Machine Learning and Intelligent Systems

Regularization

Maria A. Zuluaga

Dec 1, 2023

EURECOM - Data Science Department

Table of contents

Motivation

Motivation: Unconstrained Soft SVM

$$\arg\min_{\mathbf{w}, w_0} \underbrace{\frac{1}{2}\|\mathbf{w}\|^2}_{L2-\text{regularizer}} + C \sum_{i=1}^{N} \underbrace{\max(1 - y_i(\mathbf{w}^T \mathbf{x}_i + w_0), 0)}_{\text{hinge loss}}$$

What is the role of the regularizer?

In this part of the lecture:

- We will introduce the concept of regularization
- Go back to the linear regression setup

Linear Regression: The OLS Solution

Let us recall the closed-form solution for a linear regressor

$$\hat{\mathbf{w}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

The inversion of $\mathbf{X}^T\mathbf{X}$ can be problematic when it is poorly conditioned.

This may be often the case in practice, where $D \gg N$ often occurs

Linear Regression: The OLS Solution

Let us recall the closed-form solution for a linear regressor

$$\hat{\mathbf{w}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

The inversion of $\mathbf{X}^T\mathbf{X}$ can be problematic when it is poorly conditioned.

This may be often the case in practice, where $D \gg N$ often occurs

A solution to this is to add a small element in the diagonal:

$$\hat{\mathbf{w}} = (\mathbf{X}^T\mathbf{X} + \lambda^2\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}$$

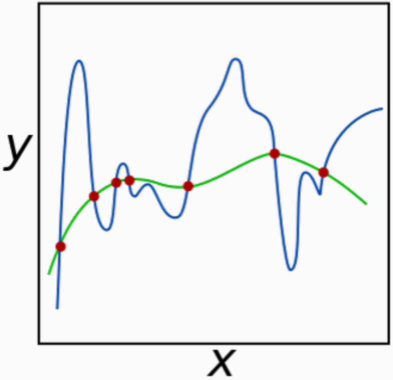
This term is denoted the **ridge regressor estimator**.

It is the solution to a **regularized quadratic cost function**.

Regularization

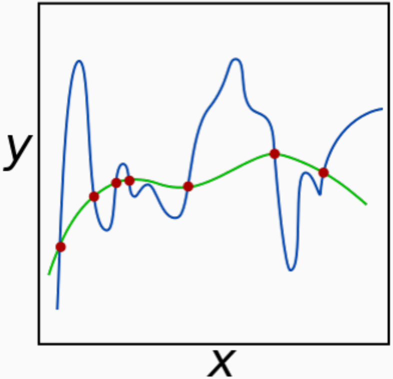
Sensitivity

- Large sensitivity can lead to poor performance of the model, i.e. poor generalization
- Very large values of **w** can make the model very sensitive



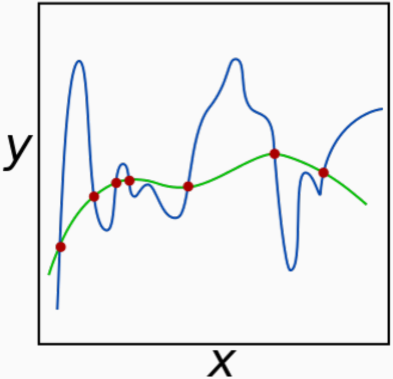
Sensitivity

- Large sensitivity can lead to poor performance of the model, i.e. poor generalization
- Very large values of $\mathbf{w}$ can make the model very sensitive
- **Intuition:** Make $\mathbf{w}_{1:D}$ small



Sensitivity

- Large sensitivity can lead to poor performance of the model, i.e. poor generalization
- Very large values of $\mathbf{w}$ can make the model very sensitive
- **Intuition:** Make $\mathbf{w}_{1:D}$ small
- Regularizers, such as in the case of the ridge regressor, allow to keep the model small



Regularizers

- We will control the size of $\mathbf{w}$ by means of a regularizer function $R : \mathbb{R}^D \to \mathbb{R}$
- $R(\mathbf{w})$ measures the size of $\mathbf{w}$

