Overfitting and control of the complexity

Guillaume Obozinski

Ecole des Ponts - ParisTech

SOCN course 2014

Outline

Risk of a predictor and PAC learning

Risk of a predictor and PAC learning

Assume now that the predictor is generated from training data D_n according to the scheme:

$$\begin{array}{rccc} \mathscr{A}: & \bigcup_{n\in\mathcal{N}} (\mathcal{X}\times\mathcal{Y})^n & \to & \mathcal{A}^{\mathcal{X}} \\ & D_n & \mapsto & \widehat{f} \end{array}$$

Risk of a predictor and PAC learning

Assume now that the predictor is generated from training data D_n according to the scheme:

$$\begin{array}{rccc} \mathscr{A}: & \bigcup_{n\in\mathcal{N}} (\mathcal{X}\times\mathcal{Y})^n & \to & \mathcal{A}^{\mathcal{X}} \\ & D_n & \mapsto & \widehat{f} \end{array}$$

As a consequence $\mathcal{R}(\widehat{f}) - \mathcal{R}(f^*)$ is a random variable.

Risk of a predictor and PAC learning

Assume now that the predictor is generated from training data D_n according to the scheme:

$$\begin{array}{rccc} \mathscr{A}: & \bigcup_{n\in\mathcal{N}} (\mathcal{X}\times\mathcal{Y})^n & \to & \mathcal{A}^{\mathcal{X}} \\ & D_n & \mapsto & \widehat{f} \end{array}$$

As a consequence $\mathcal{R}(\widehat{f}) - \mathcal{R}(f^*)$ is a random variable.

Expected Risk

$$\mathbb{E}\big[\mathcal{R}(\widehat{f})\big] - \mathcal{R}(f^*)$$

Risk of a predictor and PAC learning

Assume now that the predictor is generated from training data D_n according to the scheme:

$$\begin{array}{rccc} \mathscr{A}: & \bigcup_{n\in\mathcal{N}} (\mathcal{X}\times\mathcal{Y})^n & \to & \mathcal{A}^{\mathcal{X}} \\ & D_n & \mapsto & \widehat{f} \end{array}$$

As a consequence $\mathcal{R}(\widehat{f}) - \mathcal{R}(f^*)$ is a random variable.

Expected Risk

$$\mathbb{E}\left[\mathcal{R}(\widehat{f})\right] - \mathcal{R}(f^*)$$

Probably **A**pproximately **C**orrect Learning

Do approximately as well as the target function with very high probability

$$\mathbb{P}\Big(\mathcal{R}(\widehat{f}) - \mathcal{R}(f^*) \le \epsilon\Big) \ge 1 - \delta$$

Risk of a predictor and PAC learning

Assume now that the predictor is generated from training data D_n according to the scheme:

$$\begin{array}{rccc} \mathscr{A}: & \bigcup_{n\in\mathcal{N}} (\mathcal{X}\times\mathcal{Y})^n & \to & \mathcal{A}^{\mathcal{X}} \\ & D_n & \mapsto & \widehat{f} \end{array}$$

As a consequence $\mathcal{R}(\widehat{f}) - \mathcal{R}(f^*)$ is a random variable.

Expected Risk

$$\mathbb{E}\big[\mathcal{R}(\widehat{f})\big] - \mathcal{R}(f^*)$$

Probably **A**pproximately **C**orrect Learning

Do approximately as well as the target function with very high probability

$$\mathbb{P}\Big(\mathcal{R}(\widehat{f}) - \mathcal{R}(f^*) \leq \epsilon\Big) \geq 1 - \delta$$

$\rightarrow$ Control the convergence in probability of the excess risk.

Back to learning: facing the curse of dimensionality

Back to learning: facing the curse of dimensionality

So far we

- characterized good predictors

Back to learning: facing the curse of dimensionality

So far we

- characterized good predictors
- specified how to assess theoretically the quality of a learning scheme

Back to learning: facing the curse of dimensionality

So far we

- characterized good predictors
- specified how to assess theoretically the quality of a learning scheme

$\rightarrow$ But both rely on the unknown risk $\mathcal{R}$!

Back to learning: facing the curse of dimensionality

So far we

- characterized good predictors
- specified how to assess theoretically the quality of a learning scheme

$\rightarrow$ But both rely on the unknown risk $\mathcal{R}$!

- $\mathcal{R}$ can only be computed if we know $P_{X,Y}$

Back to learning: facing the curse of dimensionality

So far we

- characterized good predictors
- specified how to assess theoretically the quality of a learning scheme

$\rightarrow$ But both rely on the unknown risk $\mathcal{R}$!

- $\mathcal{R}$ can only be computed if we know $P_{X,Y}$

$\rightarrow$ Can we estimate/learn $P_{X,Y}$ from the training data?

Back to learning: facing the curse of dimensionality

So far we

- characterized good predictors
- specified how to assess theoretically the quality of a learning scheme

$\rightarrow$ But both rely on the unknown risk $\mathcal{R}$!

- $\mathcal{R}$ can only be computed if we know $P_{X,Y}$

$\rightarrow$ Can we estimate/learn $P_{X,Y}$ from the training data?

