ICTP, Fall 2018
Winter School on Learning and AI

Class 01: Statistical Learning & ERM

Lorenzo Rosasco

Learning from examples

- Machine Learning deals with systems that are trained from data rather than being explicitly programmed.

- These lectures are about building and understanding machine learning algorithms.

Outline

Statistical learning

ERM

Regularization

Nonlinear models

Machine learning as a paradigm

Texts

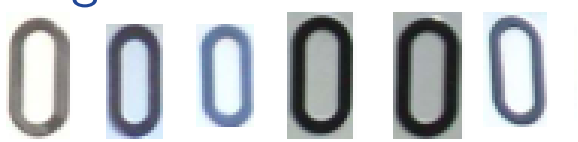
Subject	Date	Time	Body	Spam?
I has the viagra for you	03/12/1992	12:23 pm	Hi! I noticed that you are a software engineer so here's the pleasure you were looking for…	Yes
Important business	05/29/1995	01:24 pm	Give me your account number and you'll be rich. I'm totally serial	Yes
Business Plan	05/23/1996	07:19 pm	As per our conversation, here's the business plan for our new venture Warm regards…	No
Job Opportunity	02/29/1998	08:19 am	Hi !I am trying to fill a position for a PHP …	Yes
[A few thousand rows ommitted]				
Call mom	05/23/2000	02:14 pm	Call mom. She's been trying to reach you for a few days now	No

Machine learning as a paradigm

Texts

Subject	Date	Time	Body	Spam?
I has the viagra for you	03/12/1992	12:23 pm	Hi! I noticed that you are a software engineer so here's the pleasure you were looking for…	Yes
Important business	05/29/1995	01:24 pm	Give me your account number and you'll be rich. I'm totally serial	Yes
Business Plan	05/23/1996	07:19 pm	As per our conversation, here's the business plan for our new venture Warm regards…	No
Job Opportunity	02/29/1998	08:19 am	Hi !I am trying to fill a position for a PHP …	Yes
[A few thousand rows ommitted]				
Call mom	05/23/2000	02:14 pm	Call mom. She's been trying to reach you for a few days now	No

