac{1}{k^2}$ □

Corollary

$X_1, \ldots, X_m$ i.i.d variables with with mean and variance μ and σ^2 respectively then $P\left(|\frac{1}{m}\sum_{i=1}^{m} X_i - \mu| \geq \epsilon\right) \leq \frac{\sigma^2}{\epsilon^2 m}$.

Chebyshev's inequality is tight, so in order to improve it (in some respect) we need a further assumption - boundedness.

Theorem (Hoeffding inequality)

Let $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ *be the average of bounded independent random variables with* $X_i \in [a_i, b_i]$ *then*

$$P\left(\bar{X} - \mathbb{E}[\bar{X}] \geq \epsilon\right) \leq \exp\left(\frac{-2\epsilon^2 n^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right)$$

$$P\left(\mathbb{E}[\bar{X}] - \bar{X} \geq \epsilon\right) \leq \exp\left(\frac{-2\epsilon^2 n^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right)$$

We will prove a slightly weaker version where $X_i \in [0, 1]$.

Proof (restricted case).

We will prove the first inequality (second is similar). Define $S_n = X_1 + ... + X_n$ then for all $\lambda > 0$
$P(S_n \geq t) = P(\lambda S_n \geq \lambda t) = P\left(e^{\lambda S_n} \geq e^{\lambda t}\right) \overset{Markov}{\leq} e^{-\lambda t}\mathbb{E}[e^{\lambda S_n}] = e^{-\lambda t}\prod_{i=1}^{n}\mathbb{E}[e^{\lambda X_i}]$.

Let us define $\mathbb{E}[X_i] = p_i$ and $q_i = 1 - p_i$. As $e^{\lambda x}$ is convex, $e^{\lambda x} \leq xe^{\lambda} + 1 - x \Rightarrow \mathbb{E}[e^{\lambda x_i}] \leq p_i e^{\lambda} + q_i$.

Combining all we have so far we have that $P(S_n \geq t) \leq e^{-\lambda t}\prod_{i=1}^{n}(p_i e^{\lambda} + q_i)$.

By the arithmetic-geometric means inequality this is bounded by $\left(\frac{\sum(p_i e^{\lambda} + q_i)}{n}\right)^n = (pe^{\lambda} + q)^n$ for $p = \sum p_i/n$ and $q = 1 - p$.

Proof (Cont.)

$P(S_n \geq t) \leq e^{-\lambda t}(pe^{\lambda} + q)^n$ with $p = \sum p_i/n = \mathbb{E}[\bar{X}]$.

Substituting $(p+\epsilon)n$ for t we get
$P(S_n \geq (p+\epsilon)n) \leq e^{-\lambda(p+\epsilon)n}(pe^{\lambda} + q)^n$.

Optimizing λ (and some arithmetic) we get
$P(S_n \geq (p+\epsilon)n) \leq \exp\left(-(p+\epsilon)\ln\left(\frac{p+\epsilon}{p}\right) - (q-\epsilon)\ln\left(\frac{q-\epsilon}{q}\right)\right)^n$

Side note: Inside the exponent is the relative entropy/Kullback Leibler divergance $D_{KL}((p+\epsilon, q-\epsilon)||(p,q))$ between (p,q) distribution and $(p+\epsilon, q-\epsilon)$.

This is stronger then the bound we want to prove, but less convenient and therefore less used.

Proof (finished).

We have $P\left(S_n \geq (p+\epsilon)n\right) \leq \exp\left(-nf(\epsilon)\right)$ for
$f(\epsilon) = (p+\epsilon)\ln\left(\frac{p+\epsilon}{p}\right) + (q-\epsilon)\ln\left(\frac{q-\epsilon}{q}\right)$.

Derivating twice we get $f'(\epsilon) = \ln(\frac{p+\epsilon}{p}) - \ln(\frac{q-\epsilon}{q})$ and
$f''(\epsilon) = \frac{1}{(p+\epsilon)(q-\epsilon)}$.

Now $f(0) = f'(0)$ and $f''(\epsilon) \geq 4$ for all $0 < \epsilon < q$ as
$x(1-x) \leq \frac{1}{4}$ for all $0 < x < 1$.

By the Tylor theorem we have for all $0 \leq \epsilon \leq q$
$f(\epsilon) = f(0) + f'(0)t + f''(\xi)\frac{\epsilon^2}{2!} \geq 2\epsilon^2$. Plugging it in the first equation and we are done (for $\epsilon > q$ the bound is trivial). □