Problems:

- If $P_{X,Y}$ is characterized by a small number of parameters

Back to learning: facing the curse of dimensionality

So far we

- characterized good predictors
- specified how to assess theoretically the quality of a learning scheme

$\rightarrow$ But both rely on the unknown risk $\mathcal{R}$!

- $\mathcal{R}$ can only be computed if we know $P_{X,Y}$

$\rightarrow$ Can we estimate/learn $P_{X,Y}$ from the training data?

Problems:

- If $P_{X,Y}$ is characterized by a small number of parameters

$\rightarrow$ Possible to estimate $\rightarrow$ approach similar to classical statistics

Back to learning: facing the curse of dimensionality

So far we

- characterized good predictors
- specified how to assess theoretically the quality of a learning scheme

$\rightarrow$ But both rely on the unknown risk $\mathcal{R}$!

- $\mathcal{R}$ can only be computed if we know $P_{X,Y}$

$\rightarrow$ Can we estimate/learn $P_{X,Y}$ from the training data?

Problems:

- If $P_{X,Y}$ is characterized by a small number of parameters

$\rightarrow$ Possible to estimate $\rightarrow$ approach similar to classical statistics

- Learning $P_{X,Y}$ is often a more complicated than learning f !!

Back to learning: facing the curse of dimensionality

So far we

- characterized good predictors
- specified how to assess theoretically the quality of a learning scheme

$\rightarrow$ But both rely on the unknown risk $\mathcal{R}$!

- $\mathcal{R}$ can only be computed if we know $P_{X,Y}$

$\rightarrow$ Can we estimate/learn $P_{X,Y}$ from the training data?

Problems:

- If $P_{X,Y}$ is characterized by a small number of parameters

$\rightarrow$ Possible to estimate $\rightarrow$ approach similar to classical statistics

- Learning $P_{X,Y}$ is often a more complicated than learning f !!

$\rightarrow$ We should not try and solve a more complicated problem than the initial learning problem.

Back to learning: facing the curse of dimensionality

So far we

- characterized good predictors
- specified how to assess theoretically the quality of a learning scheme

$\rightarrow$ But both rely on the unknown risk $\mathcal{R}$!

- $\mathcal{R}$ can only be computed if we know $P_{X,Y}$

$\rightarrow$ Can we estimate/learn $P_{X,Y}$ from the training data?

Problems:

- If $P_{X,Y}$ is characterized by a small number of parameters

$\rightarrow$ Possible to estimate $\rightarrow$ approach similar to classical statistics

- Learning $P_{X,Y}$ is often a more complicated than learning f !!

$\rightarrow$ We should not try and solve a more complicated problem than the initial learning problem.

- $\mathcal{X} \times \mathcal{Y}$ is typically a high dimensional space.

$\rightarrow$ Density estimation requires an amount of data which grows exponentially with the number of dimensions

Back to learning: facing the curse of dimensionality

So far we

- characterized good predictors
- specified how to assess theoretically the quality of a learning scheme

$\rightarrow$ But both rely on the unknown risk $\mathcal{R}$!

- $\mathcal{R}$ can only be computed if we know $P_{X,Y}$

$\rightarrow$ Can we estimate/learn $P_{X,Y}$ from the training data?

Problems:

- If $P_{X,Y}$ is characterized by a small number of parameters

$\rightarrow$ Possible to estimate $\rightarrow$ approach similar to classical statistics

- Learning $P_{X,Y}$ is often a more complicated than learning f !!

$\rightarrow$ We should not try and solve a more complicated problem than the initial learning problem.

- $\mathcal{X} \times \mathcal{Y}$ is typically a high dimensional space.

$\rightarrow$ Density estimation requires an amount of data which grows exponentially with the number of dimensions

This is the **Curse of dimensionality**

Curse of dimensionality

Exponential grow of "volume" with dimensions

Histograms

Construct a histogram for $X \in [0, 1]$ with 10 bins

Curse of dimensionality

Exponential grow of "volume" with dimensions

Histograms

Construct a histogram for $X \in [0, 1]$ with 10 bins

$\rightarrow$ possible with 100 observations

Curse of dimensionality

Exponential grow of “volume” with dimensions

Histograms

Construct a histogram for $X \in [0, 1]$ with 10 bins

$\rightarrow$ possible with 100 observations

Construct a histogram for $X \in [0, 1]^{10}$

Curse of dimensionality

Exponential grow of "volume" with dimensions

Histograms

Construct a histogram for $X \in [0, 1]$ with 10 bins

$\rightarrow$ possible with 100 observations

Construct a histogram for $X \in [0, 1]^{10}$

$\rightarrow$ size et number of bin ?

Curse of dimensionality

Exponential grow of "volume" with dimensions

Histograms

Construct a histogram for $X \in [0,1]$ with 10 bins

$\rightarrow$ possible with 100 observations

Construct a histogram for $X \in [0,1]^{10}$

$\rightarrow$ size et number of bin ?

$\rightarrow$ a priori impossible with 100 or even with 10^6 observations !