Regularizers

- We will control the size of $\mathbf{w}$ by means of a regularizer function $R : \mathbb{R}^D \to \mathbb{R}$
- $R(\mathbf{w})$ measures the size of $\mathbf{w}$
- Examples:

$$R(\mathbf{w}) = \sum_{i=1}^{D} w_i^2 \qquad (L2 \text{ regularization})$$

$$R(\mathbf{w}) = \sum_{i=1}^{D} |w_i| \qquad (L1 \text{ regularization})$$

Using Regularization

- Using a regularizer accounts to adding the term to the loss function:

$$\mathcal{L}(\mathbf{w}) + \lambda R(\mathbf{w})$$

- $\lambda \geq 0$ is a hyper-parameter, denoted the regularization parameter

- It controls the strength of the regularization

Example: Ridge Regression

Linear regression objective function:

$$\arg\min_{\mathbf{w}} \mathcal{L}(\mathbf{w}) = \arg\min_{\mathbf{w}} \frac{1}{N} (\mathbf{y} - \mathbf{X}\mathbf{w})^T (\mathbf{y} - \mathbf{X}\mathbf{w})$$

Example: Ridge Regression

Linear regression objective function:

$$\arg\min_{\mathbf{w}} \mathcal{L}(\mathbf{w}) = \arg\min_{\mathbf{w}} \frac{1}{N} (\mathbf{y} - \mathbf{X}\mathbf{w})^T (\mathbf{y} - \mathbf{X}\mathbf{w})$$

If we include L2-regularization the new objective becomes:

$$\arg\min_{\mathbf{w}} \mathcal{L}(\mathbf{w}) + \lambda R(\mathbf{w}) = \arg\min_{\mathbf{w}} \ (\mathbf{y} - \mathbf{X}\mathbf{w})^T (\mathbf{y} - \mathbf{X}\mathbf{w}) + \lambda \mathbf{w}^T \mathbf{w}$$

To find $\hat{\mathbf{w}}$ we proceed in the same way we did with simple linear regression

Ridge Regression - Derivation

$$\frac{\partial \mathcal{L}}{\partial \mathbf{w}} = \frac{\partial}{\partial \mathbf{w}}\left((\mathbf{y} - \mathbf{X}\mathbf{w})^T(\mathbf{y} - \mathbf{X}\mathbf{w}) + \lambda \mathbf{w}^T\mathbf{w}\right) = 0$$

Cheat Sheet Notes

Manipulation:

$$(\mathbf{AB})^T = \mathbf{B}^T\mathbf{A}^T$$
$$(\mathbf{a} + \mathbf{b})^T\mathbf{C} = \mathbf{a}^T\mathbf{C} + \mathbf{b}^T\mathbf{C}$$

Derivatives:

$$\mathbf{x}^T\mathbf{B} \to \mathbf{B}$$
$$\mathbf{x}^T\mathbf{B}\mathbf{x} \to 2\mathbf{B}\mathbf{x}$$
$$\mathbf{A}\mathbf{A}^{-1} = \mathbf{I}$$

Ridge Regression - Derivation

$$\frac{\partial \mathcal{L}}{\partial \mathbf{w}} = \frac{\partial}{\partial \mathbf{w}}\left((\mathbf{y} - \mathbf{X}\mathbf{w})^T(\mathbf{y} - \mathbf{X}\mathbf{w}) + \lambda \mathbf{w}^T\mathbf{w}\right) = 0$$
$$= 2\mathbf{X}^T\mathbf{X}\mathbf{w} - 2\mathbf{X}^T\mathbf{y} + 2\lambda\mathbf{w} = 0$$

Cheat Sheet Notes

Manipulation:

$$(\mathbf{AB})^T = \mathbf{B}^T\mathbf{A}^T$$
$$(\mathbf{a} + \mathbf{b})^T\mathbf{C} = \mathbf{a}^T\mathbf{C} + \mathbf{b}^T\mathbf{C}$$

Derivatives:

$$\mathbf{x}^T\mathbf{B} \rightarrow \mathbf{B}$$
$$\mathbf{x}^T\mathbf{B}\mathbf{x} \rightarrow 2\mathbf{B}\mathbf{x}$$
$$\mathbf{A}\mathbf{A}^{-1} = \mathbf{I}$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{w}} = \frac{\partial}{\partial \mathbf{w}}\left((\mathbf{y} - \mathbf{X}\mathbf{w})^T(\mathbf{y} - \mathbf{X}\mathbf{w}) + \lambda \mathbf{w}^T\mathbf{w}\right) = 0$$
$$= 2\mathbf{X}^T\mathbf{X}\mathbf{w} - 2\mathbf{X}^T\mathbf{y} + 2\lambda\mathbf{w} = 0$$

