Mathematical Foundations of ML

Philipp Grohs

universität wien
Faculty of Mathematics

OWA Seminar, Oct. 2018

Short Reading List

1. Felipe Cucker and Ding Yuan Zhou: Learning Theory: An Approximation Theory Viewpoint, 2001
2. Luc Devroye, Laszlo Gyorfi, Gabor Lugosi: A Probabilistic Theory of Pattern Recognition; Springer, 2013.
3. Aurelien Geron: Hands-On Machine Learning with Scikit-Learn and TensorFlow; O'Reilley, 2017
4. Brian Steele and John Chandler and Swarna Reddy: Algorithms for Data Science; Springer, 2017

Syllabus

1. Basic Concepts
2. Mathematical Foundations of General Regression Problems
3. Reproducing Kernel Hilbert Spaces
4. Classification
5. Dimensionality Reduction
6. (Kernel) Support Vector Machine

1. Mathematical Foundations of Machine Learning

1.1 Basic Concepts

Definition of Learning

Definition [Mitchell (1997)]

"A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E"

The Task T

Classification

Compute $f : \mathbb{R}^n \to \{1, \ldots, k\}$ which maps data $x \in \mathbb{R}^n$ to a category in $\{1, \ldots, k\}$. Alternative: Compute $f : \mathbb{R}^n \to \mathbb{R}^k$ which maps data $x \in \mathbb{R}^n$ to a histogram with respect to k categories.

The Task T

Classification

Compute $f : \mathbb{R}^n \to \{1, \ldots, k\}$ which maps data $x \in \mathbb{R}^n$ to a category in $\{1, \ldots, k\}$. Alternative: Compute $f : \mathbb{R}^n \to \mathbb{R}^k$ which maps data $x \in \mathbb{R}^n$ to a histogram with respect to k categories.

The Task T

Classification

Compute $f : \mathbb{R}^n \to \{1, \dots, k\}$ which maps data $x \in \mathbb{R}^n$ to a category in $\{1, \dots, k\}$. Alternative: Compute $f : \mathbb{R}^n \to \mathbb{R}^k$ which maps data $x \in \mathbb{R}^n$ to a histogram with respect to k categories.

$x =$  $\mapsto f(x) = 5.$

The Task T

Regression

Predict a numerical value $f : \mathbb{R}^n \to \mathbb{R}$.

The Task T

Regression

Predict a numerical value $f : \mathbb{R}^n \to \mathbb{R}$.

- Expected claim of insured person

The Task T

Regression

Predict a numerical value $f : \mathbb{R}^n \to \mathbb{R}$.

- Expected claim of insured person
- Algorithmic trading

The Task T

Density Estimation

Estimate a probability density $p : \mathbb{R}^n \to \mathbb{R}_+$ which can be interpreted as a probability distribution on the space that the examples were drawn from.

The Task T

Density Estimation

Estimate a probability density $p : \mathbb{R}^n \to \mathbb{R}_+$ which can be interpreted as a probability distribution on the space that the examples were drawn from.

- Useful for many tasks in data processing, for example if we observe corrupted data $\tilde{x}$ we may estimate the original x as the argmax of $p(\tilde{x}|x)$.

The Experience E

The experience typically consists of a dataset which consists of many examples (aka data points).

The Experience E

The experience typically consists of a dataset which consists of many examples (aka data points).

- If these data points are labeled (for example in the classification problem, if we know the classifier of our given data points) we speak of *supervised learning*.

The Experience E

The experience typically consists of a dataset which consists of many examples (aka data points).

- If these data points are labeled (for example in the classification problem, if we know the classifier of our given data points) we speak of *supervised learning*.
- If these data points are not labeled (for example in the classification problem, the algorithm would have to find the clusters itself from the given dataset) we speak of *unsupervised learning*.

The Performance Measure P

In classification problems this is typically the *accuracy*, i.e., the proportion of examples for which the model produces the correct output.

The Performance Measure P

In classification problems this is typically the *accuracy*, i.e., the proportion of examples for which the model produces the correct output.

- Often the given dataset is split into a *training set* on which the algorithm operates and a *test set* on which its performance is measured.

An Example: Linear Regression

An Example: Linear Regression

The Task

Regression: Predict $\widehat{f} : \mathbb{R}^d \to \mathbb{R}$.

An Example: Linear Regression

The Task

Regression: Predict $\widehat{f} : \mathbb{R}^d \to \mathbb{R}$.

The Experience

Training data $((x_i^{train}, y_i^{train}))_{i=1}^m$ with $y_i^{train} \sim \widehat{f}(x_i^{train})$

An Example: Linear Regression

The Task

Regression: Predict $\widehat{f} : \mathbb{R}^d \to \mathbb{R}$.

The Experience

Training data $((x_i^{train}, y_i^{train}))_{i=1}^m$ with $y_i^{train} \sim \widehat{f}(x_i^{train})$

The Performance Measure

Given test data $((x_i^{test}, y_i^{test}))_{i=1}^n$ we evaluate the performance of an estimator $f : \mathbb{R}^d \to \mathbb{R}$ as the *mean squared error*

$$\frac{1}{n} \sum_{i=1}^{n} |f(x_i^{test}) - y_i^{test}|^2.$$

An Example: Linear Regression

The Computer Program

Define a *Hypothesis Space*

$$\mathcal{H} = \mathsf{span}\{\varphi_1, \ldots, \varphi_l\} \subset C(\mathbb{R}^d)$$

An Example: Linear Regression

The Computer Program

Define a *Hypothesis Space*

$$\mathcal{H} = \mathsf{span}\{\varphi_1, \ldots, \varphi_l\} \subset C(\mathbb{R}^d)$$

and, given training data

$$\mathbf{z} = (x_i, y_i)_{i=1}^m,$$

An Example: Linear Regression

The Computer Program

Define a *Hypothesis Space*

$$\mathcal{H} = \mathsf{span}\{\varphi_1, \ldots, \varphi_l\} \subset C(\mathbb{R}^d)$$

and, given training data

$$\mathbf{z} = (x_i, y_i)_{i=1}^m,$$

define the *empirical risk*

$$\mathcal{E}_{\mathbf{z}}(f) := \frac{1}{m} \sum_{i=1}^m (f(x_i) - y_i)^2.$$

An Example: Linear Regression

The Computer Program

Define a *Hypothesis Space*

$$\mathcal{H} = \mathsf{span}\{\varphi_1, \ldots, \varphi_l\} \subset C(\mathbb{R}^d)$$

and, given training data

$$\mathbf{z} = (x_i, y_i)_{i=1}^m,$$

define the *empirical risk*

$$\mathcal{E}_{\mathbf{z}}(f) := \frac{1}{m}\sum_{i=1}^m (f(x_i) - y_i)^2.$$

We let our algorithm find the minimizer (a.k.a. *empirical regression function*)

$$\widehat{f}_{\mathcal{H},\mathbf{z}} := \operatorname{argmin}_{f\in\mathcal{H}} \mathcal{E}_{\mathbf{z}}(f).$$

An Example: Linear Regression

Computing the Empirical Target Function

An Example: Linear Regression

Computing the Empirical Target Function

- Let

$$\mathbf{A} = (\varphi_j(x_i))_{i,j} \in \mathbb{R}^{m \times l}.$$

An Example: Linear Regression

Computing the Empirical Target Function

- Let
$$\mathbf{A} = (\varphi_j(x_i))_{i,j} \in \mathbb{R}^{m \times l}.$$

- Every $f \in \mathcal{H}$ can be written as $\sum_{i=1}^{m} w_i \varphi_i$ and we denote $\mathbf{w} := (w_i)_{i=1}^{l}$.

An Example: Linear Regression

Computing the Empirical Target Function

- Let
$$\mathbf{A} = (\varphi_j(x_i))_{i,j} \in \mathbb{R}^{m \times l}.$$

- Every $f \in \mathcal{H}$ can be written as $\sum_{i=1}^{m} w_i \varphi_i$ and we denote $\mathbf{w} := (w_i)_{i=1}^{l}$.
- We let $\mathbf{y} := (y_i)_{i=1}^{m}$.

An Example: Linear Regression

Computing the Empirical Target Function

- Let
$$\mathbf{A} = (\varphi_j(x_i))_{i,j} \in \mathbb{R}^{m \times l}.$$

- Every $f \in \mathcal{H}$ can be written as $\sum_{i=1}^{m} w_i \varphi_i$ and we denote $\mathbf{w} := (w_i)_{i=1}^{l}$.
- We let $\mathbf{y} := (y_i)_{i=1}^{m}$.
- We get that
$$\mathcal{E}_{\mathbf{z}}(f) = \|\mathbf{A}\mathbf{w} - \mathbf{y}\|^2 .$$

An Example: Linear Regression

Computing the Empirical Target Function

- Let
$$\mathbf{A} = (\varphi_j(x_i))_{i,j} \in \mathbb{R}^{m \times l}.$$
- Every $f \in \mathcal{H}$ can be written as $\sum_{i=1}^{m} w_i \varphi_i$ and we denote $\mathbf{w} := (w_i)_{i=1}^{l}$.
- We let $\mathbf{y} := (y_i)_{i=1}^{m}$.
- We get that
$$\mathcal{E}_{\mathbf{z}}(f) = \|\mathbf{A}\mathbf{w} - \mathbf{y}\|^2 .$$
- A minimizer is given by $\mathbf{w}_* := \mathbf{A}^\dagger \mathbf{y}$, and we get our estimate
$$f_* := \sum_{i=1}^{l} (\mathbf{w}_*)_i \varphi_i.$$

Proof.

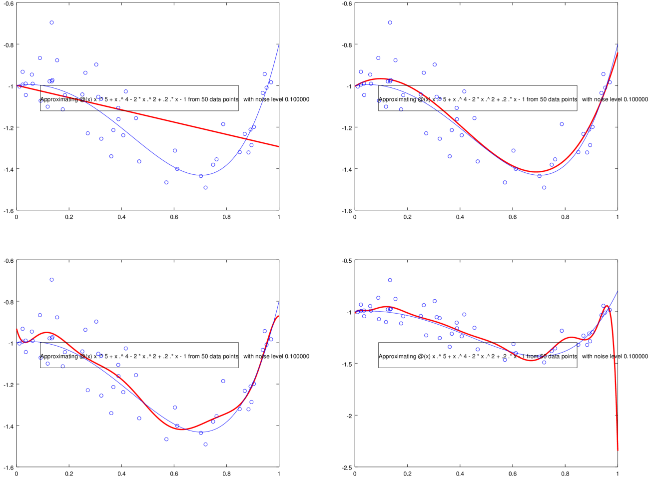
We want to minimize the function

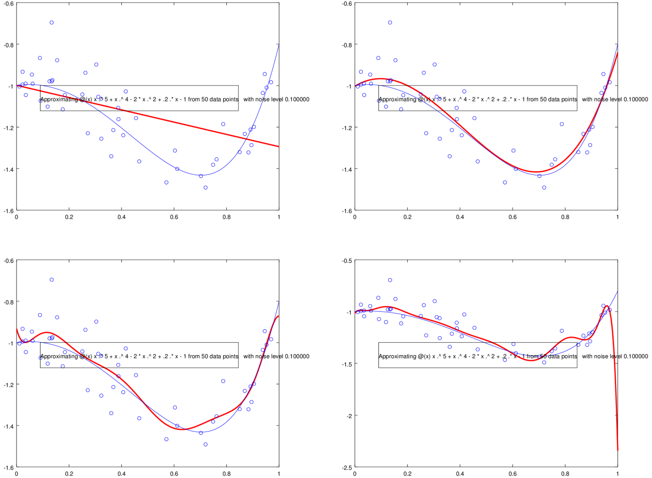
$$\mathcal{X}(\mathbf{w}) := \mathbf{w} \mapsto \|\mathbf{A}\mathbf{w} - \mathbf{y}\|^2,$$

which is (more or less...) equivalent to setting its first derivative to zero. It holds that

$$\frac{d\mathcal{X}(\mathbf{w})}{d\mathbf{w}} = 2\mathbf{A}^\dagger(\mathbf{A}\mathbf{w} - \mathbf{y}),$$

which, if set to zero, are precisely the normal equations. $\square$


Approximating @(x) x .^ 5 + x .^ 4 - 2 * x .^ 2 + .2 .* x - 1 from 50 data points with noise level 0.100000
-0.6
-0.8
-1
-1.2
-1.4
-1.6
-0.5
-1.5
-2
-2.5
0
0.2
0.4
0.6
0.8
1



Degree too low: underfitting. Degree to high: overfitting!

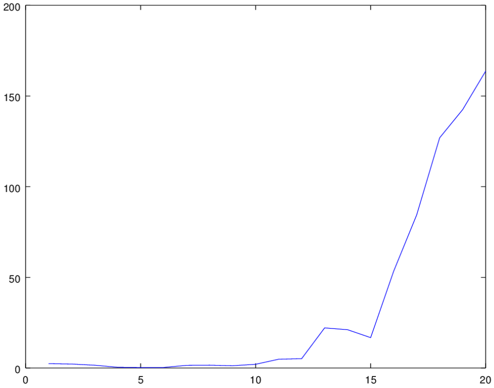


Figure: Error with Polynomial Degree

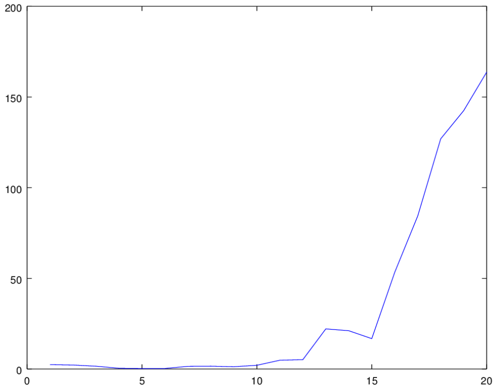


Figure: Error with Polynomial Degree

Bias-Variance Problem

"Capacity" of the hypothesis space has to be adapted to the complexity of the target function and the sample size!