Curse of dimensionality

Exponential grow of "volume" with dimensions

Histograms

Construct a histogram for $X \in [0,1]$ with 10 bins

$\rightarrow$ possible with 100 observations

Construct a histogram for $X \in [0,1]^{10}$

$\rightarrow$ size et number of bin ?

$\rightarrow$ a priori impossible with 100 or even with 10^6 observations !

Curse of dimensionality

Exponential grow of “volume” with dimensions

Histograms

Construct a histogram for $X \in [0,1]$ with 10 bins

$\rightarrow$ possible with 100 observations

Construct a histogram for $X \in [0,1]^{10}$

$\rightarrow$ size et number of bin ?

$\rightarrow$ a priori impossible with 100 or even with 10^6 observations !

Model for SNPs

SNP: Single-Nucleotide Polymorphism

- Correspond to 90% of human genetic variation

Curse of dimensionality

Exponential grow of “volume” with dimensions

Histograms

Construct a histogram for $X \in [0,1]$ with 10 bins

$\rightarrow$ possible with 100 observations

Construct a histogram for $X \in [0,1]^{10}$

$\rightarrow$ size et number of bin ?

$\rightarrow$ a priori impossible with 100 or even with 10^6 observations !

Model for SNPs

SNP: Single-Nucleotide Polymorphism

- Correspond to 90% of human genetic variation
- Number of loci $k > 10^5$

Curse of dimensionality

Exponential grow of "volume" with dimensions

Histograms

Construct a histogram for $X \in [0,1]$ with 10 bins

$\rightarrow$ possible with 100 observations

Construct a histogram for $X \in [0,1]^{10}$

$\rightarrow$ size et number of bin ?

$\rightarrow$ a priori impossible with 100 or even with 10^6 observations !

Model for SNPs

SNP: Single-Nucleotide Polymorphism

- Correspond to 90% of human genetic variation
- Number of loci $k > 10^5$
- Number of configurations $> 2^{10^5}$...

Curse of dimensionality

Exponential grow of "volume" with dimensions

Histograms

Construct a histogram for $X \in [0, 1]$ with 10 bins

$\rightarrow$ possible with 100 observations

Construct a histogram for $X \in [0, 1]^{10}$

$\rightarrow$ size et number of bin ?

$\rightarrow$ a priori impossible with 100 or even with 10^6 observations !

Model for SNPs

SNP: Single-Nucleotide Polymorphism

- Correspond to 90% of human genetic variation
- Number of loci $k > 10^5$
- Number of configurations $> 2^{10^5}$...

Outline

1. Empirical Risk Minimization
2. Polynomial regression and overfitting
3. Regularization
4. Complexity

Empirical Risk Minimization

Empirical Risk Minimization

Idea: Replace the population distribution of the data by the empirical distribution of the training data. Given a training set $\{(x_1, y_1), \ldots, (x_n, y_n)\}$, we define the

Empirical Risk Minimization

Idea: Replace the population distribution of the data by the empirical distribution of the training data. Given a training set $\{(x_1, y_1), \ldots, (x_n, y_n)\}$, we define the

Empirical Risk

$$\widehat{\mathcal{R}}_n(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(f(x_i), y_i)$$

Empirical Risk Minimization

Idea: Replace the population distribution of the data by the empirical distribution of the training data. Given a training set $\{(x_1, y_1), \ldots, (x_n, y_n)\}$, we define the

Empirical Risk

$$\widehat{\mathcal{R}}_n(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(f(x_i), y_i)$$

Empirical Risk Minimization principle

- consists in minimizing the empirical risk.

Empirical Risk Minimization

Idea: Replace the population distribution of the data by the empirical distribution of the training data. Given a training set $\{(x_1, y_1), \dots, (x_n, y_n)\}$, we define the

Empirical Risk

$$\widehat{\mathcal{R}}_n(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(f(x_i), y_i)$$

Empirical Risk Minimization principle

- consists in minimizing the empirical risk.

Problem: The target function for the empirical risk is only defined at the training points.

Learning as an ill-posed problem

A problem is **well-posed** in the sense of Hadamard if

- It admits a solution

Learning as an ill-posed problem

A problem is **well-posed** in the sense of Hadamard if

- It admits a solution
- This solution is *unique*

Learning as an ill-posed problem

A problem is **well-posed** in the sense of Hadamard if

- It admits a solution
- This solution is *unique*
- The solution depends continuously on the problem parameters for an appropriate topology.

Learning as an ill-posed problem

A problem is **well-posed** in the sense of Hadamard if

- It admits a solution
- This solution is *unique*
- The solution depends continuously on the problem parameters for an appropriate topology.

Learning as formulated is

- underconstrained
- with by essence incomplete information

and thus ill-posed.

Learning as an ill-posed problem

A problem is **well-posed** in the sense of Hadamard if

- It admits a solution
- This solution is *unique*
- The solution depends continuously on the problem parameters for an appropriate topology.

Learning as formulated is

- underconstrained
- with by essence incomplete information

and thus ill-posed.
Introduce an *inductive bias* by restricting the **hypothesis space** and/or using **regularization**.