Solving for $\mathbf{w}$

Cheat Sheet Notes

Manipulation:

$$(\mathbf{AB})^T = \mathbf{B}^T\mathbf{A}^T$$
$$(\mathbf{a} + \mathbf{b})^T\mathbf{C} = \mathbf{a}^T\mathbf{C} + \mathbf{b}^T\mathbf{C}$$

Derivatives:

$$\mathbf{x}^T\mathbf{B} \rightarrow \mathbf{B}$$
$$\mathbf{x}^T\mathbf{B}\mathbf{x} \rightarrow 2\mathbf{B}\mathbf{x}$$
$$\mathbf{A}\mathbf{A}^{-1} = \mathbf{I}$$

Ridge Regression - Derivation

$$\frac{\partial \mathcal{L}}{\partial \mathbf{w}} = \frac{\partial}{\partial \mathbf{w}}\left((\mathbf{y} - \mathbf{X}\mathbf{w})^T(\mathbf{y} - \mathbf{X}\mathbf{w}) + \lambda \mathbf{w}^T\mathbf{w}\right) = 0$$
$$= 2\mathbf{X}^T\mathbf{X}\mathbf{w} - 2\mathbf{X}^T\mathbf{y} + 2\lambda\mathbf{w} = 0$$

Solving for $\mathbf{w}$

$$2\mathbf{X}^T\mathbf{X}\mathbf{w} + 2\lambda\mathbf{w} = 2\mathbf{X}^T\mathbf{y}$$

Cheat Sheet Notes

Manipulation:

$$(\mathbf{AB})^T = \mathbf{B}^T\mathbf{A}^T$$
$$(\mathbf{a} + \mathbf{b})^T\mathbf{C} = \mathbf{a}^T\mathbf{C} + \mathbf{b}^T\mathbf{C}$$

Derivatives:

$$\mathbf{x}^T\mathbf{B} \rightarrow \mathbf{B}$$
$$\mathbf{x}^T\mathbf{B}\mathbf{x} \rightarrow 2\mathbf{B}\mathbf{x}$$
$$\mathbf{A}\mathbf{A}^{-1} = \mathbf{I}$$

Ridge Regression - Derivation

$$\frac{\partial \mathcal{L}}{\partial \mathbf{w}} = \frac{\partial}{\partial \mathbf{w}} \left((\mathbf{y} - \mathbf{X}\mathbf{w})^T (\mathbf{y} - \mathbf{X}\mathbf{w}) + \lambda \mathbf{w}^T \mathbf{w} \right) = 0$$
$$= 2\mathbf{X}^T\mathbf{X}\mathbf{w} - 2\mathbf{X}^T\mathbf{y} + 2\lambda\mathbf{w} = 0$$

Solving for $\mathbf{w}$

$$2\mathbf{X}^T\mathbf{X}\mathbf{w} + 2\lambda\mathbf{w} = 2\mathbf{X}^T\mathbf{y}$$
$$(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})\mathbf{w} = \mathbf{X}^T\mathbf{y}$$

Cheat Sheet Notes

Manipulation:

$$(\mathbf{AB})^T = \mathbf{B}^T\mathbf{A}^T$$
$$(\mathbf{a} + \mathbf{b})^T\mathbf{C} = \mathbf{a}^T\mathbf{C} + \mathbf{b}^T\mathbf{C}$$

Derivatives:

$$\mathbf{x}^T\mathbf{B} \rightarrow \mathbf{B}$$
$$\mathbf{x}^T\mathbf{B}\mathbf{x} \rightarrow 2\mathbf{B}\mathbf{x}$$
$$\mathbf{A}\mathbf{A}^{-1} = \mathbf{I}$$

Ridge Regression - Derivation

$$\frac{\partial \mathcal{L}}{\partial \mathbf{w}} = \frac{\partial}{\partial \mathbf{w}}\left((\mathbf{y} - \mathbf{X}\mathbf{w})^T(\mathbf{y} - \mathbf{X}\mathbf{w}) + \lambda \mathbf{w}^T\mathbf{w}\right) = 0$$
$$= 2\mathbf{X}^T\mathbf{X}\mathbf{w} - 2\mathbf{X}^T\mathbf{y} + 2\lambda\mathbf{w} = 0$$

Solving for $\mathbf{w}$

$$2\mathbf{X}^T\mathbf{X}\mathbf{w} + 2\lambda\mathbf{w} = 2\mathbf{X}^T\mathbf{y}$$
$$(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})\mathbf{w} = \mathbf{X}^T\mathbf{y}$$
$$(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})\mathbf{w} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}$$