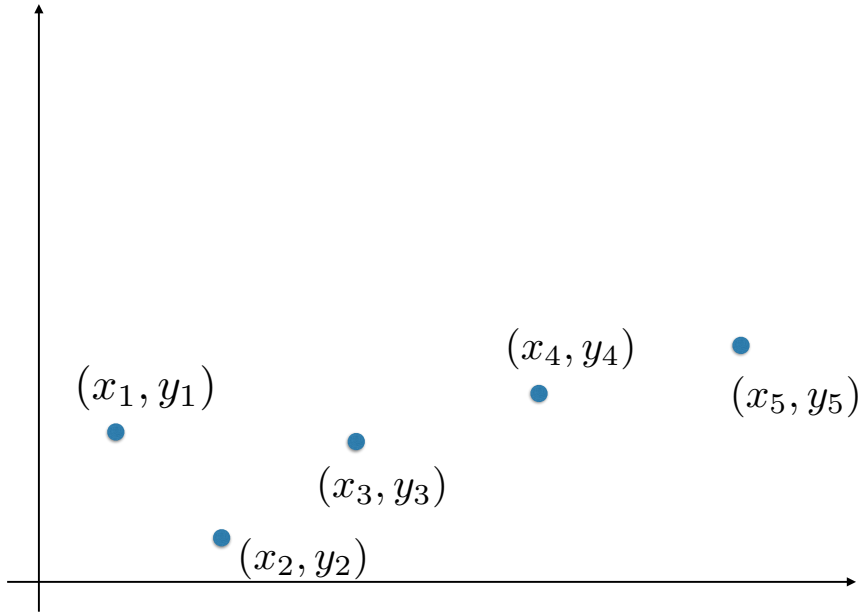
Images



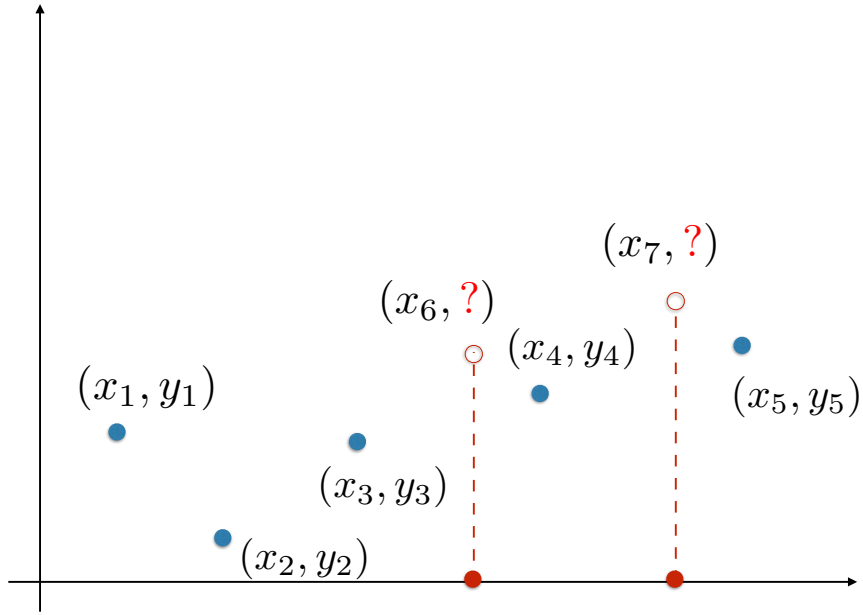
Data: $(x_1, y_1), \ldots, (x_n, y_n)$

Note: x_i's can be *high/huge* dimensional!

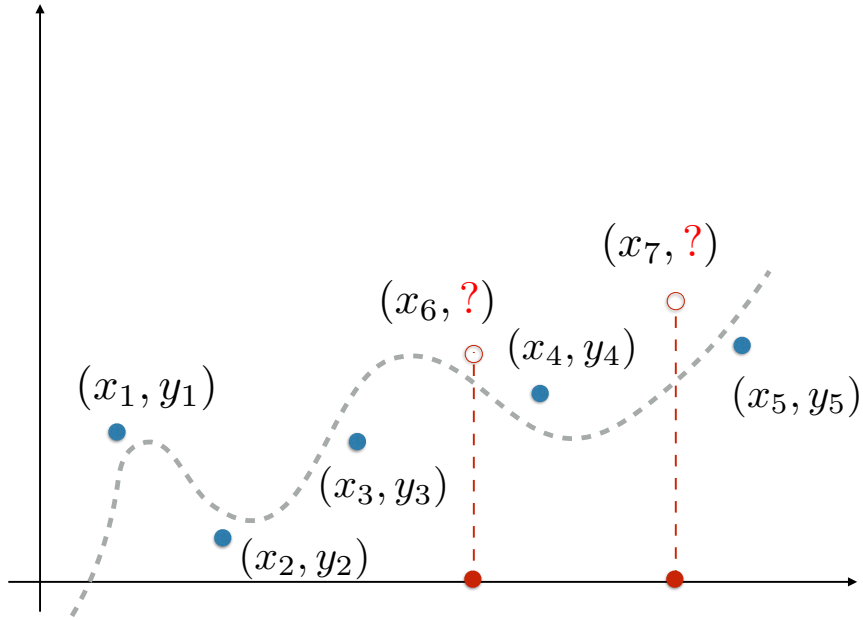
Learning functions



Learning functions



Learning functions



Learning is about inference

Problem: **given** $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ **find** $f(x_{\text{new}}) \sim y_{\text{new}}$

Statistical learning

- $X \times Y$ probability space, with measure P.

Define the **expected risk**:

$$L(f) = \mathbb{E}_{(x,y)\sim P}[(y - f(x))^2]$$

Statistical learning

- $X \times Y$ probability space, with measure P.

Define the **expected risk**:

$$L(f) = \mathbb{E}_{(x,y)\sim P}[(y - f(x))^2]$$

Problem: Solve

$$\min_{f:X\to Y} L(f),$$

given only

$$S_n = (x_1, y_1), \ldots, (x_n, y_n) \sim P^n,$$

sampled i.i.d. with P fixed, but unknown.