Hypothesis space

For both computational and statistical reasons, it is necessary to consider to restrict the set of predictors or the set of hypotheses considered. Given a hypothesis space $S \subset \mathcal{Y}^{\mathcal{X}}$ considered the constrained ERM problem

$$\min_{f \in S} \widehat{\mathcal{R}}_n(f)$$

- linear functions
- polynomial functions
- spline functions
- multiresolution approximation spaces (wavelet)

Outline

Polynomial regression and overfitting

Polynomial regression: an instance of linear regression

Model of the form $Y = w_0 + w_1 X + w_2 X^2 + \ldots + w_p X^p + \varepsilon$

Polynomial regression: an instance of linear regression

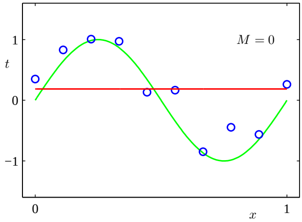
Model of the form $Y = w_0 + w_1 X + w_2 X^2 + \ldots + w_p X^p + \varepsilon$

$$\min_{\mathbf{w}} \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - (w_0 + w_1 x_i + w_2 x_i^2 + \ldots + w_p x_i^p) \right)^2$$

Polynomial regression: an instance of linear regression

Model of the form $Y = w_0 + w_1 X + w_2 X^2 + \ldots + w_p X^p + \varepsilon$

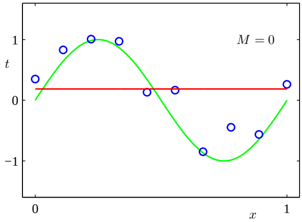
$$\min_{\mathbf{w}} \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - (w_0 + w_1 x_i + w_2 x_i^2 + \ldots + w_p x_i^p)\right)^2$$

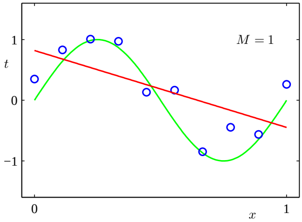


Polynomial regression: an instance of linear regression

Model of the form $Y = w_0 + w_1 X + w_2 X^2 + \ldots + w_p X^p + \varepsilon$

$$\min_{\mathbf{w}} \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - \left(w_0 + w_1 x_i + w_2 x_i^2 + \ldots + w_p x_i^p \right) \right)^2$$

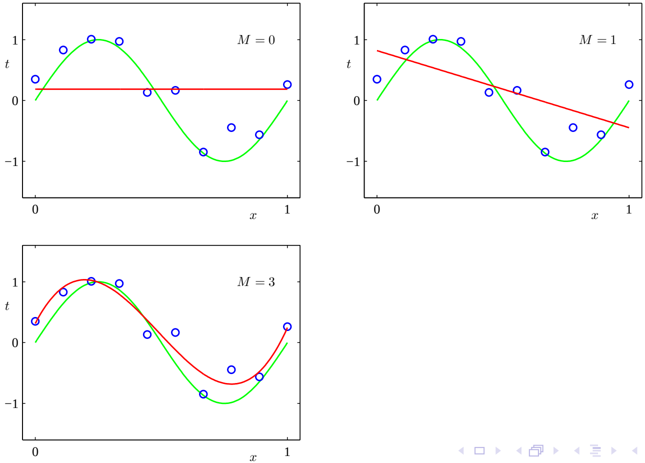




Polynomial regression: an instance of linear regression

Model of the form $Y = w_0 + w_1 X + w_2 X^2 + \ldots + w_p X^p + \varepsilon$

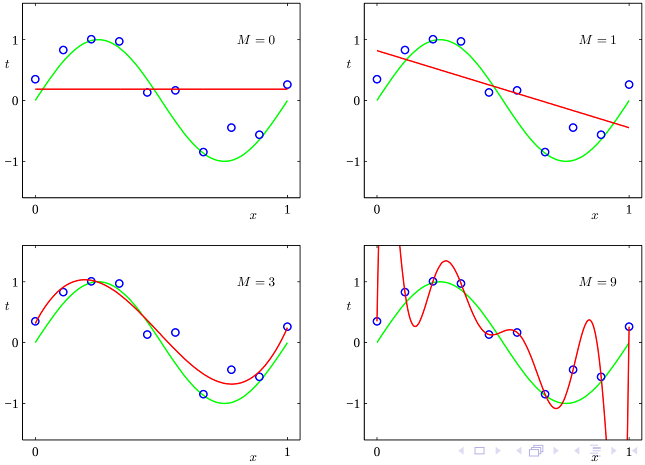
$$\min_{\mathbf{w}} \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - \left(w_0 + w_1 x_i + w_2 x_i^2 + \ldots + w_p x_i^p \right) \right)^2$$



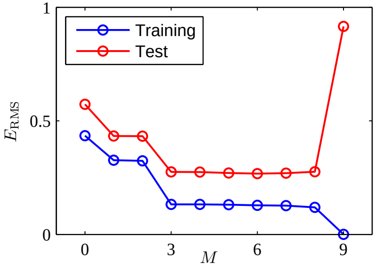
Polynomial regression: an instance of linear regression