1.2 Mathematical Foundations of General Regression Problems

1.2.1 Basic Definitions

The Mathematical Learning Problem

The Mathematical Learning Problem

Let $(\Sigma, \mathcal{G}, \mathbb{P})$ probability space. Given (Borel measurable) random vectors $X : \Sigma \to \mathbb{R}^d$, $Y : \Sigma \to \mathbb{R}^k$ with $\mathrm{im}(X) \subseteq \Omega$ for $\Omega \subset \mathbb{R}^d$ compact.

The Mathematical Learning Problem

Let $(\Sigma, \mathcal{G}, \mathbb{P})$ probability space. Given (Borel measurable) random vectors $X : \Sigma \to \mathbb{R}^d$, $Y : \Sigma \to \mathbb{R}^k$ with $\mathrm{im}(X) \subseteq \Omega$ for $\Omega \subset \mathbb{R}^d$ compact.

For any (Borel measurable) $f : \Omega \to \mathbb{R}^k$ define the *least squares error*

$$\mathcal{E}(f) := \mathbb{E}[(f(X) - Y)^2].$$

The Mathematical Learning Problem

Let $(\Sigma, \mathcal{G}, \mathbb{P})$ probability space. Given (Borel measurable) random vectors $X : \Sigma \to \mathbb{R}^d$, $Y : \Sigma \to \mathbb{R}^k$ with $\mathrm{im}(X) \subseteq \Omega$ for $\Omega \subset \mathbb{R}^d$ compact.

For any (Borel measurable) $f : \Omega \to \mathbb{R}^k$ define the *least squares error*

$$\mathcal{E}(f) := \mathbb{E}[(f(X) - Y)^2].$$

The learning problem asks for the function $\widehat{f}$ which minimizes $\mathcal{E}$.

The Mathematical Learning Problem

Let $(\Sigma, \mathcal{G}, \mathbb{P})$ probability space. Given (Borel measurable) random vectors $X : \Sigma \to \mathbb{R}^d$, $Y : \Sigma \to \mathbb{R}^k$ with $\mathrm{im}(X) \subseteq \Omega$ for $\Omega \subset \mathbb{R}^d$ compact.
For any (Borel measurable) $f : \Omega \to \mathbb{R}^k$ define the *least squares error*

$$\mathcal{E}(f) := \mathbb{E}[(f(X) - Y)^2].$$

The learning problem asks for the function $\widehat{f}$ which minimizes $\mathcal{E}$.

Example 1: Regression

Example 1: Regression

- Suppose our data is generated from noisy observations

$$y = f(x) + \xi,$$

where ξ is a r.v. with $\mathbb{E}(\xi) = 0$.

Example 1: Regression

- Suppose our data is generated from noisy observations

$$y = f(x) + \xi,$$

  where ξ is a r.v. with $\mathbb{E}(\xi) = 0$.
- Let $X : \Omega \to \mathbb{R}$ be a r.v. (independent of ξ) and let $Y := f(X) + \xi$.

Example 1: Regression

- Suppose our data is generated from noisy observations

$$y = f(x) + \xi,$$

where ξ is a r.v. with $\mathbb{E}(\xi) = 0$.

- Let $X : \Omega \to \mathbb{R}$ be a r.v. (independent of ξ) and let $Y := f(X) + \xi$.
- We have that

$$\begin{aligned}
\mathcal{E}(g) &= \mathbb{E}[(g(X) - Y)^2] = \mathbb{E}[(g(X) - f(X) - \xi)^2] \\
&= \mathbb{E}[f(X) - g(X)^2] + 2\mathbb{E}[(g(X) - f(X))\xi] + \mathbb{E}\xi^2 \\
&= \mathbb{E}[f(X) - g(X)^2] + \mathbb{E}\xi^2 \\
&= \|f - g\|^2_{L^2(\mathbb{R}^d, d\mathbb{P}_X)} + \mathbb{V}[\xi].
\end{aligned}$$

Example 1: Regression

- Suppose our data is generated from noisy observations

$$y = f(x) + \xi,$$

where ξ is a r.v. with $\mathbb{E}(\xi) = 0$.

- Let $X : \Omega \to \mathbb{R}$ be a r.v. (independent of ξ) and let $Y := f(X) + \xi$.
- We have that

$$\begin{aligned}
\mathcal{E}(g) &= \mathbb{E}[(g(X) - Y)^2] = \mathbb{E}[(g(X) - f(X) - \xi)^2] \\
&= \mathbb{E}[f(X) - g(X)^2] + 2\mathbb{E}[(g(X) - f(X))\xi] + \mathbb{E}\xi^2 \\
&= \mathbb{E}[f(X) - g(X)^2] + \mathbb{E}\xi^2 \\
&= \|f - g\|^2_{L^2(\mathbb{R}^d, d\mathbb{P}_X)} + \mathbb{V}[\xi].
\end{aligned}$$

- The learning problem finds f!

Example 2: Classifications

Example 2: Classifications

- Suppose that there is a function f which maps a matrix $x \in [0,1]^{256\times 256}$ to a histogram $f(x) \in \mathbb{R}_+^{10}$. We consider the vector $f(x)/\sum_{i=1}^{10} f(x)_i$ as a histogram describing which digit the image x represents.

Example 2: Classifications

- Suppose that there is a function f which maps a matrix $x \in [0,1]^{256\times 256}$ to a histogram $f(x) \in \mathbb{R}_+^{10}$. We consider the vector $f(x)/\sum_{i=1}^{10} f(x)_i$ as a histogram describing which digit the image x represents.
- Let (X, Y) be random vectors on $\mathbb{R}^{256\times 256} \times \mathbb{R}_+^{10}$ which generate the measurement data we get to see ((X, Y) will not be known to us!!!)

Example 2: Classifications

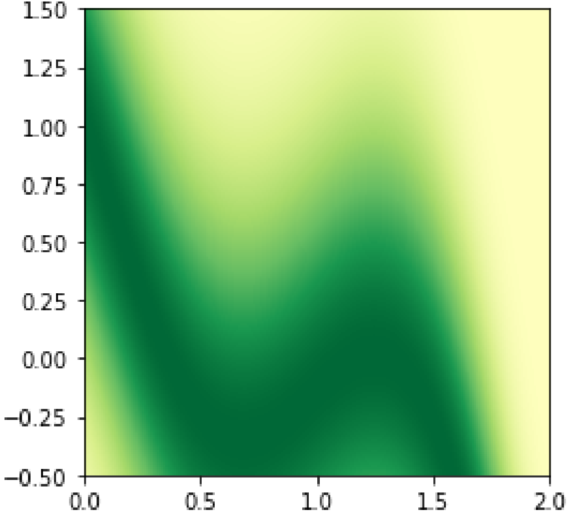
- Suppose that there is a function f which maps a matrix $x \in [0,1]^{256\times 256}$ to a histogram $f(x) \in \mathbb{R}_+^{10}$. We consider the vector $f(x)/\sum_{i=1}^{10} f(x)_i$ as a histogram describing which digit the image x represents.
- Let (X,Y) be random vectors on $\mathbb{R}^{256\times 256} \times \mathbb{R}_+^{10}$ which generate the measurement data we get to see ((X,Y) will not be known to us!!!)
- Now, a function f as above will in general not exist for our problem. But we can look for the function $\widehat{f}$ which minimizes the least squares error $\mathcal{E}$ – this will be the optimal explanation of the measurements in terms of a functional relation between X and Y!

A New Look

Suppose that our training data consists of samples according to a given data distribution (X, Y)

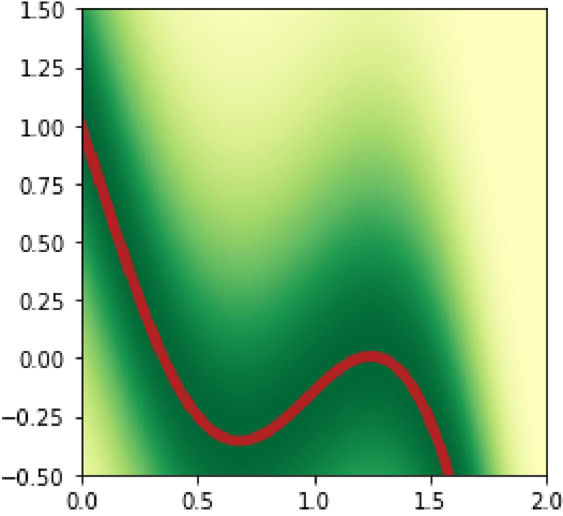
A New Look

Suppose that our training data consists of samples according to a given data distribution (X, Y)



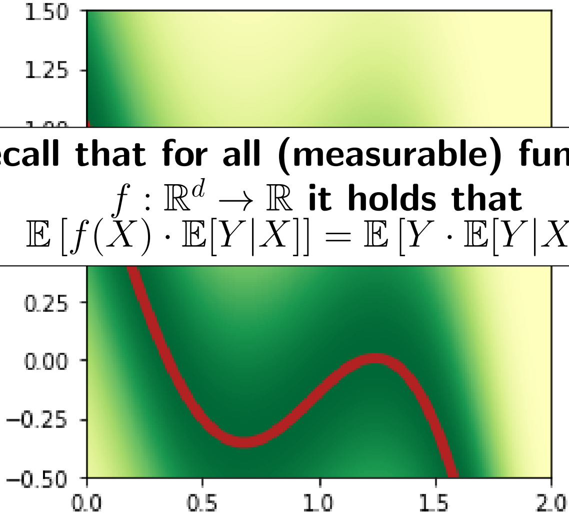
A New Look

If we knew the data distribution (X, Y), the best functional relation between X and Y would simply be $\mathbb{E}[Y|X = x]$!



A New Look

If we knew the data distribution (X, Y), the best functional relation between X and Y would simply be $\mathbb{E}[Y|X = x]$!



Recall that for all (measurable) functions $f : \mathbb{R}^d \to \mathbb{R}$ it holds that $\mathbb{E}\left[f(X) \cdot \mathbb{E}[Y|X]\right] = \mathbb{E}\left[Y \cdot \mathbb{E}[Y|X]\right]$.

Regression Function

Theorem (Main Regression Theorem)

Let $\widehat{f} := \mathbb{E}[Y|X]$ be the regression function and $\sigma^2 := \mathcal{E}(\widehat{f})$. It holds that

$$\mathcal{E}(f) = \|f - \widehat{f}\|^2_{L^2(\mathbb{R}^d, d\mathbb{P}_X)} + \sigma^2$$

Regression Function

Theorem (Main Regression Theorem)

Let $\widehat{f} := \mathbb{E}[Y|X]$ be the regression function and $\sigma^2 := \mathcal{E}(\widehat{f})$. It holds that

$$\mathcal{E}(f) = \|f - \widehat{f}\|^2_{L^2(\mathbb{R}^d, d\mathbb{P}_X)} + \sigma^2$$

Proof.

$$\mathcal{E}(f) = \mathbb{E}[(f(X) - Y)^2] = \mathbb{E}[(\widehat{f}(X) - Y)^2] + 2\underbrace{\mathbb{E}[(f(X) - \widehat{f}(X)) \cdot (\widehat{f}(X) - Y)]}_{=0} + \mathbb{E}[(f(X) - \widehat{f}(X)^2].$$

□

Regression Function

Theorem (Main Regression Theorem)

Let $\widehat{f} := \mathbb{E}[Y|X]$ be the regression function and $\sigma^2 := \mathcal{E}(\widehat{f})$. It holds that

$$\mathcal{E}(f) = \|f - \widehat{f}\|^2_{L^2(\mathbb{R}^d, d\mathbb{P}_X)} + \sigma^2$$

Proof.

$$\mathcal{E}(f) = \mathbb{E}[(f(X) - Y)^2] = \mathbb{E}[(\widehat{f}(X) - Y)^2] + \\ 2\underbrace{\mathbb{E}[(f(X) - \widehat{f}(X)) \cdot (\widehat{f}(X) - Y)]}_{=0} + \mathbb{E}[(f(X) - \widehat{f}(X)^2].$$

□

Corollary

The regression function solves the learning problem!

Regression Function

Theorem (Main Regression Theorem)

Let $\widehat{f} := \mathbb{E}[Y|X]$ be the regression function and $\sigma^2 := \mathcal{E}(\widehat{f})$. It holds that

$$\mathcal{E}(f) = \|f - \widehat{f}\|^2_{L^2(\mathbb{R}^d, d\mathbb{P}_X)} + \sigma^2$$

Proof.

Are we done?

$$\mathcal{E}(f) = \mathbb{E}[(f(X) - Y)^2] = \mathbb{E}[(\widehat{f}(X) - Y)^2] + 2\underbrace{\mathbb{E}[(f(X) - \widehat{f}(X)) \cdot (\widehat{f}(X) - Y)]}_{=0} + \mathbb{E}[(f(X) - \widehat{f}(X)^2].$$

□

Corollary

The regression function solves the learning problem!

Regression Function

Theorem (Main Regression Theorem)

Let $\widehat{f} := \mathbb{E}[Y|X]$ be the regression function and $\sigma^2 := \mathcal{E}(\widehat{f})$. It holds that

$$\mathcal{E}(f) = \|f - \widehat{f}\|^2_{L^2(\mathbb{R}^d, d\mathbb{P}_X)} + \sigma^2$$

Proof.

Are we done?