Learning algorithms

Learning algorithm

$$S_n \to \widehat{f} = \widehat{f}_{S_n}.$$

How to measure the error of an estimate?

Excess risk:

$$L(\widehat{f}) - \min_{f:X\to Y} L(f).$$

Quality of a solution

Consistency: For any $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}\left(L(\widehat{f}) - \min_{f:X\to Y} L(f) \geq \epsilon \right) = 0.$$

Finite sample bounds: For any $\epsilon > 0, n \in \mathbb{N}$,

$$\mathbb{P}\left(L(\widehat{f}) - \min_{f:X\to Y} L(f) \geq \epsilon \right) \leq \delta_P(n, \epsilon).$$

Learning algorithms design

How can we design a learning algorithm?

Outline

Empirical risk minimization

Replace

$$\min_{f:X\to Y} \mathbb{E}_{(x,y)\sim P}[(y - w^\top x)^2],$$

by

$$\min_{w\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n}(y_i - w^\top x_i)^2.$$

Minimize

- an empirical approximate objective,
- over manageable functions[1].

[1] Linear functions are the conceptual building block of most functions.

Matrices and linear systems

Let $\widehat{X} \in \mathbb{R}^{nd}$ and $\widehat{y} \in \mathbb{R}^{n}$. Then

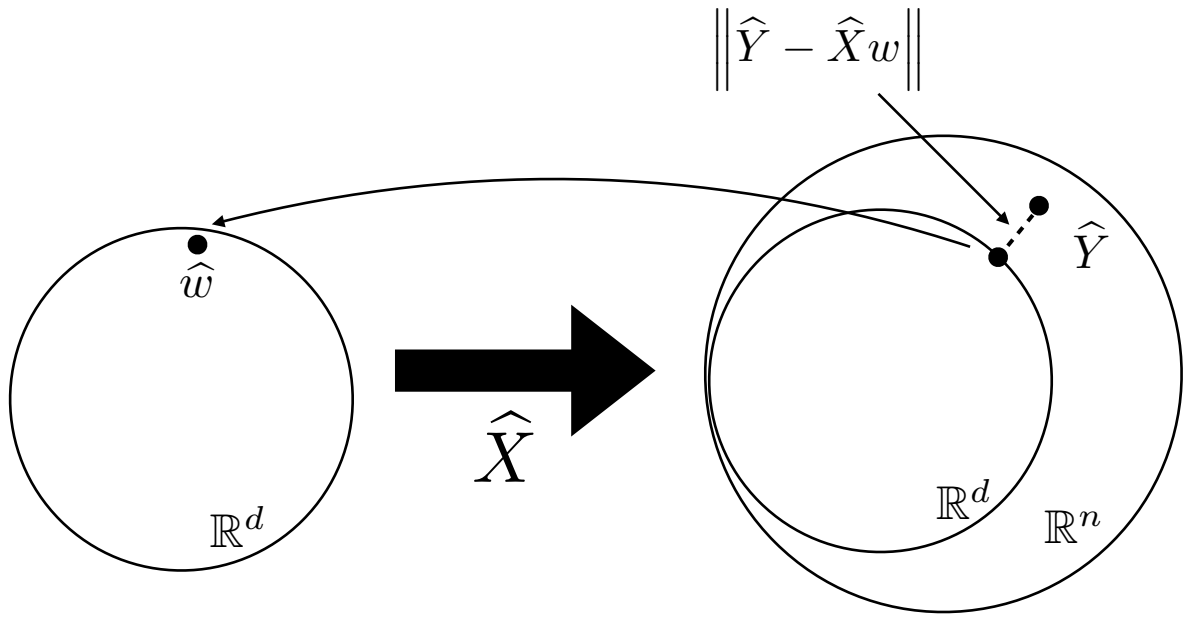
$$\frac{1}{n}\sum_{i=1}^{n}(y_i - w^\top x_i)^2 = \frac{1}{n}\left\|\widehat{y} - \widehat{X}w\right\|^2.$$