Model of the form $Y = w_0 + w_1 X + w_2 X^2 + \ldots + w_p X^p + \varepsilon$

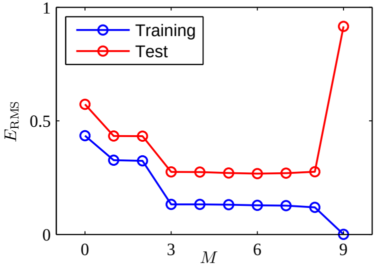
$$\min_{\mathbf{w}} \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - \left(w_0 + w_1 x_i + w_2 x_i^2 + \ldots + w_p x_i^p\right)\right)^2$$

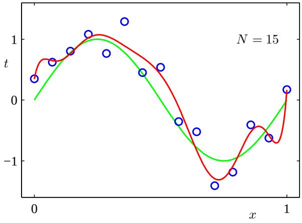


Overfitting: symptoms and characteristics

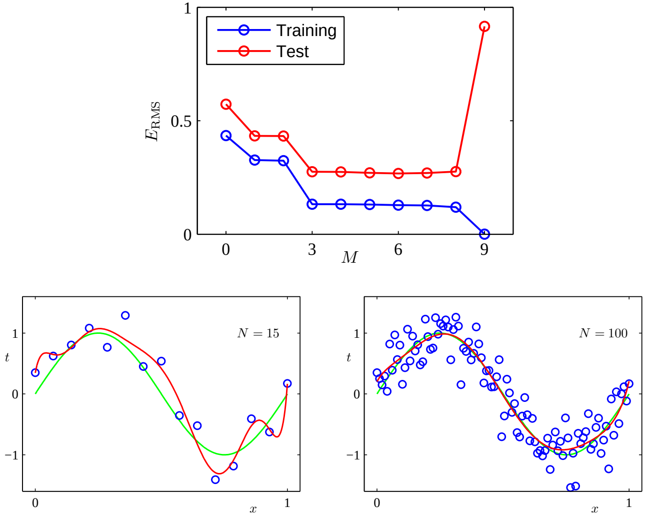


Overfitting: symptoms and characteristics





Overfitting: symptoms and characteristics



Outline

Regularization

Tikhonov regularization

$$\min_{f \in S} \widehat{\mathcal{R}}_n(f) + \lambda \|f\|^2$$

Tikhonov regularization

$$\min_{f \in S} \widehat{\mathcal{R}}_n(f) + \lambda \|f\|^2$$

- λ is the regularization coefficient or hyperparameter

Tikhonov regularization

$$\min_{f \in S} \widehat{\mathcal{R}}_n(f) + \lambda \|f\|^2$$

- λ is the regularization coefficient or hyperparameter

Is the problem now well-posed?

If $\widehat{\mathcal{R}}_n$ is convex

Tikhonov regularization

$$\min_{f \in S} \widehat{\mathcal{R}}_n(f) + \lambda \|f\|^2$$

- λ is the regularization coefficient or hyperparameter

Is the problem now well-posed?

If $\widehat{\mathcal{R}}_n$ is convex

$\Rightarrow$ The objective is strongly convex and coercive for any $\lambda > 0$

Tikhonov regularization

$$\min_{f \in S} \widehat{\mathcal{R}}_n(f) + \lambda \|f\|^2$$

- λ is the regularization coefficient or hyperparameter

Is the problem now well-posed?

If $\widehat{\mathcal{R}}_n$ is convex

$\Rightarrow$ The objective is strongly convex and coercive for any $\lambda > 0$

$\Rightarrow$ The solution exists and is unique.

Tikhonov regularization

$$\min_{f \in S} \widehat{\mathcal{R}}_n(f) + \lambda \|f\|^2$$

- λ is the regularization coefficient or hyperparameter

Is the problem now well-posed?

If $\widehat{\mathcal{R}}_n$ is convex

$\Rightarrow$ The objective is strongly convex and coercive for any $\lambda > 0$

$\Rightarrow$ The solution exists and is unique.

$\Rightarrow$ $\lambda \mapsto \widehat{f}_\lambda$ is a continuous function

Tikhonov regularization

$$\min_{f \in S} \widehat{\mathcal{R}}_n(f) + \lambda \|f\|^2$$

- λ is the regularization coefficient or hyperparameter

Is the problem now well-posed?

If $\widehat{\mathcal{R}}_n$ is convex

$\Rightarrow$ The objective is strongly convex and coercive for any $\lambda > 0$

$\Rightarrow$ The solution exists and is unique.

$\Rightarrow$ $\lambda \mapsto \widehat{f}_\lambda$ is a continuous function

If $\widehat{\mathcal{R}}_n$ is bounded below

Tikhonov regularization

$$\min_{f \in S} \widehat{\mathcal{R}}_n(f) + \lambda \|f\|^2$$

- λ is the regularization coefficient or hyperparameter

Is the problem now well-posed?