We don't know (X, Y)!!!

$$\mathcal{E}(f) = \mathbb{E}[($$

$$2\underbrace{\mathbb{E}[(f(X) - \widehat{f}(X)) \cdot (\widehat{f}(X) - Y)]}_{=0} + \mathbb{E}[(f(X) - \widehat{f}(X)^2].$$

□

Corollary

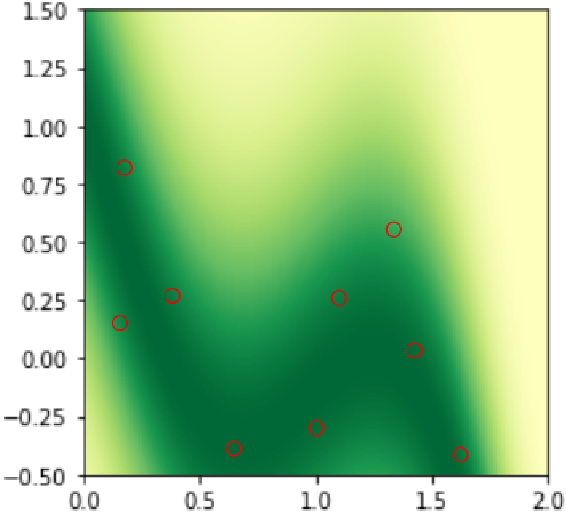
The regression function solves the learning problem!

The Actual Problem

We only have samples.

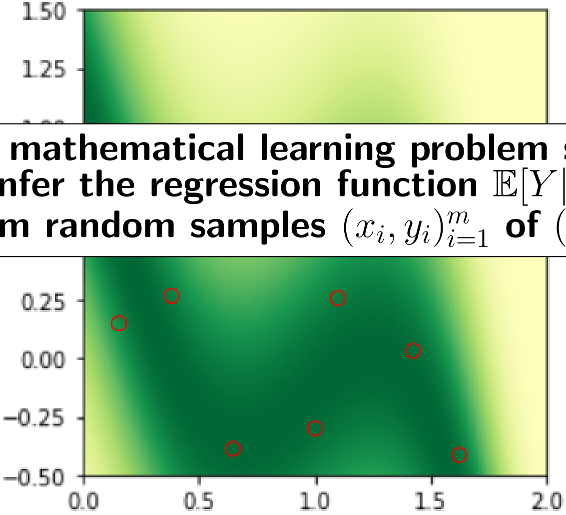
The Actual Problem

We only have samples.



The Actual Problem

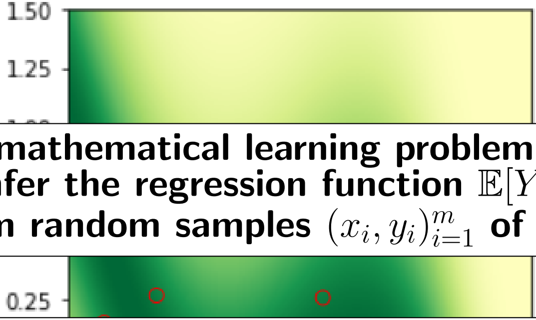
We only have samples.



A mathematical learning problem seeks to infer the regression function $\mathbb{E}[Y|X = x]$ from random samples $(x_i, y_i)_{i=1}^m$ of (X, Y).

The Actual Problem

We only have samples.



A mathematical learning problem seeks to infer the regression function $\mathbb{E}[Y|X=x]$ from random samples $(x_i, y_i)_{i=1}^m$ of (X, Y).

More generally we would like to minimize

$$\mathbb{E}[\mathcal{L}(f(X), Y)]$$

with general loss function.

$\mathcal{L}(y, y') = (y - y')^2 \rightsquigarrow$ **quadratic loss**

$\mathcal{L}(y, y') = y\log(y') + (1-y)\log(1-y') \rightsquigarrow$ **cross-entropy loss.**

1.2.2 Empirical Minimization and Hypothesis Space

Sampling

Empirical Error

Given $\mathbf{z} = ((X^{(1)}, Y^{(1)}), \ldots, (X^{(m)}, Y^{(m)}))$ be i.i.d. with $(X^{(1)}, Y^{(1)}) \sim (X, Y)$. Define the *empirical error*

$$\mathcal{E}_{\mathbf{z}}(f) := \frac{1}{m} \sum_{i=1}^{m} (f(X^{(i)}) - Y^{(i)})^2.$$

Sampling

Empirical Error

Given $\mathbf{z} = ((X^{(1)}, Y^{(1)}), \dots, (X^{(m)}, Y^{(m)}))$ be i.i.d. with $(X^{(1)}, Y^{(1)}) \sim (X, Y)$. Define the *empirical error*

$$\mathcal{E}_{\mathbf{z}}(f) := \frac{1}{m} \sum_{i=1}^{m} (f(X^{(i)}) - Y^{(i)})^2.$$

Given $\mathbf{z}$ the empirical error can actually be computed!

Sampling

Empirical Error

Given $\mathbf{z} = ((X^{(1)}, Y^{(1)}), \ldots, (X^{(m)}, Y^{(m)}))$ be i.i.d. with $(X^{(1)}, Y^{(1)}) \sim (X, Y)$. Define the *empirical error*

$$\mathcal{E}_{\mathbf{z}}(f) := \frac{1}{m} \sum_{i=1}^{m} (f(X^{(i)}) - Y^{(i)})^2.$$

Given $\mathbf{z}$ the empirical error can actually be computed!

Defect

The defect of f is defined as

$$L_{\mathbf{z}}(f) := \mathcal{E}(f) - \mathcal{E}_{\mathbf{z}}(f).$$

Sampling

Empirical Error

Given $\mathbf{z} = ((X^{(1)}, Y^{(1)}), \ldots, (X^{(m)}, Y^{(m)}))$ be i.i.d. with $(X^{(1)}, Y^{(1)}) \sim (X, Y)$. Define the *empirical error*

$$\mathcal{E}_{\mathbf{z}}(f) := \frac{1}{m} \sum_{i=1}^{m} (f(X^{(i)}) - Y^{(i)})^2.$$

Given $\mathbf{z}$ the empirical error can actually be computed!

Defect

The defect of f is defined as

$$L_{\mathbf{z}}(f) := \mathcal{E}(f) - \mathcal{E}_{\mathbf{z}}(f).$$

Can we control the defect? If yes, we actually have some hope of approximating the regression function.

Data Generating Distribution

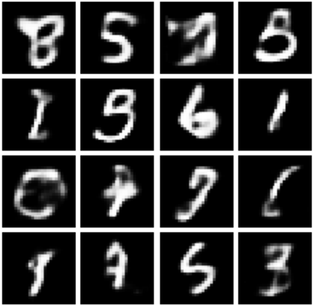
We suppose that there exists a probability distribution on $\mathbb{R}^{784}$ that randomly generates handwritten digits.

Data Generating Distribution

We suppose that there exists a probability distribution on $\mathbb{R}^{784}$ that randomly generates handwritten digits.

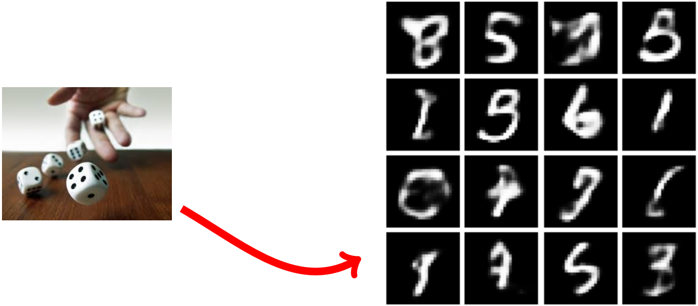
Data Generating Distribution

We suppose that there exists a probability distribution on $\mathbb{R}^{784}$ that randomly generates handwritten digits.



Data Generating Distribution

We suppose that there exists a probability distribution on $\mathbb{R}^{784}$ that randomly generates handwritten digits.



$\rightsquigarrow$ **Variational Autoencoder Demo**

Concentration Inequalities

Bernstein Inequality

Suppose that $(\xi^{(i)})_{i=1}^m$ i.i.d. with $\xi^{(1)} \sim \xi$ with mean $\mathbb{E}(\xi) = \mu$ and $\mathbb{V}(\xi) = \sigma^2$. Suppose that $|\xi - \mu| \leq M$ with probability 1. Then

$$\mathbb{P}\left\{\left|\frac{1}{m}\sum_{i=1}^m \xi^{(i)} - \mu\right| \geq \varepsilon\right\} \leq 2e^{-\frac{m\varepsilon^2}{2\left(\sigma^2 + \frac{1}{3}M\varepsilon\right)}}.$$

Bounding the Defect

Theorem A

Let $f : \mathbb{R}^d \to \mathbb{R}^k$ and let $\sigma_f^2 = \mathbb{V}[(f(X) - Y)^2]$. Suppose that $|f(X) - Y| \leq M$ almost everywhere. Then, for all $\varepsilon > 0$,

$$\mathbb{P}\left\{|L_{\mathbf{z}}(f)| \leq \varepsilon\right\} \geq 1 - 2e^{-\frac{m\varepsilon^2}{2\left(\sigma_f^2 + \frac{1}{3}M\varepsilon\right)}}.$$

Bounding the Defect

Theorem A

Let $f : \mathbb{R}^d \to \mathbb{R}^k$ and let $\sigma_f^2 = \mathbb{V}[(f(X) - Y)^2]$. Suppose that $|f(X) - Y| \leq M$ almost everywhere. Then, for all $\varepsilon > 0$,

$$\mathbb{P}\left\{|L_{\mathbf{z}}(f)| \leq \varepsilon\right\} \geq 1 - 2e^{-\frac{m\varepsilon^2}{2\left(\sigma_f^2 + \frac{1}{3}M\varepsilon\right)}}.$$

Proof.

Apply Bernstein Inequality to $\xi = (f(X) - Y)^2$. □

Bounding the Defect

Theorem A

Let $f : \mathbb{R}^d \to \mathbb{R}^k$ and let $\sigma_f^2 = \mathbb{V}[(f(X) - Y)^2]$. Suppose that $|f(X) - Y| \leq M$ almost everywhere. Then, for all $\varepsilon > 0$,

$$\mathbb{P}\left\{|L_{\mathbf{z}}(f)| \leq \varepsilon\right\} \geq 1 - 2e^{-\frac{m\varepsilon^2}{2\left(\sigma_f^2 + \frac{1}{3}M\varepsilon\right)}}.$$

Proof.

Apply Bernstein Inequality to $\xi = (f(X) - Y)^2$. $\square$

Are we done?? We could just minimize the empirical error and bound the defect...

Bounding the Defect

Theorem A

Let $f : \mathbb{R}^d \to \mathbb{R}^k$ and let $\sigma_f^2 = \mathbb{V}[(f(X) - Y)^2]$. Suppose that $|f(X) - Y| \leq M$ almost everywhere. Then, for all $\varepsilon > 0$,

$$\mathbb{P}\left\{|L_{\mathbf{z}}(f)| \leq \varepsilon\right\} \geq 1 - 2e^{-\frac{m\varepsilon^2}{2\left(\sigma_f^2 + \frac{1}{3}M\varepsilon\right)}}.$$

Proof.

Apply Bernstein Inequality to $\xi = (f(X) - Y)^2$. $\square$

Are we done?? We could just minimize the empirical error and bound the defect...

Any f vanishing on the sample points makes the empirical error vanish!!!

Hypothesis Space

Definition

Let $\mathcal{H}$ be a compact subset of the Banach space $\{f : X \to Y, \text{ continuous}\}$ with norm $\|f\| := \max_{x \in X} |f(x)|$. We call $\mathcal{H}$ *hypothesis space* or *model space*.

Hypothesis Space

Definition

Let $\mathcal{H}$ be a compact subset of the Banach space $\{f : X \to Y, \text{ continuous}\}$ with norm $\|f\| := \max_{x \in X} |f(x)|$. We call $\mathcal{H}$ *hypothesis space* or *model space*.

Best Approximation in $\mathcal{H}$

Define the *best approximation in $\mathcal{H}$* via

$$\widehat{f}_{\mathcal{H}} := \operatorname{argmin}_{f \in \mathcal{H}} \mathcal{E}(f) = \operatorname{argmin}_{f \in \mathcal{H}} \|\widehat{f} - f\|_{L^2(\mathbb{R}^d, d\mathbb{P}_X)}.$$

Hypothesis Space

Definition

Let $\mathcal{H}$ be a compact subset of the Banach space $\{f : X \to Y, \text{ continuous}\}$ with norm $\|f\| := \max_{x \in X} |f(x)|$. We call $\mathcal{H}$ *hypothesis space* or *model space*.

Best Approximation in $\mathcal{H}$

Define the *best approximation in* $\mathcal{H}$ via

$$\widehat{f}_{\mathcal{H}} := \operatorname{argmin}_{f \in \mathcal{H}} \mathcal{E}(f) = \operatorname{argmin}_{f \in \mathcal{H}} \|\widehat{f} - f\|_{L^2(\mathbb{R}^d, d\mathbb{P}_X)}.$$

Empirical Regression Function

Given $\mathbf{z}$ define the *empirical regression function* as

$$\widehat{f}_{\mathcal{H},\mathbf{z}} := \operatorname{argmin}_{f \in \mathcal{H}} \mathcal{E}_{\mathbf{z}}(f).$$

Hypothesis Space

Definition

Let $\mathcal{H}$ be a compact subset of the Banach space $\{f : X \to Y, \text{ continuous}\}$ with norm $\|f\| := \max_{x\in X} |f(x)|$. We call $\mathcal{H}$ *hypothesis space* or *model space*.