This is the least squares problem associated to the linear system

$$\widehat{X}w = \widehat{y}.$$

Overdetermined lin. syst.

$$n > d$$



$$\nexists\ \widehat{w} \quad \text{s.t.} \quad \widehat{X}w = \widehat{y}$$

Least squares solutions

From the optimality conditions

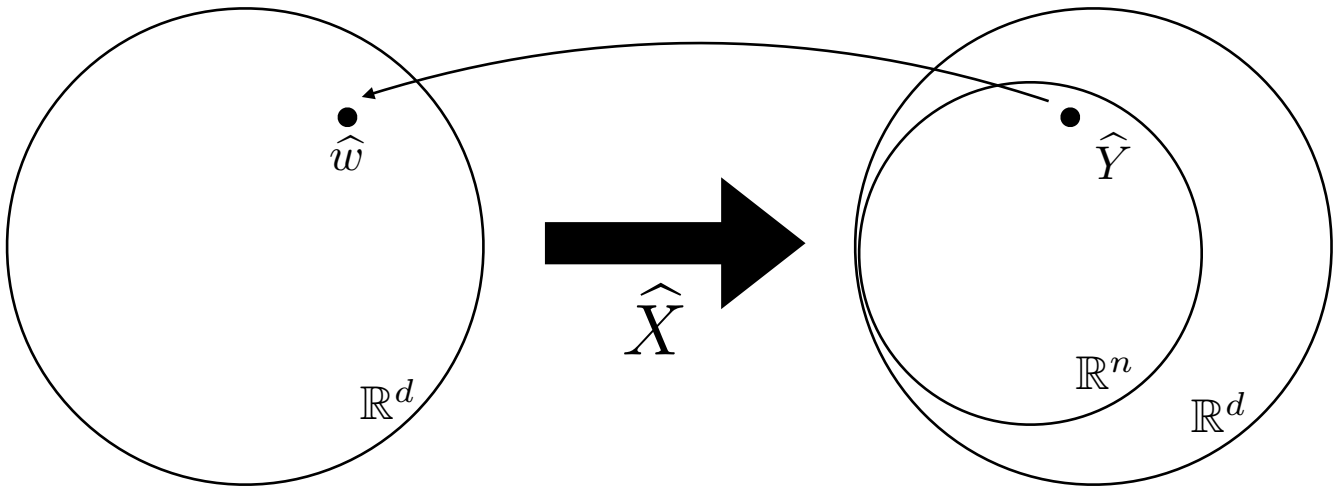
$$\nabla_w \frac{1}{n} \left\| \widehat{y} - \widehat{X} w \right\|^2 = 0$$

we can derive the normal equation

$$\widehat{X}^\top \widehat{X} w = \widehat{X}^\top \widehat{y} \qquad \Leftrightarrow \qquad \widehat{w} = (\widehat{X}^\top \widehat{X})^{-1} \widehat{X}^\top \widehat{y}.$$

Underdetermined lin. syst.

$$n < d$$



$$\exists\ \widehat{w} \quad \text{s.t.} \quad \widehat{X}w = \widehat{y}$$

possibly not unique...

Minimal norm solution

There can be many solutions

$$\widehat{X}\widehat{w} = \widehat{y}, \quad \text{and} \quad \widehat{X}w_0 = 0 \quad \Rightarrow \widehat{X}(\widehat{w} + w_0) = \widehat{y}.$$

Consider

$$\min_{w \in \mathbb{R}^d} \|w\|^2, \qquad \text{subj. to} \qquad \widehat{X}w = \widehat{y}.$$

Using the method of Lagrange multipliers, the solution is

$$\widehat{w} = \widehat{X}^\top (\widehat{X}\widehat{X}^\top)^{-1}\widehat{y}.$$

Pseudoinverse

$$\widehat{w} = \widehat{X}^{\dagger}\widehat{y}$$

For $n > d$, (independent columns)

$$\widehat{X}^{\dagger} = (\widehat{X}^{\top}\widehat{X})^{-1}\widehat{X}^{\top}.$$

For $n < d$, (independent rows)

$$\widehat{X}^{\dagger} = \widehat{X}^{\top}(\widehat{X}\widehat{X}^{\top})^{-1}.$$

Spectral view

Consider the SVD of $\widehat{X}$

$$\widehat{X} = USV^\top \quad \Leftrightarrow \quad \widehat{X}w = \sum_{j=1}^{r} \sigma_j (v_j^\top w) u_j,$$

here $r \leq n \wedge d$ is the rank of $\widehat{X}$.