If $\widehat{\mathcal{R}}_n$ is convex

$\Rightarrow$ The objective is strongly convex and coercive for any $\lambda > 0$

$\Rightarrow$ The solution exists and is unique.

$\Rightarrow$ $\lambda \mapsto \widehat{f}_\lambda$ is a continuous function

If $\widehat{\mathcal{R}}_n$ is bounded below

$\Rightarrow$ The objective is coercive for any $\lambda > 0$

Tikhonov regularization

$$\min_{f \in S} \widehat{\mathcal{R}}_n(f) + \lambda \|f\|^2$$

- λ is the regularization coefficient or hyperparameter

Is the problem now well-posed?

If $\widehat{\mathcal{R}}_n$ is convex

$\Rightarrow$ The objective is strongly convex and coercive for any $\lambda > 0$

$\Rightarrow$ The solution exists and is unique.

$\Rightarrow$ $\lambda \mapsto \widehat{f}_\lambda$ is a continuous function

If $\widehat{\mathcal{R}}_n$ is bounded below

$\Rightarrow$ The objective is coercive for any $\lambda > 0$

$\Rightarrow$ At least a solution exists

Tikhonov regularization

$$\min_{f \in S} \widehat{\mathcal{R}}_n(f) + \lambda \|f\|^2$$

- λ is the regularization coefficient or hyperparameter

Is the problem now well-posed?

If $\widehat{\mathcal{R}}_n$ is convex

$\Rightarrow$ The objective is strongly convex and coercive for any $\lambda > 0$

$\Rightarrow$ The solution exists and is unique.

$\Rightarrow$ $\lambda \mapsto \widehat{f}_\lambda$ is a continuous function

If $\widehat{\mathcal{R}}_n$ is bounded below

$\Rightarrow$ The objective is coercive for any $\lambda > 0$

$\Rightarrow$ At least a solution exists

If $\widehat{\mathcal{R}}_n$ is $\mathcal{C}^2$ with bounded curvature

Tikhonov regularization

$$\min_{f \in S} \widehat{\mathcal{R}}_n(f) + \lambda \|f\|^2$$

- λ is the regularization coefficient or hyperparameter

Is the problem now well-posed?

If $\widehat{\mathcal{R}}_n$ is convex

$\Rightarrow$ The objective is strongly convex and coercive for any $\lambda > 0$

$\Rightarrow$ The solution exists and is unique.

$\Rightarrow$ $\lambda \mapsto \widehat{f}_\lambda$ is a continuous function

If $\widehat{\mathcal{R}}_n$ is bounded below

$\Rightarrow$ The objective is coercive for any $\lambda > 0$

$\Rightarrow$ At least a solution exists

If $\widehat{\mathcal{R}}_n$ is $\mathcal{C}^2$ with bounded curvature

$\Rightarrow$ Regularization eliminates small local minima.

Tikhonov regularization

$$\min_{f \in S} \widehat{\mathcal{R}}_n(f) + \lambda \|f\|^2$$

- λ is the regularization coefficient or hyperparameter

Is the problem now well-posed?

If $\widehat{\mathcal{R}}_n$ is convex

$\Rightarrow$ The objective is strongly convex and coercive for any $\lambda > 0$

$\Rightarrow$ The solution exists and is unique.

$\Rightarrow$ $\lambda \mapsto \widehat{f}_\lambda$ is a continuous function

If $\widehat{\mathcal{R}}_n$ is bounded below

$\Rightarrow$ The objective is coercive for any $\lambda > 0$

$\Rightarrow$ At least a solution exists

If $\widehat{\mathcal{R}}_n$ is $\mathcal{C}^2$ with bounded curvature

$\Rightarrow$ Regularization eliminates small local minima.

Ridge regression

Is obtained by applying Tikhonov regularization to OLS regression.

$$\min_{\mathbf{w}\in\mathbb{R}^p} \frac{1}{2n}\|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 + \lambda\|\mathbf{w}\|_2^2$$

- Problem now strongly convex thus well-posed

Ridge regression

Is obtained by applying Tikhonov regularization to OLS regression.

$$\min_{\mathbf{w}\in\mathbb{R}^p} \frac{1}{2n}\|\mathbf{y}-\mathbf{X}\mathbf{w}\|_2^2 + \lambda\|\mathbf{w}\|_2^2$$

- Problem now strongly convex thus well-posed
- Thus with unique solution:

$$\hat{\mathbf{w}}^{(\text{ridge})} = (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{y}$$

Ridge regression

Is obtained by applying Tikhonov regularization to OLS regression.

$$\min_{\mathbf{w}\in\mathbb{R}^p} \frac{1}{2n}\|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 + \lambda\|\mathbf{w}\|_2^2$$

- Problem now strongly convex thus well-posed
- Thus with unique solution:

$$\hat{\mathbf{w}}^{(\text{ridge})} = (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{y}$$

- Shrinkage effect

Ridge regression

Is obtained by applying Tikhonov regularization to OLS regression.

$$\min_{\mathbf{w}\in\mathbb{R}^p} \frac{1}{2n}\|\mathbf{y}-\mathbf{X}\mathbf{w}\|_2^2 + \lambda\|\mathbf{w}\|_2^2$$