Best Approximation in $\mathcal{H}$

Define the *best approximation in $\mathcal{H}$* via

$$\widehat{f}_{\mathcal{H}} := \operatorname{argmin}_{f\in\mathcal{H}} \mathcal{E}(f) = \operatorname{argmin}_{f\in\mathcal{H}} \|\widehat{f} - f\|_{L^2(\mathbb{R}^d, d\mathbb{P}_X)}.$$

Empirical Regression Function

Given $\mathbf{z}$ define the *empirical regression function* as

$$\widehat{f}_{\mathcal{H},\mathbf{z}} := \operatorname{argmin}_{f\in\mathcal{H}} \mathcal{E}_{\mathbf{z}}(f).$$

 The empirical regression function can be computed!

1.2.3. Bias-Variance Decomposition

Generalization- and Approximation Error

Theorem (Bias-Variance Decomposition)

It holds that

$$\|\widehat{f}_{\mathcal{H},\mathbf{z}} - \widehat{f}\|^2_{L^2(\mathbb{R}^d, d\mathbb{P}_X)} = \left(\mathcal{E}(\widehat{f}_{\mathcal{H},\mathbf{z}}) - \mathcal{E}(\widehat{f}_{\mathcal{H}})\right) + \|\widehat{f}_{\mathcal{H}} - \widehat{f}\|^2_{L^2(\mathbb{R}^d, d\mathbb{P}_X)}.$$

The first term is called *generalization error* and the second term is called *approximation error*.

Generalization- and Approximation Error

Theorem (Bias-Variance Decomposition)

It holds that

$$\|\widehat{f}_{\mathcal{H},\mathbf{z}} - \widehat{f}\|^2_{L^2(\mathbb{R}^d, d\mathbb{P}_X)} = \Big(\mathcal{E}(\widehat{f}_{\mathcal{H},\mathbf{z}}) - \mathcal{E}(\widehat{f}_{\mathcal{H}})\Big) + \|\widehat{f}_{\mathcal{H}} - \widehat{f}\|^2_{L^2(\mathbb{R}^d, d\mathbb{P}_X)}.$$

The first term is called *generalization error* and the second term is called *approximation error*.

Proof.

By the Main Regression Theorem

$$\begin{aligned}
\|\widehat{f}_{\mathcal{H},\mathbf{z}} - \widehat{f}\|^2_{L^2(\mathbb{R}^d, d\mathbb{P}_X)} &= \mathcal{E}(\widehat{f}_{\mathcal{H},\mathbf{z}}) - \mathcal{E}(\widehat{f}) \\
&= \mathcal{E}(\widehat{f}_{\mathcal{H},\mathbf{z}}) - \mathcal{E}(\widehat{f}_{\mathcal{H}}) + \mathcal{E}(\widehat{f}_{\mathcal{H}}) - \mathcal{E}(\widehat{f}) \\
&= \Big(\mathcal{E}(\widehat{f}_{\mathcal{H},\mathbf{z}}) - \mathcal{E}(\widehat{f}_{\mathcal{H}})\Big) + \|\widehat{f}_{\mathcal{H}} - \widehat{f}\|^2_{L^2(\mathbb{R}^d, d\mathbb{P}_X)}.
\end{aligned}$$

$\square$

Generalization- and Approximation Error

Theorem (Bias-Variance Decomposition)

It holds that

$$\|\widehat{f}_{\mathcal{H},\mathbf{z}} - \widehat{f}\|^2_{L^2(\mathbb{R}^d, d\mathbb{P}_X)} = \Big(\mathcal{E}(\widehat{f}_{\mathcal{H},\mathbf{z}}) - \mathcal{E}(\widehat{f}_{\mathcal{H}})\Big) + \|\widehat{f}_{\mathcal{H}} - \widehat{f}\|^2_{L^2(\mathbb{R}^d, d\mathbb{P}_X)}.$$

The first term is called *generalization error* and the second term is called *approximation error*.

Proof.

By the M

$$\|\widehat{f}_{\mathcal{H},}$$

Our goal is to make the empirical error

$$\|\widehat{f}_{\mathcal{H},\mathbf{z}} - \widehat{f}\|^2_{L^2(\mathbb{R}^d, d\mathbb{P}_X)}$$

as small as possible.

$$= \Big(\mathcal{E}(\widehat{f}_{\mathcal{H},\mathbf{z}}) - \mathcal{E}(\widehat{f}_{\mathcal{H}})\Big) + \|\widehat{f}_{\mathcal{H}} - \widehat{f}\|^2_{L^2(\mathbb{R}^d, d\mathbb{P}_X)}.$$

□

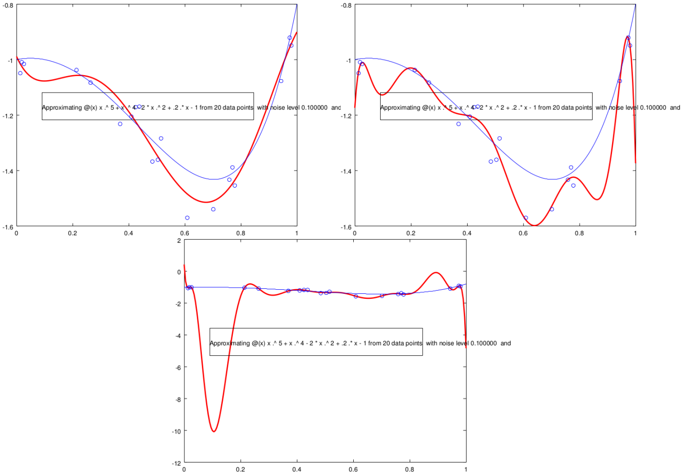


Figure: Blue: $f_{\mathcal{H}}$, Red: $f_{\mathcal{H},\mathbf{z}}$, $m = 10$, $\mathcal{H} =$ polynomials of degree $5, 15, 20$ (from top left to bottom).

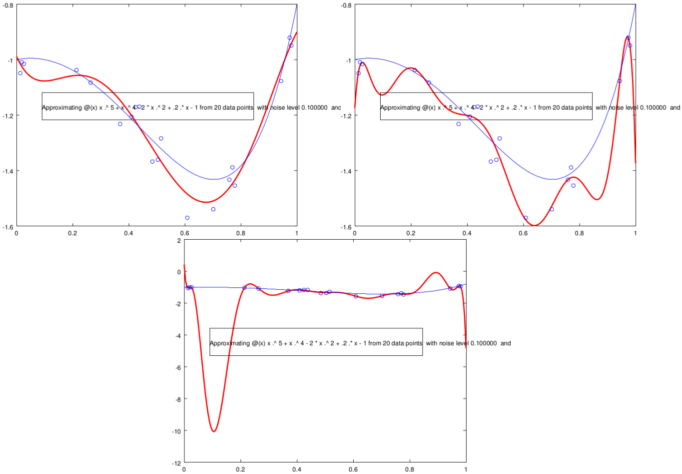


Figure: Blue: $f_{\mathcal{H}}$, Red: $f_{\mathcal{H},\mathbf{z}}$, $m = 10$, $\mathcal{H} =$ polynomials of degree $5, 15, 20$ (from top left to bottom).

If $\mathcal{H}$ is too complex, the sampling error increases.

The Bias-Variance Trade-Off

If we keep the sample size m fixed and enlarge the hypothesis space $\mathcal{H}$, the approximation error will certainly decrease, BUT the sample error will increase – this is exactly what we observed experimentally!

The Bias-Variance Trade-Off

If we keep the sample size m fixed and enlarge the hypothesis space $\mathcal{H}$, the approximation error will certainly decrease, BUT the sample error will increase – this is exactly what we observed experimentally!

Bishop [Neural Networks for Pattern Recognition (1995)]

> *"A model which is too simple, or too inflexible, will have a large bias, while one which has too much flexibility in relation to the particular data set will have a large variance. Bias and variance are complementary quantities, and the best generalization is obtained when we have the best compromise between the conflicting requirements of small bias and small variance."*

The Bias-Variance Trade-Off

If we keep the sample size m fixed and enlarge the hypothesis space $\mathcal{H}$, the approximation error will certainly decrease, BUT the sample error will increase – this is exactly what we observed experimentally!

Bishop [Neural Networks for Pattern Recognition (1995)]

> *"A model which is too simple, or too inflexible, will have a large bias, while one which has too much flexibility in relation to the particular data set will have a large variance. Bias and variance are complementary quantities, and the best generalization is obtained when we have the best compromise between the conflicting requirements of small bias and small variance."*

Bias-Variance Problem

What are the precise relations between the number of samples m and the "capacity" of our hypothesis space $\mathcal{H}$?

1.2.4 Bounds on the Generalization Error $\mathcal{E}(\widehat{f}_{\mathcal{H},\mathbf{z}}) - \mathcal{E}(\widehat{f}_{\mathcal{H}})$.

Covering Numbers

Definition

Let S be a metric space and $s > 0$. Define the *covering number* $\mathcal{N}(S, s)$ to be the minimal $l \in \mathcal{N}$ such that there exist l disks in S with radius s covering S.

Covering Numbers

Definition

Let S be a metric space and $s > 0$. Define the *covering number* $\mathcal{N}(S, s)$ to be the minimal $l \in \mathcal{N}$ such that there exist l disks in S with radius s covering S.

Covering Numbers

Definition

Let S be a metric space and $s > 0$. Define the *covering number* $\mathcal{N}(S, s)$ to be the minimal $l \in \mathcal{N}$ such that there exist l disks in S with radius s covering S.

Scaling of $\mathcal{N}(S, s)$ with s is a measure of complexity of S termed *metric entropy*.

Abstract Analysis of Generalization Error

Theorem B

Let $\mathcal{H} \subset C(X)$ be a hypothesis class. Assume that for all $f \in \mathcal{H}$ it holds that $|f(X) - Y| < M$ a.e. Then, for all $\varepsilon > 0$,

$$\mathbb{P}\left(\sup_{f\in\mathcal{H}} |L_{\mathbf{z}}(f)| \le \varepsilon\right) \ge 1 - \mathcal{N}(\mathcal{H}, \frac{\varepsilon}{8M})2e^{-\frac{m\varepsilon^2}{4(2\sigma^2+\frac{1}{3}M^2\varepsilon)}},$$

where $\sigma^2 := \sup_{f\in\mathcal{H}} \sigma_f^2$.

Proof.

First show that for all f, g with $\|f - g\| \leq \tau$ it holds that

$$|\mathcal{E}(f) - \mathcal{E}(g)| \leq 2M\tau \quad \text{and} \quad |\mathcal{E}_{\mathbf{z}}(f) - \mathcal{E}_{\mathbf{z}}(g)| \leq 2M\tau.$$

Cover $\mathcal{H}$ with balls $(U_i)_{i=1}^{\mathcal{N}(\mathcal{H},\epsilon/(8M))}$ with center f_i of radius $\frac{\epsilon}{8M}$. By the estimate above it holds that

$$\left(\sup_{f\in U_i} |\mathcal{E}_{\mathbf{z}}(f) - \mathcal{E}(f)| > \varepsilon\right) \Rightarrow \left(|\mathcal{E}_{\mathbf{z}}(f_i) - \mathcal{E}(f_i)| > \varepsilon/2\right)$$

Then by this fact and Theorem A it holds that

$$\begin{aligned}\mathbb{P}\left(\sup_{f\in\mathcal{H}} |L_{\mathbf{z}}(f)| > \varepsilon\right) &\leq \sum_{i=1}^{\mathcal{N}(\mathcal{H},\epsilon/(8M))} \mathbb{P}\left(\sup_{f\in U_i} |\mathcal{E}_{\mathbf{z}}(f) - \mathcal{E}(f)| > \varepsilon\right) \\ &\leq \sum_{i=1}^{\mathcal{N}(\mathcal{H},\epsilon/(8M))} \mathbb{P}\left(|\mathcal{E}_{\mathbf{z}}(f_i) - \mathcal{E}(f_i)| > \varepsilon/2\right) \\ &\leq \mathcal{N}(\mathcal{H}, \epsilon/(8M)) 2e^{-\frac{m\varepsilon^2}{4(2\sigma^2 + \frac{1}{3}M^2\varepsilon)}}.\end{aligned}$$

$\square$

Abstract Analysis of Generalization Error

Lemma

Let $\varepsilon > 0$ and $0 < \delta < 1$ such that

$$\mathbb{P}(\sup_{f \in \mathcal{H}} |L_{\mathbf{z}}(f)| \leq \varepsilon) \geq 1 - \delta.$$

Then

$$\mathbb{P}(\mathcal{E}(\widehat{f}_{\mathcal{H},\mathbf{z}}) - \mathcal{E}(\widehat{f}_{\mathcal{H}}) \leq 2\varepsilon) \geq 1 - \delta.$$

Abstract Analysis of Generalization Error

Lemma

Let $\varepsilon > 0$ and $0 < \delta < 1$ such that

$$\mathbb{P}(\sup_{f \in \mathcal{H}} |L_{\mathbf{z}}(f)| \leq \varepsilon) \geq 1 - \delta.$$

Then

$$\mathbb{P}(\mathcal{E}(\widehat{f}_{\mathcal{H},\mathbf{z}}) - \mathcal{E}(\widehat{f}_{\mathcal{H}}) \leq 2\varepsilon) \geq 1 - \delta.$$

Proof.