Then,

$$\widehat{w}^\dagger = \widehat{X}^\dagger \widehat{y} = \sum_{j=1}^{r} \frac{1}{\sigma_j} (u_j^\top \widehat{y}) v_j.$$

Pseudoinverse and bias

$$\widehat{w}^\dagger = \widehat{X}^\dagger \widehat{y} = \sum_{j=1}^{r} \frac{1}{\sigma_j} (u_j^\top \widehat{y}) v_j.$$

$(v_j)_j$ are principal components of $\widehat{X}$: OLS "likes" principal components.

Not all linear functions are the same for OLS!

The pseudoinverse introduces a bias towards certain solutions.

Terminology: regularization or pseudosolutions?

- In signal processing, minimal norm solutions are called regularization.
- In classical regularization theory, they are called pseudosolutions.
- Regularization refers to a family of *stable* solutions converging to pseudosolutions...

Stability and regularization

In practice

$$\min_{w\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n}(y_i - w^\top x_i)^2 \qquad \Leftrightarrow \qquad \widehat{X}^\top \widehat{X} w = \widehat{X}^\top \widehat{y}.$$

In theory

$$\min_{w\in\mathbb{R}^d} \mathbb{E}_{(x,y)\sim P}[(y - w^\top x)^2] \qquad \Leftrightarrow \qquad \mathbb{E}_{x\sim P_X}[xx^\top]w = \mathbb{E}_{(x,y)\sim P}[xy].$$

The solution

$$\widehat{w}^\dagger = \widehat{X}^\dagger \widehat{y} = \sum_{j=1}^{r} \frac{1}{\sigma_j}(u_j^\top \widehat{y}) v_j.$$

might not be robust since we replace

$$\mathbb{E}_{(x,y)\sim P}[xy] \mapsto \widehat{X}^\top \widehat{y} \qquad \text{and} \qquad \mathbb{E}_{x\sim P_X}[xx^\top] \mapsto \widehat{X}^\top \widehat{X}$$

Outline

From OLS to ridge regression

Recall, it also holds,

$$\widehat{X}^{\dagger} = \lim_{\lambda \to 0_+} (\widehat{X}^\top \widehat{X} + \lambda I)^{-1} \widehat{X}^\top = \lim_{\lambda \to 0_+} \widehat{X}^\top (\widehat{X}\widehat{X}^\top + \lambda I)^{-1}.$$

Consider for $\lambda > 0$,

$$\widehat{w}^{\lambda} = (\widehat{X}^\top \widehat{X} + \lambda I)^{-1} \widehat{X}^\top \widehat{y}.$$

This is called ridge regression.

Spectral view on ridge regression

$$\widehat{w}^{\lambda} = (\widehat{X}^{\top}\widehat{X} + \lambda I)^{-1}\widehat{X}^{\top}\widehat{y}$$

Considering the SVD of $\widehat{X}$,

$$\widehat{w}^{\lambda} = \sum_{j=1}^{r} \frac{\sigma_j}{\sigma_j^2 + \lambda}(u_j^{\top}\widehat{y})v_j.$$

Ridge regression as filtering

$$\widehat{w}^{\lambda} = \sum_{j=1}^{r} \frac{\sigma_j}{\sigma_j^2 + \lambda} (u_j^{\top} \widehat{y}) v_j$$

The function

$$F(s) = \frac{\sigma}{\sigma^2 + \lambda},$$

acts as a low pass filter (low frequencies= principal components).

- For s small, $F(\sigma) \approx \sigma/\lambda$.
- For s big, $F(\sigma) \approx 1/\sigma$.