- Problem now strongly convex thus well-posed
- Thus with unique solution:

$$\hat{\mathbf{w}}^{(\text{ridge})} = (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{y}$$

- Shrinkage effect
- Regularization improves the conditioning number of the Hessian

Ridge regression

Is obtained by applying Tikhonov regularization to OLS regression.

$$\min_{\mathbf{w} \in \mathbb{R}^p} \frac{1}{2n} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 + \lambda \|\mathbf{w}\|_2^2$$

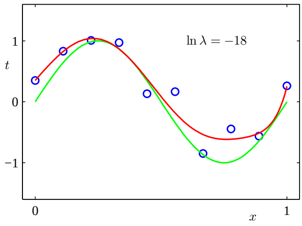
- Problem now strongly convex thus well-posed
- Thus with unique solution:

$$\hat{\mathbf{w}}^{(\text{ridge})} = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^\top \mathbf{y}$$

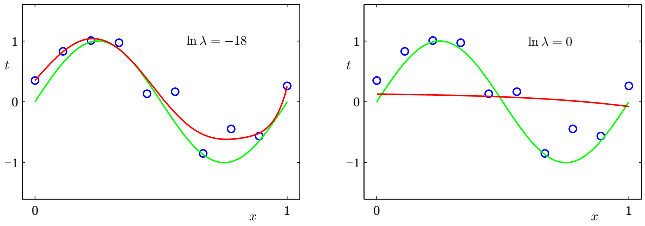
- Shrinkage effect
- Regularization improves the conditioning number of the Hessian

$\Rightarrow$ Problem now easier to solve computationally

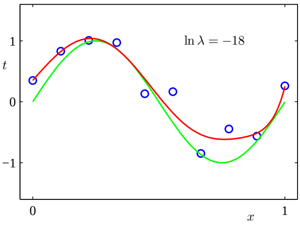
Polynomial regression with ridge

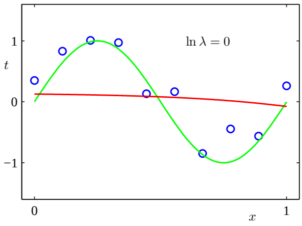


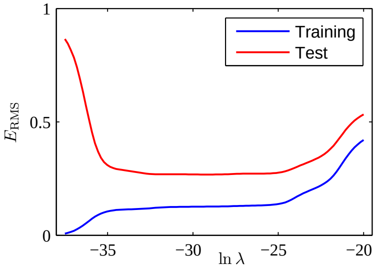
Polynomial regression with ridge



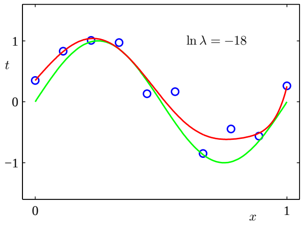
Polynomial regression with ridge

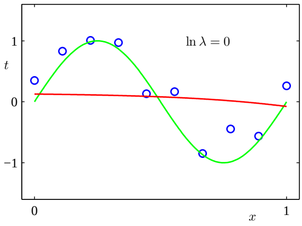


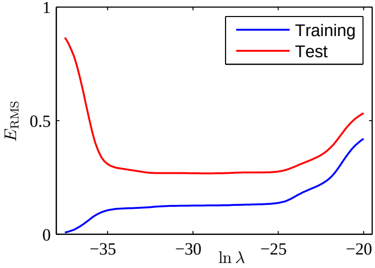




Polynomial regression with ridge







Outline

Complexity

Controlling the complexity of the hypothesis space

Explicit control

- number of variables
- maximal degree for polynomial functions
- degree and number of knots for spline functions
- maximal resolution in wavelet approximations.
- bandwidth in RKHS

The complexity is fixed.

Controlling the complexity of the hypothesis space

Explicit control

- number of variables
- maximal degree for polynomial functions
- degree and number of knots for spline functions
- maximal resolution in wavelet approximations.
- bandwidth in RKHS

The complexity is fixed.

Implicit control with regularization.

Controlling the complexity of the hypothesis space

Explicit control

- number of variables
- maximal degree for polynomial functions
- degree and number of knots for spline functions
- maximal resolution in wavelet approximations.
- bandwidth in RKHS

The complexity is fixed.

Implicit control with regularization.
The complexity of the predictor results from a compromise between fitting and increasing complexity.

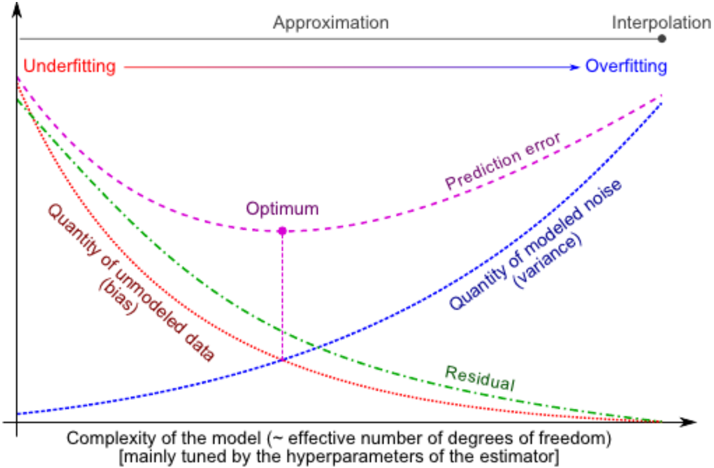
Problem of model selection: How to choose the level of complexity?