Suppose that $\sup_{f \in \mathcal{H}} |L_{\mathbf{z}}(f)| \leq \varepsilon$. Then $|\mathcal{E}_{\mathbf{z}}(\widehat{f}_{\mathcal{H},\mathbf{z}}) - \mathcal{E}(\widehat{f}_{\mathcal{H},\mathbf{z}})| \leq \varepsilon$, $|\mathcal{E}_{\mathbf{z}}(\widehat{f}_{\mathcal{H}}) - \mathcal{E}(\widehat{f}_{\mathcal{H}})| \leq \varepsilon$ and $\mathcal{E}_{\mathbf{z}}(\widehat{f}_{\mathcal{H},\mathbf{z}}) - \mathcal{E}_{\mathbf{z}}(\widehat{f}_{\mathcal{H}}) \leq 0$. It follows that

$$\begin{aligned}
\mathcal{E}(\widehat{f}_{\mathcal{H},\mathbf{z}}) - \mathcal{E}(\widehat{f}_{\mathcal{H}}) &= \mathcal{E}(\widehat{f}_{\mathcal{H},\mathbf{z}}) - \mathcal{E}_{\mathbf{z}}(\widehat{f}_{\mathcal{H},\mathbf{z}}) + \mathcal{E}_{\mathbf{z}}(\widehat{f}_{\mathcal{H},\mathbf{z}}) - \mathcal{E}_{\mathbf{z}}(\widehat{f}_{\mathcal{H}}) + \\
&\quad \mathcal{E}_{\mathbf{z}}(\widehat{f}_{\mathcal{H}}) - \mathcal{E}(\widehat{f}_{\mathcal{H}}) \\
&\leq |\mathcal{E}_{\mathbf{z}}(\widehat{f}_{\mathcal{H},\mathbf{z}}) - \mathcal{E}(\widehat{f}_{\mathcal{H},\mathbf{z}})| + |\mathcal{E}_{\mathbf{z}}(\widehat{f}_{\mathcal{H}}) - \mathcal{E}(\widehat{f}_{\mathcal{H}})| \leq 2\epsilon.
\end{aligned}$$

$\square$

Abstract Analysis of Generalization Error

Theorem C

Let $\mathcal{H}$ be a hypothesis class. Assume that for all $f \in \mathcal{H}$ it holds that $|f(X) - Y| < M$ a.e. Then, for all $\varepsilon > 0$,

$$\mathbb{P}\left(\mathcal{E}(\widehat{f}_{\mathcal{H},\mathbf{z}}) - \mathcal{E}(\widehat{f}_{\mathcal{H}}) \leq \varepsilon\right) \geq 1 - \mathcal{N}(\mathcal{H}, \frac{\varepsilon}{16M})2e^{-\frac{m\varepsilon^2}{8(2\sigma^2 + \frac{1}{3}M^2\varepsilon)}},$$

where $\sigma^2 := \sup_{f\in\mathcal{H}} \sigma_f^2$.

Abstract Analysis of Generalization Error

Theorem C

Let $\mathcal{H}$ be a hypothesis class. Assume that for all $f \in \mathcal{H}$ it holds that $|f(X) - Y| < M$ a.e. Then, for all $\varepsilon > 0$,

$$\mathbb{P}\left(\mathcal{E}(\widehat{f}_{\mathcal{H},\mathbf{z}}) - \mathcal{E}(\widehat{f}_{\mathcal{H}}) \leq \varepsilon\right) \geq 1 - \mathcal{N}(\mathcal{H}, \frac{\varepsilon}{16M})2e^{-\frac{m\varepsilon^2}{8(2\sigma^2 + \frac{1}{3}M^2\varepsilon)}},$$

where $\sigma^2 := \sup_{f \in \mathcal{H}} \sigma_f^2$.

Proof.

Apply Lemma and Theorem B with $\epsilon \leftrightarrow \epsilon/2$. $\square$

Abstract Analysis of Generalization Error

Question

Given $\varepsilon, \delta > 0$, how many samples m do we need such that the probability that the generalization error is $\leq \varepsilon$ is at least $1 - \delta$?

Abstract Analysis of Generalization Error

Question

Given $\varepsilon, \delta > 0$, how many samples m do we need such that the probability that the generalization error is $\leq \varepsilon$ is at least $1 - \delta$?

Answer

By the previous theorem it suffices to choose

$$m \geq \frac{8(4\sigma^2 + \frac{1}{3}M^2\varepsilon)}{\varepsilon^2}\left(\ln(2\mathcal{N}(\mathcal{H}, \frac{\varepsilon}{16M})) + \ln(\frac{1}{\delta})\right).$$

Abstract Analysis of Generalization Error

Question

Given $\varepsilon, \delta > 0$, how many samples m do we need such that the probability that the generalization error is $\leq \varepsilon$ is at least $1 - \delta$?

Answer

By the previous theorem it suffices to choose

$$m \geq \frac{8(4\sigma^2 + \frac{1}{3}M^2\varepsilon)}{\varepsilon^2}\left(\ln(2\mathcal{N}(\mathcal{H}, \frac{\varepsilon}{16M})) + \ln(\frac{1}{\delta})\right).$$

Question

How to bound the covering number?

1.2.5 A Simple Example

Linear Regression

Recall

$$\mathcal{H}_{l,R} = \mathsf{span}\,\{\varphi_1, \ldots, \varphi_l\} \cap \{f \in C(\Omega) : \|f\| \leq R\} \subset C(\Omega).$$

Linear Regression

Recall

$$\mathcal{H}_{l,R} = \text{span}\,\{\varphi_1, \dots, \varphi_l\} \cap \{f \in C(\Omega) : \|f\| \leq R\} \subset C(\Omega).$$

Theorem

Let $T := \| \sum_{j=1}^{l} |\varphi_j| \|$. Then

$$\ln(\mathcal{N}(\mathcal{H}_R, \eta)) \leq l \cdot \ln\left(\frac{4RT}{\eta}\right).$$

Linear Regression

Recall

$$\mathcal{H}_{l,R} = \text{span}\{\varphi_1, \dots, \varphi_l\} \cap \{f \in C(\Omega) : \|f\| \leq R\} \subset C(\Omega).$$

Theorem

Let $T := \|\sum_{j=1}^{l} |\varphi_j|\|$. Then

$$\ln(\mathcal{N}(\mathcal{H}_R, \eta)) \leq l \cdot \ln\left(\frac{4RT}{\eta}\right).$$

In the motivational section on linear regression we have seen that $f_{\mathcal{H},\mathbf{z}}$ can be found by solving an l-dimensional linear system.

Analysis of Linear Regression

Theorem

Suppose that we have the approximation error estimate

$$\inf_{f \in \mathcal{H}_{l,R}} \|\widehat{f} - f\|^2_{L^2(\mathbb{R}^d, d\mathbb{P}_X)} \leq \frac{\epsilon}{2}.$$

Then

$$m \gtrsim \frac{\left(l \cdot \mathrm{polylog}(\epsilon) + \ln(\frac{1}{\delta})\right)}{\epsilon^2}$$

independent training samples suffice to get an empirical error l with probability $\geq 1 - \delta$.

More Advanced Topics

More Advanced Topics

- General Loss function (see for example Devroye, Gyorfi, Lugosi: A Probabilistic Theory of Pattern Recognition)

More Advanced Topics

- General Loss function (see for example Devroye, Gyorfi, Lugosi: A Probabilistic Theory of Pattern Recognition)
- Other (sharper) complexity measures such as Rademacher complexity, VC dimension, empirical oscillation, ... (see for example Devroye, Gyorfi, Lugosi: A Probabilistic Theory of Pattern Recognition)

More Advanced Topics

- General Loss function (see for example Devroye, Gyorfi, Lugosi: A Probabilistic Theory of Pattern Recognition)
- Other (sharper) complexity measures such as Rademacher complexity, VC dimension, empirical oscillation, ... (see for example Devroye, Gyorfi, Lugosi: A Probabilistic Theory of Pattern Recognition)
- Weaker conditions on (X, Y) (we essentially require bounded noise!!) (see for example Mendelson: Learning without Concentration)

More Advanced Topics

- General Loss function (see for example Devroye, Gyorfi, Lugosi: A Probabilistic Theory of Pattern Recognition)
- Other (sharper) complexity measures such as Rademacher complexity, VC dimension, empirical oscillation, ... (see for example Devroye, Gyorfi, Lugosi: A Probabilistic Theory of Pattern Recognition)
- Weaker conditions on (X, Y) (we essentially require bounded noise!!) (see for example Mendelson: Learning without Concentration)
- Better learning procedures than ERM (see for example Mendelson: An Optimal Unrestricted Learnning Procedure)

More Advanced Topics

- General Loss function (see for example Devroye, Gyorfi, Lugosi: A Probabilistic Theory of Pattern Recognition)
- Other (sharper) complexity measures such as Rademacher complexity, VC dimension, empirical oscillation, ... (see for example Devroye, Gyorfi, Lugosi: A Probabilistic Theory of Pattern Recognition)
- Weaker conditions on (X, Y) (we essentially require bounded noise!!) (see for example Mendelson: Learning without Concentration)
- Better learning procedures than ERM (see for example Mendelson: An Optimal Unrestricted Learnning Procedure)
- Better sampling procedures (see for example Cohen, Migliorati: Optimal Weighted Least Squares Methods)

1.3 Reproducing Kernel Hilbert Spaces (RKHS)

1.3.1 Definition

Reproducing Kernel Hilbert Spaces (RKHS)

Motivation

Suppose we have a 'similarity measure' $K : \Omega \times \Omega \to \mathbb{R}$ on Ω and we would like to do things like nearest neighbour search, PCA etc. with respect to this measure of similarity. One idea is to associate each $x \in \Omega$ with a *feature map* Φ_x which is an element of a high-dimensional inner-product-space, but which 'linearizes' the similarity measure in the sense that

$$K(x, x') = \langle \Phi_x, \Phi_{x'} \rangle.$$

Reproducing Kernel Hilbert Spaces (RKHS)

Motivation

Suppose we have a 'similarity measure' $K : \Omega \times \Omega \to \mathbb{R}$ on Ω and we would like to do things like nearest neighbour search, PCA etc. with respect to this measure of similarity. One idea is to associate each $x \in \Omega$ with a *feature map* Φ_x which is an element of a high-dimensional inner-product-space, but which 'linearizes' the similarity measure in the sense that

$$K(x, x') = \langle \Phi_x, \Phi_{x'} \rangle.$$

Question

Which conditions on K guarantee the existence of a feature map?

Mercer Theorem

Definition

$K : \Omega \times \Omega \to \mathbb{R}$ is symmetric if $K(x, x') = K(x', x)$ for all $x, x' \in \Omega$. Let $\mathbf{x} = \{x_1, \ldots, x_k\} \subset \Omega$ and $K[\mathbf{x}] \in \mathbb{R}^{k \times k}$ with entries $K(x_i, x_j)$ the *Gramian* of K at $\mathbf{x}$. K is called *positive semidefinite* if its Gramian is always positive semidefinite. K is called a *Mercer kernel* if it is symmetric, positive semidefinite and continuous.

Mercer Theorem

Definition

$K : \Omega \times \Omega \to \mathbb{R}$ is symmetric if $K(x, x') = K(x', x)$ for all $x, x' \in \Omega$. Let $\mathbf{x} = \{x_1, \dots, x_k\} \subset \Omega$ and $K[\mathbf{x}] \in \mathbb{R}^{k \times k}$ with entries $K(x_i, x_j)$ the *Gramian* of K at $\mathbf{x}$. K is called *positive semidefinite* if its Gramian is always positive semidefinite. K is called a *Mercer kernel* if it is symmetric, positive semidefinite and continuous.

Theorem

There exists a unique Hilbert space $\mathcal{H}_K$ of functions on Ω satisfying

1. The functions $K_x : x' \mapsto K(x, x')$ are in $\mathcal{H}_K$,
2. the span of the K_x's is dense, and
3. for all $f \in \mathcal{H}_K$ we have $f(x) = \langle f, K_k \rangle$.

Mercer Theorem

Definition

$K : \Omega \times \Omega \to \mathbb{R}$ is symmetric if $K(x, x') = K(x', x)$ for all $x, x' \in \Omega$. Let $\mathbf{x} = \{x_1, \ldots, x_k\} \subset \Omega$ and $K[\mathbf{x}] \in \mathbb{R}^{k \times k}$ with entries $K(x_i, x_j)$ the *Gramian* of K at $\mathbf{x}$. K is called *positive semidefinite* if its Gramian is always positive semidefinite. K is called a *Mercer kernel* if it is symmetric, positive semidefinite and continuous.

Theorem

There exists a unique Hilbert space $\mathcal{H}_K$ of functions on Ω satisfying

1. The functions $K_x : x' \mapsto K(x, x')$ are in $\mathcal{H}_K$,
2. the span of the K_x's is dense, and
3. for all $f \in \mathcal{H}_K$ we have $f(x) = \langle f, K_k \rangle$.

In particular, $K(x, x') = \langle K_x, K_{x'} \rangle$ and in this sense, the RKHS $\mathcal{H}_K$ can be regarded as a feature space.

Proof.

Consider finite sums

$$f(x) = \sum_{i=1}^{m} w_i K(x_i, x), \quad g(x) = \sum_{i=1}^{m} v_i K(x_i, x)$$

with inner product $\langle f, g \rangle = \mathbf{w}^T K[\mathbf{x}]\mathbf{v}$ and complete. $\square$

Examples I

Dot-Product Kernels

Let Ω be the ball of radius T in $\mathbb{R}^d$ and $K(x, x') = \sum_{d=1}^{\infty} a_d(x \cdot x')$, where $a_d \geq 0$ for all d and $\sum_d a_d T^{2d} < \infty$. Then K is a mercer kernel on Ω called a *dot product kernel*.

Examples I

Dot-Product Kernels

Let Ω be the ball of radius T in $\mathbb{R}^d$ and $K(x, x') = \sum_{d=1}^{\infty} a_d(x \cdot x')$, where $a_d \geq 0$ for all d and $\sum_d a_d T^{2d} < \infty$. Then K is a mercer kernel on Ω called a *dot product kernel*.