Ridge regression as ERM

$$\widehat{w}^{\lambda} = (\widehat{X}^{\top}\widehat{X} + \lambda I)^{-1}\widehat{X}^{\top}\widehat{y}$$

is the solution of

$$\min_{w \in \mathbb{R}^d} \underbrace{\left\|\widehat{y} - \widehat{X}w\right\|^2 + \lambda \|w\|^2}_{\widehat{L}_{\lambda}(w)}.$$

It follows from,

$$\Delta \widehat{L}_{\lambda}(w) = -\frac{2}{n}\widehat{X}^{\top}(\widehat{y} - \widehat{X}w) + 2\lambda w = 2(\frac{1}{n}\widehat{X}^{\top}\widehat{X} + \lambda I)w - \frac{2}{n}\widehat{X}^{\top}\widehat{y}.$$

Different views on regularization

$$\widehat{w} = \widehat{X}^{\dagger}\widehat{y} \qquad\qquad \widehat{w}_{\lambda} = (\widehat{X}^{\top}\widehat{X} + \lambda I)^{-1}\widehat{X}^{\top}\widehat{y}$$

$$\min_{w \in \mathbb{R}^d \text{ s.t. } \widehat{X}w = \widehat{y}} \|w\|^2 \qquad\qquad \min_{w \in \mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n}(y_i - w^{\top}x_i)^2 + \lambda \|w\|^2$$

- Introduces a *bias* towards certain solutions: small norm/principal components,
- controls the *stability* of the solution .

Complexity of ridge regression

Back to computations.

Solving

$$\widehat{w}^{\lambda} = (\widehat{X}^{\top}\widehat{X} + \lambda I)^{-1}\widehat{X}^{\top}\widehat{y}$$

requires essentially (using a direct solver)

- time $O(nd^2 + d^3)$,
- memory $O(nd \vee d^2)$.

What if $n \ll d$?

Representer theorem in disguise

A simple observation

Using SVD we can see that

$$(\widehat{X}^\top \widehat{X} + \lambda I)^{-1} \widehat{X}^\top = \widehat{X}^\top (\widehat{X}\widehat{X}^\top + \lambda I)^{-1}$$

More on complexity

Then

$$\widehat{w}^{\lambda} = \widehat{X}^{\top}(\widehat{X}\widehat{X}^{\top} + \lambda I)^{-1}\widehat{y}.$$

requires essentially (using a direct solver)

- time $O(n^2 d + n^3)$,
- memory $O(nd \vee n^2)$.

Representer theorem

Note that

$$\widehat{w}^{\lambda} = \widehat{X}^{\top} \underbrace{(\widehat{X}\widehat{X}^{\top} + \lambda I)^{-1}\widehat{y}}_{c \in \mathbb{R}^n} = \sum_{i=1}^{n} x_i c_i.$$

The coefficients vector is a linear combination of the input points.

Then

$$\hat{f}^{\lambda}(x) = x^{\top}\widehat{w}^{\lambda} = x^{\top}\widehat{X}^{\top} c = \sum_{i=1}^{n} x^{\top} x_i c_i$$

The function we obtain is a linear combination of *inner products*.

This will be the key to nonparametric learning.

Outline

Nonlinear models

What about non linear models?

Nonlinear features

So far

$$f(x) = w^\top x, \qquad \widehat{w} = (\widehat{X}^\top \widehat{X} + \lambda I)^{-1} \widehat{X}^\top \widehat{y}.$$

Now consider

$$f(x) = w^\top \Phi(x) = \sum_{i=1}^{p} w^j \varphi_j(x) \qquad \widehat{w}^\lambda = (\widehat{\Phi}^\top \widehat{\Phi} + n\lambda I)^{-1} \widehat{\Phi}^\top \widehat{y}$$

with $\widehat{\Phi} \in \mathbb{R}^{np}$ such that $(\widehat{\Phi})_{ij} = \varphi_j(x_i)$

Examples

$$f(x) = w^\top \Phi(x) = \sum_{i=1}^{p} w^j \varphi_j(x)$$

- Consider $X = \mathbb{R}$ and $\Phi(x) = (x^2, x, 1)$, hence f polynomial.
- Fourier basis,
- wave-lets + their variations,
- vision: SIFT, HOG
- audio: MFCC
- ...