Risk decomposition: approximation-estimation trade-off

$$\underbrace{\mathcal{R}(\widehat{f}_S) - \mathcal{R}(f^*)}_{\text{excess risk}} = \underbrace{\mathcal{R}(\widehat{f}_S) - \mathcal{R}(f_S^*)}_{\text{estimation error}} + \underbrace{\mathcal{R}(f_S^*) - \mathcal{R}(f^*)}_{\text{approximation error}}$$

- Sometimes also called “bias-variance tradeoff

Approximation-estimation tradeoff



Bias-variance decomposition of a predictor

$$\mathbb{E}\left[(Z-c)^2\right] = \underbrace{\mathbb{E}\left[(Z-\mathbb{E}[Z])^2\right]}_{\text{variance}} + \underbrace{(\mathbb{E}[Z]-c)^2}_{\text{squared bias}}.$$

Bias-variance decomposition of a predictor

$$\mathbb{E}\left[(Z-c)^2\right] = \underbrace{\mathbb{E}\left[(Z-\mathbb{E}[Z])^2\right]}_{\text{variance}} + \underbrace{(\mathbb{E}[Z]-c)^2}_{\text{squared bias}}.$$

$$\mathbb{E}_{D_n}\left[\left(\widehat{f}(x) - f(x)\right)^2\right] =$$

Bias-variance decomposition of a predictor

$$\mathbb{E}\left[(Z-c)^2\right] = \underbrace{\mathbb{E}\left[(Z-\mathbb{E}[Z])^2\right]}_{\text{variance}} + \underbrace{(\mathbb{E}[Z]-c)^2}_{\text{squared bias}}.$$

$$\mathbb{E}_{D_n}\left[\left(\widehat{f}(x) - f(x)\right)^2\right] = \mathbb{E}_{D_n}\left[\left(\widehat{f}(x) - \mathbb{E}_{D_n}[\widehat{f}(x)]\right)^2\right] + \left(\mathbb{E}_{D_n}[\widehat{f}(x)] - f(x)\right)^2$$

Bias-variance decomposition of a predictor

$$\mathbb{E}\left[(Z-c)^2\right] = \underbrace{\mathbb{E}\left[(Z-\mathbb{E}[Z])^2\right]}_{\text{variance}} + \underbrace{(\mathbb{E}[Z]-c)^2}_{\text{squared bias}}.$$

$$\mathbb{E}_{D_n}\left[\left(\widehat{f}(x)-f(x)\right)^2\right] = \mathbb{E}_{D_n}\left[\left(\widehat{f}(x)-\mathbb{E}_{D_n}[\widehat{f}(x)]\right)^2\right] + \left(\mathbb{E}_{D_n}[\widehat{f}(x)]-f(x)\right)^2$$

$$\begin{aligned} \mathbb{E}\left[\mathcal{E}(\widehat{f})\right] &= \mathbb{E}_{D_n,X}\left[\mathcal{R}(\widehat{f})\right] - \mathcal{R}(f^*) \\ &= \mathbb{E}[(\widehat{f}(X)-f^*(X))^2] \end{aligned}$$

Bias-variance decomposition of a predictor

$$\mathbb{E}\left[(Z-c)^2\right] = \underbrace{\mathbb{E}\left[(Z-\mathbb{E}[Z])^2\right]}_{\text{variance}} + \underbrace{(\mathbb{E}[Z]-c)^2}_{\text{squared bias}}.$$

$$\mathbb{E}_{D_n}\left[\left(\widehat{f}(x)-f(x)\right)^2\right] = \mathbb{E}_{D_n}\left[\left(\widehat{f}(x)-\mathbb{E}_{D_n}[\widehat{f}(x)]\right)^2\right] + \left(\mathbb{E}_{D_n}[\widehat{f}(x)]-f(x)\right)^2$$

$$\begin{aligned}
\mathbb{E}\left[\mathcal{E}(\widehat{f})\right] &= \mathbb{E}_{D_n,X}\left[\mathcal{R}(\widehat{f})\right] - \mathcal{R}(f^*) \\
&= \mathbb{E}[(\widehat{f}(X)-f^*(X))^2] \\
&= \underbrace{\mathbb{E}[(\widehat{f}(X)-\mathbb{E}[\widehat{f}(X)|X])^2]}_{\text{variance of } \widehat{f}} + \underbrace{\mathbb{E}[(\mathbb{E}[\widehat{f}(X)|X]-f^*(X))^2]}_{\text{bias of } \widehat{f}}
\end{aligned}$$

with $f^*(X) = \mathbb{E}[Y|X]$.