Example

Suppose that Ω is as above and $K(x, x') = 1 + x \cdot x'$. Then $\{1, x_1, \ldots, x_n\}$ constitutes on ONB of $\mathcal{H}_K$.

Examples II

Translation-Invariant Kernels

Suppose that $k : \mathbb{R}^d \to \mathbb{R}$ is such that its Fourier transform is real-valued and non-negative. Then $K(x, x') := k(x - x')$ is a mercer kernel, called a *translation-invariant kernel*.

Examples II

Translation-Invariant Kernels

Suppose that $k : \mathbb{R}^d \to \mathbb{R}$ is such that its Fourier transform is real-valued and non-negative. Then $K(x, x') := k(x - x')$ is a mercer kernel, called a *translation-invariant kernel*.

Example

Let $k = \chi_{[-1,1]} * \chi_{[-1,1]}$ the cardinal B-spline of degree one. Then

$$K(x, x') = \begin{cases} 1 - \frac{|x-x'|}{2} & |x - x'| \leq 2 \\ 0 & \text{else} \end{cases} .$$

Examples III

Radial Basis Functions (RBF)

Suppose that $f : \mathbb{R}_+ \to \mathbb{R}$ is completely monotonic (i.e. $(-1)^k f^{(k)} \geq 0$). Then $K(x, x') := f(|x - x'|^2)$ is a mercer kernel, called a *RBF kernel*.

Examples III

Radial Basis Functions (RBF)

Suppose that $f : \mathbb{R}_+ \to \mathbb{R}$ is completely monotonic (i.e. $(-1)^k f^{(k)} \geq 0$). Then $K(x, x') := f(|x - x'|^2)$ is a mercer kernel, called a *RBF kernel*.

Example

A Gaussian $f(t) := e^{-t/c^2}$ and an inverse multiquadric $(c^2 + |t|)^{-\alpha}, \alpha > 0$ are completely monotonic and define corresponding RBF kernels.

Covering Numbers

Theorem

For $R > 0$ denote B_R the ball of radius R in a RKHS $\mathcal{H}_K$. Then B_R is a compact subset of $C(\Omega)$ and thus a valid hypothesis space.

Covering Numbers

Theorem

For $R > 0$ denote B_R the ball of radius R in a RKHS $\mathcal{H}_K$. Then B_R is a compact subset of $C(\Omega)$ and thus a valid hypothesis space.

Theorem

If $K \in C^s$ then

$$\ln(\mathcal{N}(B_R, \eta)) \leq C \cdot \mathsf{diam}(X)^n \|K\|_{C^s}^{n/s} \left(\frac{R}{\eta}\right)^{2n/s}.$$

Covering Numbers

Theorem

For $R > 0$ denote B_R the ball of radius R in a RKHS $\mathcal{H}_K$. Then B_R is a compact subset of $C(\Omega)$ and thus a valid hypothesis space.

Theorem

If $K \in C^s$ then

$$\ln(\mathcal{N}(B_R, \eta)) \leq C \cdot \mathrm{diam}(X)^n \|K\|_{C^s}^{n/s} \left(\frac{R}{\eta}\right)^{2n/s}.$$

For specific kernels (such as Gaussian RBF kernels), much better results exist.

1.3.2 Computation of the Empirical Regression Function

Computational Issues

Question

How can we determine $f_{\mathcal{H},\mathbf{z}}$?

Computational Issues

Question

How can we determine $f_{\mathcal{H},\mathbf{z}}$?

Let $\mathcal{H}_{K,\mathbf{z}} := \text{span}\{K_{x_1}, \ldots, K_{x_m}\}$ and $P : \mathcal{H}_K \to \mathcal{H}_{K,\mathbf{z}}$ the orthogonal projection. Then, since $f(x_i) = \langle f, K_{x_i} \rangle$ $= \langle P(f), K_{x_i} \rangle = P(f)(x_i)$ we have $\mathcal{E}_{\mathbf{z}}(f) = \mathcal{E}_{\mathbf{z}}(P(f))$!!!

Computational Issues

Question

How can we determine $f_{\mathcal{H},\mathbf{z}}$?

Let $\mathcal{H}_{K,\mathbf{z}} := \text{span}\{K_{x_1}, \ldots, K_{x_m}\}$ and $P : \mathcal{H}_K \to \mathcal{H}_{K,\mathbf{z}}$ the orthogonal projection. Then, since $f(x_i) = \langle f, K_{x_i} \rangle$ $= \langle P(f), K_{x_i} \rangle = P(f)(x_i)$ we have $\mathcal{E}_{\mathbf{z}}(f) = \mathcal{E}_{\mathbf{z}}(P(f))$!!!

Theorem

We have $f_{\mathcal{H},\mathbf{z}} = \sum_{i=1}^m c_i^* K_{x_i}$, where (c_i^*) is a minimizer of

$$\frac{1}{m} \sum_{j=1}^m \left(\sum_{i=1}^m c_i K(x_i, x_j) - y_j \right)^2 \quad \text{s.t.} \quad c^T K[\mathbf{x}] c \leq R^2.$$

Computational Issues

Question

How can we determine $f_{\mathcal{H},\mathbf{z}}$?

Let $\mathcal{H}_{K,\mathbf{z}} := \text{span}\{K_{x_1}, \ldots, K_{x_m}\}$ and $P : \mathcal{H}_K \to \mathcal{H}_{K,\mathbf{z}}$ the orthogonal projection. Then, since $f(x_i) = \langle f, K_{x_i} \rangle$ $= \langle P(f), K_{x_i} \rangle = P(f)(x_i)$ we have $\mathcal{E}_{\mathbf{z}}(f) = \mathcal{E}_{\mathbf{z}}(P(f))$!!!

Theorem

We have $f_{\mathcal{H},\mathbf{z}} = \sum_{i=1}^m c_i^* K_{x_i}$, where (c_i^*) is a minimizer of

$$\frac{1}{m} \sum_{j=1}^m \left(\sum_{i=1}^m c_i K(x_i, x_j) - y_j \right)^2 \quad \text{s.t.} \quad c^T K[\mathbf{x}] c \leq R^2.$$

This is a convex quadratic program that can be efficiently solved by interior point methods!

Computational Issues

Question

How can we determine $f_{\mathcal{H},\mathbf{z}}$?

Let $\mathcal{H}_{K,\mathbf{z}} := \text{span}\{K_{x_1}, \ldots, K_{x_m}\}$ and $P : \mathcal{H}_K \to \mathcal{H}_{K,\mathbf{z}}$ the orthogonal projection. Then, since $f(x_i) = \langle f, K_{x_i} \rangle$ $= \langle P(f), K_{x_i} \rangle = P(f)(x_i)$ we have $\mathcal{E}_{\mathbf{z}}(f) = \mathcal{E}_{\mathbf{z}}(P(f))$!!!

Theorem

We have $f_{\mathcal{H},\mathbf{z}} = \sum_{i=1}^m c_i^* K_{x_i}$, where (c_i^*) is a minimizer of

$$\frac{1}{m} \sum_{j=1}^m \left(\sum_{i=1}^m c_i K(x_i, x_j) - y_j \right)^2 \quad \text{s.t.} \quad c^T K[\mathbf{x}] c \leq R^2.$$

This is a convex quadratic program that can be efficiently solved by interior point methods! Check out `http://cvxr.com/cvx/`!

1.3.3 A Bayesian Interpretation

Bayes' Theorem

Bayes' Theorem

- Suppose that the probability of measuring $f \in \mathcal{H}_K$ is equal to $\mathbb{P}(f) = c \cdot \exp(-\|f\|^2_{\mathcal{H}_K})$.

Bayes' Theorem

- Suppose that the probability of measuring $f \in \mathcal{H}_K$ is equal to $\mathbb{P}(f) = c \cdot \exp(-\|f\|^2_{\mathcal{H}_K})$.
- Suppose that we observe a function f, corrupted with Gaussian noise.

Bayes' Theorem

- Suppose that the probability of measuring $f \in \mathcal{H}_K$ is equal to $\mathbb{P}(f) = c \cdot \exp(-\|f\|^2_{\mathcal{H}_K})$.
- Suppose that we observe a function f, corrupted with Gaussian noise. Then the probability of making the observations $\mathbf{z} = ((x_1, y_1), \ldots, (x_m, y_m))$, from the signal f is equal to

$$\mathbb{P}(\mathbf{z}|f) = d \cdot \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{m}(f(x_i) - y_i)^2\right).$$

Bayes' Theorem

- Suppose that the probability of measuring $f \in \mathcal{H}_K$ is equal to $\mathbb{P}(f) = c \cdot \exp(-\|f\|^2_{\mathcal{H}_K})$.
- Suppose that we observe a function f, corrupted with Gaussian noise. Then the probability of making the observations $\mathbf{z} = ((x_1, y_1), \ldots, (x_m, y_m))$, from the signal f is equal to

$$\mathbb{P}(\mathbf{z}|f) = d \cdot \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{m}(f(x_i) - y_i)^2\right).$$

- *Bayes' Theorem* yields that

$$\mathbb{P}(f|\mathbf{z}) = \frac{\mathbb{P}(\mathbf{z}|f) \cdot \mathbb{P}(f)}{\mathbb{P}(\mathbf{z})}$$

Bayes' Theorem

- Suppose that the probability of measuring $f \in \mathcal{H}_K$ is equal to $\mathbb{P}(f) = c \cdot \exp(-\|f\|^2_{\mathcal{H}_K})$.
- Suppose that we observe a function f, corrupted with Gaussian noise. Then the probability of making the observations $\mathbf{z} = ((x_1, y_1), \ldots, (x_m, y_m))$, from the signal f is equal to

$$\mathbb{P}(\mathbf{z}|f) = d \cdot \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{m}(f(x_i) - y_i)^2\right).$$

- *Bayes' Theorem* yields that

$$\mathbb{P}(f|\mathbf{z}) = \frac{\mathbb{P}(\mathbf{z}|f) \cdot \mathbb{P}(f)}{\mathbb{P}(\mathbf{z})} \sim \exp(-\frac{1}{2\sigma^2}\sum_{i=1}^{m}(f(x_i) - y_i)^2 - \|f\|^2_{\mathcal{H}_K}).$$

Maximum A Posteriori (MAP) Estimate

MAP Estimate

The MAP Estimate maximizes the a posteriori probability $\mathbb{P}(f|\mathbf{z})$, given an a priori distribution on f and on the noise.

Maximum A Posteriori (MAP) Estimate

MAP Estimate

The MAP Estimate maximizes the a posteriori probability $\mathbb{P}(f|\mathbf{z})$, given an a priori distribution on f and on the noise.

For the a priori distribution $\mathbb{P}(f) = c \cdot \exp(-\|f\|^2_{\mathcal{H}_K})$ and Gaussian noise, the solution of the regularized least squares problem is also the MAP estimate!

1.4 Classification

The Classification Problem

The Classification Problem

We now aim at classifying data into two classes and thus look for $f : \Omega \to \{-1, 1\}$. Therefore, let's put $Y = \{-1, 1\}$ and $Z := \Omega \times Y$.

The Classification Problem

We now aim at classifying data into two classes and thus look for $f : \Omega \to \{-1, 1\}$. Therefore, let's put $Y = \{-1, 1\}$ and $Z := \Omega \times Y$.

Misclassification Error

Given a distribution (X, Y) and $f : X \to Y$, define the *misclassification error* as

$$\mathcal{R}(f) := \mathbb{P}_{(X,Y)}(f(X) \neq Y).$$

The Classification Problem

We now aim at classifying data into two classes and thus look for $f : \Omega \to \{-1, 1\}$. Therefore, let's put $Y = \{-1, 1\}$ and $Z := \Omega \times Y$.

Misclassification Error

Given a distribution (X, Y) and $f : X \to Y$, define the *misclassification error* as

$$\mathcal{R}(f) := \mathbb{P}_{(X,Y)}(f(X) \neq Y).$$

The classification problem asks to minimize the misclassification error.

Bayes Rule

Bayes Rule

Define the *Bayes rule* as

$$\widehat{f_c} := \mathrm{sgn}(\widehat{f}).$$

Bayes Rule

Bayes Rule

Define the *Bayes rule* as

$$\widehat{f}_c := \mathsf{sgn}(\widehat{f}).$$

Theorem

The Bayes rule minimizes the misclassification error

Bayes Rule

Bayes Rule

Define the *Bayes rule* as

$$\widehat{f}_c := \mathsf{sgn}(\widehat{f}).$$

Theorem

The Bayes rule minimizes the misclassification error

We can re–use everything!

Bayes Rule

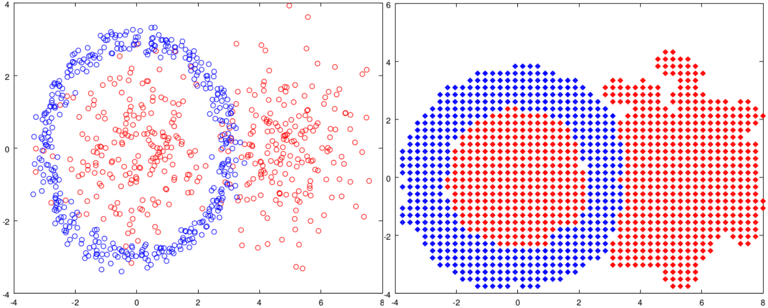


Figure: Bayes Rule for Gaussian Kernel regression. Left: sample data. Right: Estimate using Bayes Rule.