Representer theorem a little less in disguise

Analogously to before

$$\widehat{w}^{\lambda} = \widehat{\Phi}^{\top} c = \sum_{i=1}^{n} \Phi(x_i) c_i \qquad \Leftrightarrow \qquad \hat{f}^{\lambda}(x) = \sum_{i=1}^{n} \Phi(x)^{\top} \Phi(x_i) c_i$$

$$c = (\widehat{\Phi}\widehat{\Phi}^{\top} + \lambda I)^{-1} \widehat{y}, \qquad (\widehat{\Phi}\widehat{\Phi}^{\top})_{ij} = \Phi(x_i)^{\top} \Phi(x_j)$$

$$\Phi(x)^{\top} \Phi(\bar{x}) = \sum_{s=1}^{p} \varphi_s(x) \varphi_s(\bar{x}).$$

We can consider $p = \infty$, as long as the series converges!!!

An observation

For $X = \mathbb{R}$ consider

$$\varphi_j(x) = x^{j-1} e^{-x^2\gamma} \sqrt{\frac{(2\gamma)^{(j-1)}}{(j-1)!}}, \qquad j = 2, \ldots, \infty$$

with $\varphi_1(x) = 1$.

Then

$$\begin{aligned}
\sum_{j=1}^{\infty} \varphi_j(x)\varphi_j(\bar{x}) &= \sum_{j=1}^{\infty} x^{j-1} e^{-x^2\gamma} \sqrt{\frac{(2\gamma)^{j-1}}{(j-1)!}} \bar{x}^{j-1} e^{-\bar{x}^2\gamma} \sqrt{\frac{(2\gamma)^{j-1}}{(j-1)!}} \\
&= e^{-x^2\gamma} e^{-\bar{x}^2\gamma} \sum_{j=1}^{\infty} \frac{(2\gamma)^{j-1}}{(j-1)!} (x\bar{x})^{j-1} = e^{-x^2\gamma} e^{-\bar{x}^2\gamma} e^{2x\bar{x}^2\gamma} \\
&= e^{-|x-\bar{x}|^2\gamma}
\end{aligned}$$

Kernel ridge regression

We have

$$\hat{f}^{\lambda}(x) = \sum_{i=1}^{n} \Phi(x)^{\top} \Phi(x_i) c_i = \sum_{i=1}^{n} k(x, x_i) c_i$$

$$c = (\widehat{K} + \lambda I)^{-1} \widehat{y}, \qquad (\widehat{K})_{ij} = \Phi(x_i)^{\top} \Phi(x_j) = k(x_i, x_j)$$

$\widehat{K}$ is the kernel matrix, the Gram (inner products) matrix of the data.

"The kernel trick"

Kernels

A kernel $k : X \times X \to \mathbb{R}$ is symmetric and positive definite.[2].

Examples

- linear $k(x, \bar{x}) = x^\top \bar{x}$
- polynomial $k(x, \bar{x}) = (x^\top \bar{x} + 1)^s$
- Gaussian $k(x, \bar{x}) = e^{-\|x-\bar{x}\|^2 \gamma}$
- kernels on probability distributions
- kernels on strings, groups, graphs...

It is natural to think of a kernel as a measure of similarity.

[2] i.e. the matrix $\hat{K}$ is positive semidefinite for all choice of points $x_1, \ldots, x_n$, i.e.

$$a^\top \hat{K} a \geq 0, \qquad \forall a \in \mathbb{R}^n.$$

Ideas related to this class

- Linear inverse problems.
- Max margin theory.
- Reproducing kernel Hilbert spaces (RKHS).
- Mercer theorem (Karhunen Loéve expansion).
- Gaussian processes.
- Cameron-Martin spaces.

Summing up

- Statistical learning
- ERM
- Penalized ERM
- Nonlinear functions

- Beyond ERM: Optimization & Implicit regularization.