Case Study: Breast Cancer Detection

```
2 1:-0.860107 2:-0.111111 3:-1 4:-1 5:-1 6:-0.777778 7:-1 8:-0.555556 9:-1 10:-1
2 1:-0.859671 2:-0.111111 3:-0.333333 4:-0.333333 5:-0.111111 6:0.333333 7:1 8:-0.555556 9:-0.777778 10:-1
2 1:-0.857807 2:-0.555556 3:-1 4:-1 5:-1 6:-0.777778 7:-0.777778 8:-0.555556 9:-1 10:-1
2 1:-0.85768 2:0.111111 3:0.555556 4:0.555556 5:-1 6:-0.555556 7:-0.333333 8:-0.555556 9:0.333333 10:-1
2 1:-0.857569 2:-0.333333 3:-1 4:-1 5:-0.555556 6:-0.777778 7:-1 8:-0.555556 9:-1 10:-1
4 1:-0.857554 2:0.555556 3:1 4:1 5:0.555556 6:0.333333 7:1 8:0.777778 9:0.333333 10:-1
2 1:-0.857408 2:-1 3:-1 4:-1 5:-1 6:-0.777778 7:1 8:-0.555556 9:-1 10:-1
2 1:-0.857339 2:-0.777778 3:-1 4:-0.777778 5:-1 6:-0.777778 7:-1 8:-0.555556 9:-1 10:-1
2 1:-0.855171 2:-0.777778 3:-1 4:-1 5:-1 6:-0.777778 7:-1 8:-1 9:-1 10:-0.111111
2 1:-0.855171 2:-0.333333 3:-0.777778 4:-1 5:-1 6:-0.777778 7:-1 8:-0.777778 9:-1 10:-1
2 1:-0.854841 2:-1 3:-1 4:-1 5:-1 6:-1 7:-1 8:-0.555556 9:-1 10:-1
2 1:-0.854709 2:-0.777778 3:-1 4:-1 5:-1 6:-0.777778 7:-1 8:-0.777778 9:-1 10:-1
4 1:-0.853868 2:-0.111111 3:-0.555556 4:-0.555556 5:-0.555556 6:-0.777778 7:-0.555556 8:-0.333333 9:-0.333333 10:-1
2 1:-0.85354 2:-1 3:-1 4:-1 5:-1 6:-0.777778 7:-0.555556 8:-0.555556 9:-1 10:-1
4 1:-0.853454 2:0.555556 3:0.333333 4:-0.111111 5:1 6:0.333333 7:0.777778 8:-0.111111 9:-0.111111 10:-0.333333
4 1:-0.852997 2:0.333333 3:-0.333333 4:0.111111 5:-0.333333 6:0.111111 7:-1 8:-0.333333 9:-0.555556 10:-1
2 1:-0.852842 2:-0.333333 3:-1 4:-1 5:-1 6:-0.777778 7:-1 8:-0.777778 9:-1 10:-1
2 1:-0.852671 2:-0.333333 3:-1 4:-1 5:-1 6:-0.777778 7:-1 8:-0.555556 9:-1 10:-1
4 1:-0.852543 2:1 3:0.333333 4:0.333333 5:0.111111 6:-0.333333 7:1 8:-0.333333 9:-1 10:-0.777778
2 1:-0.852536 2:0.111111 3:-1 4:-1 5:-1 6:-0.777778 7:-1 8:-0.555556 9:-1 10:-1
4 1:-0.851958 2:0.333333 3:-0.555556 4:-0.777778 5:1 6:-0.111111 7:1 8:-0.111111 9:-0.333333 10:-0.333333
4 1:-0.851957 2:1 3:-0.111111 4:-0.111111 5:-0.555556 6:0.111111 7:0.333333 8:0.333333 9:1 10:-1
2 1:-0.85163 2:-0.555556 3:-1 4:-1 5:-1 6:-0.777778 7:-1 8:-0.777778 9:-1 10:-1
2 1:-0.851217 2:-1 3:-1 4:-1 5:-1 6:-0.777778 7:-1 8:-0.555556 9:-1 10:-1
4 1:-0.850295 2:-0.111111 3:-0.777778 4:-0.555556 5:-0.333333 6:-0.777778 7:0.333333 8:-0.555556 9:0.111111 10:-1
2 1:-0.850198 2:-0.555556 3:-0.777778 4:-1 5:-1 6:-1 7:-1 8:-0.777778 9:-1 10:-1
2 1:-0.850107 2:-0.111111 3:-1 4:-1 5:-1 6:-0.777778 7:-1 8:-0.777778 9:-1 10:-1
2 1:-0.850038 2:-0.777778 3:-1 4:-1 5:-1 6:-0.777778 7:-1 8:-0.777778 9:-1 10:-1
2 1:-0.849517 2:-1 3:-1 4:-0.555556 5:-1 6:-0.777778 7:-1 8:-1 9:-1 10:-1
2 1:-0.849517 2:-0.555556 3:-1 4:-1 5:-1 6:-1 7:-1 8:-0.777778 9:-1 10:-1
2 1:-0.849393 2:-0.777778 3:-1 4:-1 5:-1 6:-0.777778 7:-1 8:-0.555556 9:-1 10:-1
4 1:-0.849331 2:1 3:0.333333 4:0.333333 5:-0.555556 6:0.555556 7:-0.111111 8:0.333333 9:-0.333333 10:-0.555556
2 1:-0.848968 2:-0.777778 3:-1 4:-1 5:-0.777778 6:-0.777778 7:-1 8:-0.555556 9:-1 10:-1
2 1:-0.848891 2:-0.555556 3:-1 4:-0.777778 5:-1 6:-0.777778 7:-1 8:-0.777778 9:-1 10:-1
2 1:-0.848267 2:-0.777778 3:-1 4:-1 5:-1 6:-0.777778 7:-1 8:-0.777778 9:-1 10:-1
4 1:-0.848135 2:1 3:1 4:1 5:0.555556 6:0.111111 7:-1 8:0.555556 9:0.777778 10:-1
2 1:-0.847895 2:0.111111 3:-0.777778 4:-1 5:-1 6:-1 7:-1 8:0.333333 9:-1 10:-1
4 1:-0.847478 2:-0.111111 3:-0.333333 4:-0.333333 5:0.777778 6:-0.777778 7:1 8:-0.111111 9:0.111111 10:-1
4 1:-0.846481 2:-0.777778 3:-0.111111 4:-0.555556 5:-0.555556 6:0.111111 7:0.333333 8:0.333333 9:-0.111111 10:-1
4 1:-0.845249 2:1 3:-0.333333 4:-0.555556 5:-1 6:-0.555556 7:-0.555556 8:0.111111 9:-0.111111 10:-0.777778
4 1:-0.845097 2:0.111111 3:1 4:1 5:-0.777778 6:0.555556 7:1 8:0.333333 9:-0.555556 10:-0.555556
4 1:-0.844791 2:-0.111111 3:0.111111 4:-0.111111 5:0.111111 6:1 7:-1 8:-0.555556 9:-1 10:-1
```

Classification Results based on Kernel Regression

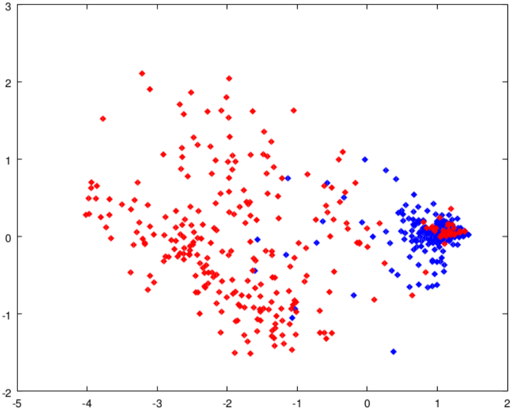
Size of dataset $m = 683$ and dimensionality of feature space $d = 10$.

Classification Results based on Kernel Regression

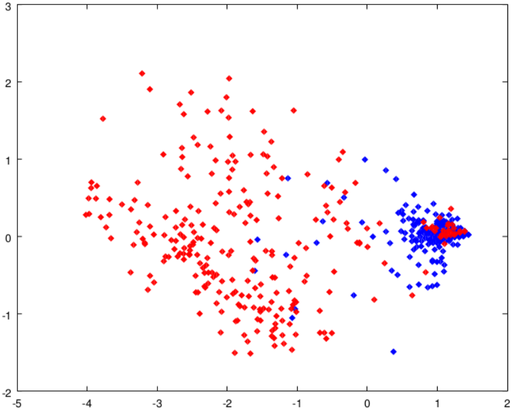
Size of dataset $m = 683$ and dimensionality of feature space $d = 10$.

We obtain 95 percent classification accuracy from only 68 training samples and linear kernel!

Visualizing Data

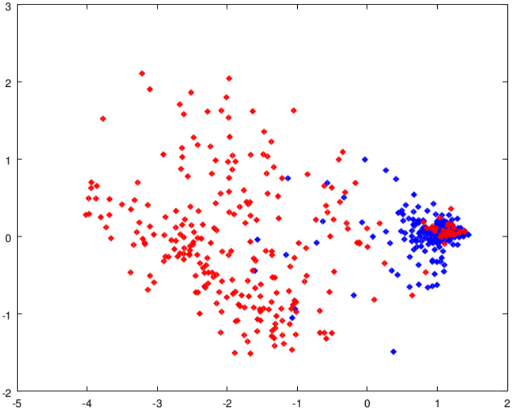


Visualizing Data



Data is very well-clustered...

Visualizing Data



Data is very well-clustered...

But how did we obtain this visualization of our 10-dimensional dataset?

1.5 Dimensionality Reduction

Dimensionality Reduction

Dimensionality Reduction Problem

Dimensionality Reduction

Dimensionality Reduction Problem

- Given dataset $\mathbf{x} = (x_i)_{i=1}^{m} \subset \mathbb{R}^d$ with d large.

Dimensionality Reduction

Dimensionality Reduction Problem

- Given dataset $\mathbf{x} = (x_i)_{i=1}^m \subset \mathbb{R}^d$ with d large.
- Goal: Construct map $\Phi : \mathbb{R}^d \to \mathbb{R}^s$ with $s << d$ such that the features of the dataset $\mathbf{x}$ are preserved under the mapping Φ.

Dimensionality Reduction

Dimensionality Reduction Problem

- Given dataset $\mathbf{x} = (x_i)_{i=1}^m \subset \mathbb{R}^d$ with d large.
- Goal: Construct map $\Phi : \mathbb{R}^d \to \mathbb{R}^s$ with $s << d$ such that the features of the dataset $\mathbf{x}$ are preserved under the mapping Φ.
- Useful for reduction in computational complexity, visualization (if $s \leq 3$) or de-noising/compression.

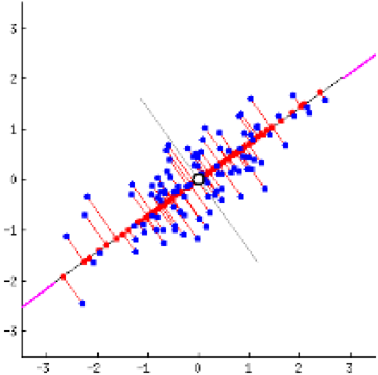
Dimensionality Reduction

Dimensionality Reduction Problem

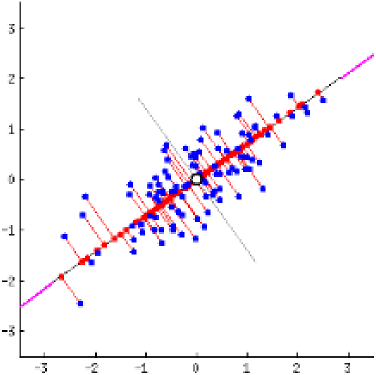
- Given dataset $\mathbf{x} = (x_i)_{i=1}^m \subset \mathbb{R}^d$ with d large.
- Goal: Construct map $\Phi : \mathbb{R}^d \to \mathbb{R}^s$ with $s << d$ such that the features of the dataset $\mathbf{x}$ are preserved under the mapping Φ.
- Useful for reduction in computational complexity, visualization (if $s \leq 3$) or de-noising/compression.

Simplest case: Φ is orthogonal projection onto affine subspace $\rightsquigarrow$ PCA.

What is a good projection?



What is a good projection?



Pick subspace which maximizes variance of the projected dataset.

PCA

PCA Problem

Look for s-dimensional affine subspace with associated orthogonal projection Φ such that the variance $\sum_{i=1}^{m} |\Phi(x_i - \frac{1}{m}\sum_{j=1}^{m} x_j)|^2$ is maximized.

PCA

PCA Problem

Look for s-dimensional affine subspace with associated orthogonal projection Φ such that the variance $\sum_{i=1}^{m} |\Phi(x_i - \frac{1}{m}\sum_{j=1}^{m} x_j)|^2$ is maximized.

PCA Solution

PCA

PCA Problem

Look for s-dimensional affine subspace with associated orthogonal projection Φ such that the variance $\sum_{i=1}^{m} |\Phi(x_i - \frac{1}{m}\sum_{j=1}^{m} x_j)|^2$ is maximized.

PCA Solution

- Suppose w.l.o.g. (why?) that the data is centered, i.e., $\frac{1}{m}\sum_{j=1}^{m} x_j = 0$.

PCA

PCA Problem

Look for s-dimensional affine subspace with associated orthogonal projection Φ such that the variance $\sum_{i=1}^{m} |\Phi(x_i - \frac{1}{m}\sum_{j=1}^{m} x_j)|^2$ is maximized.

PCA Solution

- Suppose w.l.o.g. (why?) that the data is centered, i.e., $\frac{1}{m}\sum_{j=1}^{m} x_j = 0$.
- Define the empirical covariance matrix $G = \frac{1}{m}\sum_{i=1}^{m} x_i x_i^T \in \mathbb{R}^{d\times d}$.

PCA

PCA Problem

Look for s-dimensional affine subspace with associated orthogonal projection Φ such that the variance $\sum_{i=1}^{m} |\Phi(x_i - \frac{1}{m}\sum_{j=1}^{m} x_j)|^2$ is maximized.

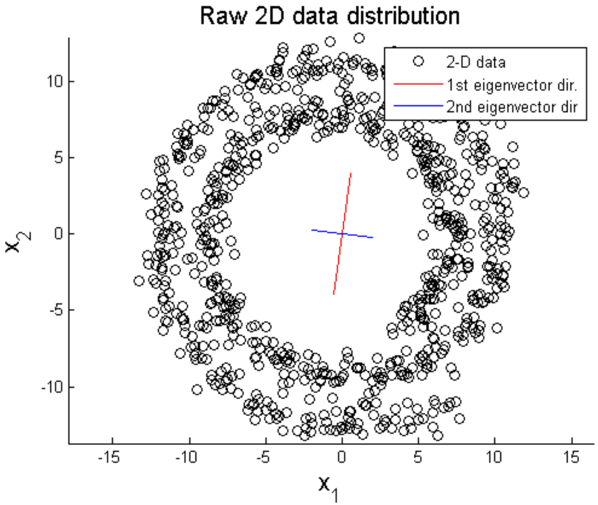
PCA Solution

- Suppose w.l.o.g. (why?) that the data is centered, i.e., $\frac{1}{m}\sum_{j=1}^{m} x_j = 0$.
- Define the empirical covariance matrix $G = \frac{1}{m}\sum_{i=1}^{m} x_i x_i^T \in \mathbb{R}^{d\times d}$.
- Then the solution is given by the subspace spanned by the first s normalized Eigenvectors $u_1, \ldots, u_s$ of G and $\Phi(x) = \sum_{l=1}^{s}(x \cdot u_l)u_l$.

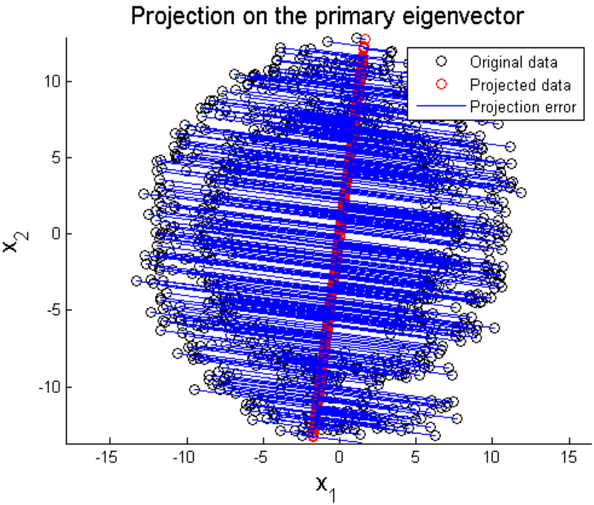
It's really simple

```
function z=pca(X)
%project data X on its
%principal components.
C=cov(X);
[U,D,pc] = svd(C);
z = center(X)*pc;
scatter(z(:,1),z(:,2));%plot 2d projection
```

When PCA fails...



When PCA fails...



Kernel PCA

Construct nonlinear Φ by applying linear PCA on RKHS $\mathcal{H}_K$ by mapping data points x_i to their feature vectors K_{x_i}!

Kernel PCA

Construct nonlinear Φ by applying linear PCA on RKHS $\mathcal{H}_K$ by mapping data points x_i to their feature vectors K_{x_i}!

This is a powerful idea, called *kernelization*!

Kernel PCA

Construct nonlinear Φ by applying linear PCA on RKHS $\mathcal{H}_K$ by mapping data points x_i to their feature vectors K_{x_i}!

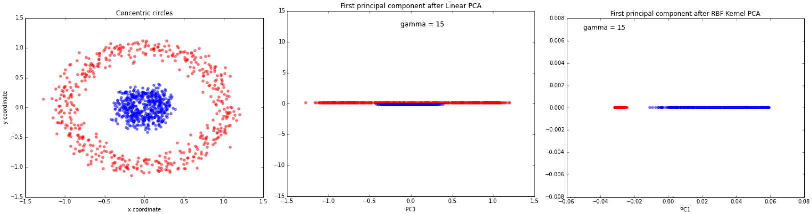
This is a powerful idea, called *kernelization*!

Kernel PCA

Define the matrix
$G = K[\mathbf{x}] - \mathbf{1}_m K[\mathbf{x}] - K[\mathbf{x}]\mathbf{1}_m + \mathbf{1}_m K[\mathbf{x}]\mathbf{1}_m \in \mathbb{R}^{m \times m}$ and $(\mathbf{1}_m)_{i,j} = \frac{1}{m}$ for $i, j \in \{1, \ldots, m\}$ and denote $u_1, \ldots, u_s$ the first s normalized (w.r.t. the inner product $u^T K[\mathbf{x}]u$) Eigenvectors of G. Then the projection Φ is defined as

$$\Phi(x) = \left(\sum_{i=1}^{m} (u_1)_i K(x_i, x), \ldots, \sum_{i=1}^{m} (u_s)_i K(x_i, x) \right)^T .$$

Example



Concentric circles
y coordinate
x coordinate
First principal component after Linear PCA
gamma = 15
PC1
First principal component after RBF Kernel PCA
gamma = 15
PC1

Kernel PCA Denoising

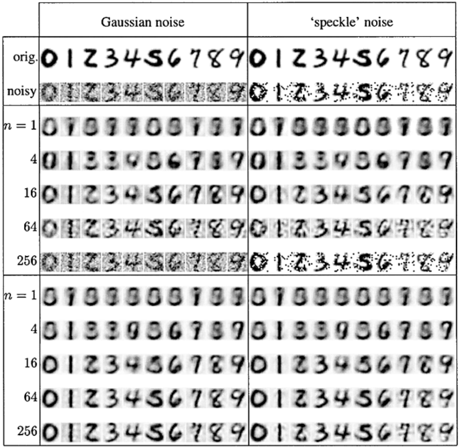


Figure: Top: Linear PCA reconstruction from n principal components. Bottom: Gaussian Kernel Reconstruction from n principal components (find z with $\|K_z - \Phi(x)\|_{\mathcal{H}_K}$ minimal).

- Literature: Sebastian Mika, Bernhard Schölkopf, Alex Smola, Klaus-Robert Müller, Matthias Scholz, Gunnar Rätsch. *Kernel PCA and De-Noising in Feature Space*. NIPS (1999).

- Literature: Sebastian Mika, Bernhard Schölkopf, Alex Smola, Klaus-Robert Müller, Matthias Scholz, Gunnar Rätsch. *Kernel PCA and De-Noising in Feature Space*. NIPS (1999).
- Other methods: Multidimensional Scaling, Isomap, Diffusion Maps, ...

- Literature: Sebastian Mika, Bernhard Schölkopf, Alex Smola, Klaus-Robert Müller, Matthias Scholz, Gunnar Rätsch. *Kernel PCA and De-Noising in Feature Space*. NIPS (1999).
- Other methods: Multidimensional Scaling, Isomap, Diffusion Maps, ...
- Try to appreciate the power of kernelization!

- Literature: Sebastian Mika, Bernhard Schölkopf, Alex Smola, Klaus-Robert Müller, Matthias Scholz, Gunnar Rätsch. *Kernel PCA and De-Noising in Feature Space*. NIPS (1999).
- Other methods: Multidimensional Scaling, Isomap, Diffusion Maps, ...
- Try to appreciate the power of kernelization!
- Go to `https://archive.ics.uci.edu/ml/datasets.html` for further datasets and play around with them!

1.6 (Kernel) Support Vector Machine (SVM)

Basic Idea

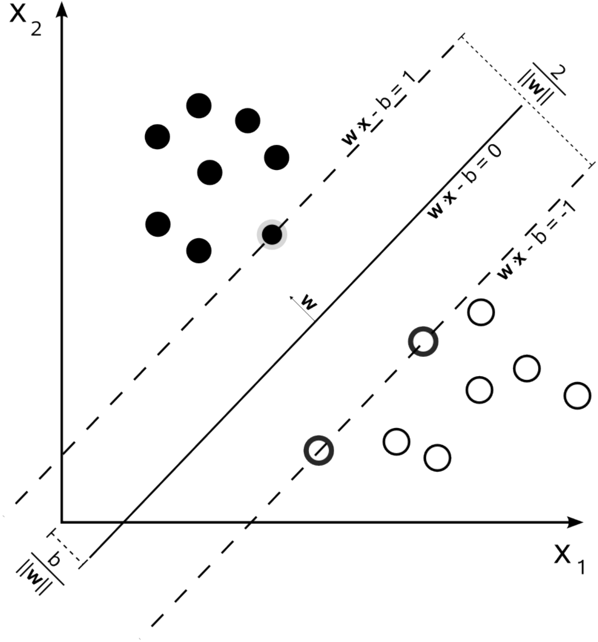
- Suppose that data points $(x_i)_{i=1}^m \subset \mathbb{R}^n$ to be classified are *linearly separable*, i.e. there exists a separating hyperplane defined by $w \in \mathbb{R}^n$, $|w| = 1$ and $b \in \mathbb{R}$ such that

$$y_i = 1 \Leftrightarrow w \cdot x_i > b.$$

- Define the *margin* of a separating hyperplane defined by w, b as above by

$$\Delta(w, b) := \min_{i=1}^{m} |w \cdot x_i - b|.$$

Try to find separating hyperplane with maximal margin!



X 2
X 1
w·x - b = 1
w·x - b = 0
w·x - b = -1
w
2/||w||
b/||w||

The SVM

The SVM problem can be formalized by the following minimization problem

$$\operatorname{argmin}_{w,b} |w| \quad \text{s.t.} \quad y_i(w \cdot x_i - b) \geq 1 \quad \text{for all} \quad i \in \{1, \ldots, m\}.$$

The SVM

The SVM problem can be formalized by the following minimization problem

$$\operatorname{argmin}_{w,b} |w| \quad \text{s.t.} \quad y_i(w \cdot x_i - b) \geq 1 \quad \text{for all} \quad i \in \{1, \ldots, m\}.$$

🙁 data is in general not linearly separable...

The SVM

The SVM problem can be formalized by the following minimization problem

$$\operatorname{argmin}_{w,b} |w| \quad \text{s.t.} \quad y_i(w \cdot x_i - b) \geq 1 \quad \text{for all} \quad i \in \{1, \ldots, m\}.$$

🙁 data is in general not linearly separable...

Soft Margin SVM

Relax to

$$\operatorname{argmin}_{w,b} \frac{1}{m} \sum_{i=1}^{m} \Phi_{hl}(y_i(w \cdot x_i - b)) + \lambda |w|^2,$$

where $\Phi_{hl}(t) := \max(0, 1-t)$, the *hinge loss*.

The SVM

The SVM problem can be formalized by the following minimization problem

$$\operatorname{argmin}_{w,b} |w| \quad \text{s.t.} \quad y_i(w \cdot x_i - b) \geq 1 \quad \text{for all} \quad i \in \{1, \ldots, m\}.$$

data is in general not linearly separable...

Soft Margin SVM

Relax to

$$\operatorname{argmin}_{w,b} \frac{1}{m} \sum_{i=1}^{m} \Phi_{hl}(y_i(w \cdot x_i - b)) + \lambda |w|^2,$$

where $\Phi_{hl}(t) := \max(0, 1 - t)$, the *hinge loss*.

 This is a convex quadratic program that can be efficiently solved!

K-SVM

K-SVM

Kernelization yields the problem

$$\operatorname{argmin}_{f\in\mathcal{H}_K,b}\frac{1}{m}\sum_{i=1}^{m}\Phi_{hl}(y_i(f(x_i)-b))+\lambda\|f\|_{\mathcal{H}_K}^2$$

K-SVM

K-SVM

Kernelization yields the problem

$$\mathrm{argmin}_{f\in\mathcal{H}_K,b}\,\frac{1}{m}\sum_{i=1}^{m}\Phi_{hl}(y_i(f(x_i)-b))+\lambda\|f\|^2_{\mathcal{H}_K}$$

Separable Measures

This provably works for separable measures ρ in the sense that there is $f_s\in\mathcal{H}_K$ with $yf(x)>0$ almost surely. It means that data is separated by the zero level set of f_s. Clearly, the Bayes classifier is that equal to $\mathrm{sgn}(f_s)$.

K-SVM

K-SVM

Kernelization yields the problem

$$\text{argmin}_{f\in\mathcal{H}_K,b}\frac{1}{m}\sum_{i=1}^{m}\Phi_{hl}(y_i(f(x_i)-b))+\lambda\|f\|_{\mathcal{H}_K}^2$$

Separable Measures

This provably works for separable measures ρ in the sense that there is $f_s\in\mathcal{H}_K$ with $yf(x)>0$ almost surely. It means that data is separated by the zero level set of f_s. Clearly, the Bayes classifier is that equal to sgn(f_s).

 For most data points the hinge loss will be zero which implies that f will be sparse in $\{K_{x_i}: i=1,\ldots,m\}$, resulting in potentially big computational savings!

- Literature: Felipe Cucker and Ding Xuan Zhou. *Learning Theory: An Approximation Theory Viewpoint*. Chapter 9.

- Literature: Felipe Cucker and Ding Xuan Zhou. *Learning Theory: An Approximation Theory Viewpoint*. Chapter 9.
- Main advantage as compared to classification method in Section 1.4 is that solution will be sparse.

- Literature: Felipe Cucker and Ding Xuan Zhou. *Learning Theory: An Approximation Theory Viewpoint*. Chapter 9.
- Main advantage as compared to classification method in Section 1.4 is that solution will be sparse.
- Experiment and Compare!

Further Userful Methods

- Gradient Boosted Trees
- Independent Component Analysis