Cheat Sheet Notes

Manipulation:

$$(\mathbf{AB})^T = \mathbf{B}^T\mathbf{A}^T$$
$$(\mathbf{a} + \mathbf{b})^T\mathbf{C} = \mathbf{a}^T\mathbf{C} + \mathbf{b}^T\mathbf{C}$$

Derivatives:

$$\mathbf{x}^T\mathbf{B} \rightarrow \mathbf{B}$$
$$\mathbf{x}^T\mathbf{B}\mathbf{x} \rightarrow 2\mathbf{B}\mathbf{x}$$
$$\mathbf{A}\mathbf{A}^{-1} = \mathbf{I}$$

Ridge Regression - Derivation

$$\frac{\partial \mathcal{L}}{\partial \mathbf{w}} = \frac{\partial}{\partial \mathbf{w}} \left((\mathbf{y} - \mathbf{X}\mathbf{w})^T (\mathbf{y} - \mathbf{X}\mathbf{w}) + \lambda \mathbf{w}^T \mathbf{w} \right) = 0$$
$$= 2\mathbf{X}^T\mathbf{X}\mathbf{w} - 2\mathbf{X}^T\mathbf{y} + 2\lambda\mathbf{w} = 0$$

Solving for $\mathbf{w}$

$$2\mathbf{X}^T\mathbf{X}\mathbf{w} + 2\lambda\mathbf{w} = 2\mathbf{X}^T\mathbf{y}$$
$$(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})\mathbf{w} = \mathbf{X}^T\mathbf{y}$$
$$(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})\mathbf{w} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}$$

which leads to

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

Ridge regressor estimate

Cheat Sheet Notes

Manipulation:

$$(\mathbf{AB})^T = \mathbf{B}^T\mathbf{A}^T$$
$$(\mathbf{a} + \mathbf{b})^T\mathbf{C} = \mathbf{a}^T\mathbf{C} + \mathbf{b}^T\mathbf{C}$$

Derivatives:

$$\mathbf{x}^T\mathbf{B} \rightarrow \mathbf{B}$$
$$\mathbf{x}^T\mathbf{B}\mathbf{x} \rightarrow 2\mathbf{B}\mathbf{x}$$
$$\mathbf{A}\mathbf{A}^{-1} = \mathbf{I}$$

Example: Lasso Regression

Linear regression objective function:

$$\arg\min_{\mathbf{w}} \mathcal{L}(\mathbf{w}) = \arg\min_{\mathbf{w}} \ (\mathbf{y} - \mathbf{X}\mathbf{w})^T (\mathbf{y} - \mathbf{X}\mathbf{w})$$

If we include L1-regularization the new objective becomes:

$$\arg\min_{\mathbf{w}} \mathcal{L}(\mathbf{w}) + \lambda R(\mathbf{w}) = \arg\min_{\mathbf{w}} \ (\mathbf{y} - \mathbf{X}\mathbf{w})^T (\mathbf{y} - \mathbf{X}\mathbf{w}) + \lambda \|\mathbf{w}\|_1$$

Unlike ridge regression, lasso regression has no closed-form solution.

The original implementation involves quadratic programming techniques from convex optimization

Regularization as a Constrained Optimization Problem

- The two examples we just presented can be reformulated as constrained optimization problems
- The regularized quadratic cost function (L2-regularization)

$$(\mathbf{y} - \mathbf{X}\mathbf{w})^T (\mathbf{y} - \mathbf{X}\mathbf{w}) + \lambda \mathbf{w}^T \mathbf{w}$$

Regularization as a Constrained Optimization Problem

- The two examples we just presented can be reformulated as constrained optimization problems
- The regularized quadratic cost function (L2-regularization)

$$(\mathbf{y} - \mathbf{X}\mathbf{w})^T (\mathbf{y} - \mathbf{X}\mathbf{w}) + \lambda \mathbf{w}^T \mathbf{w}$$

can be reformulated as

$$\begin{aligned} \arg\min_{\mathbf{w}} \quad & (\mathbf{y} - \mathbf{X}\mathbf{w})^T (\mathbf{y} - \mathbf{X}\mathbf{w}) \\ \text{subject to} \quad & \mathbf{w}^T \mathbf{w} \leq K \end{aligned}$$

Ridge Regression with Constraint Definition

Regularization as a Constrained Optimization Problem

Similarly, the regularized quadratic cost function with L1 penalty

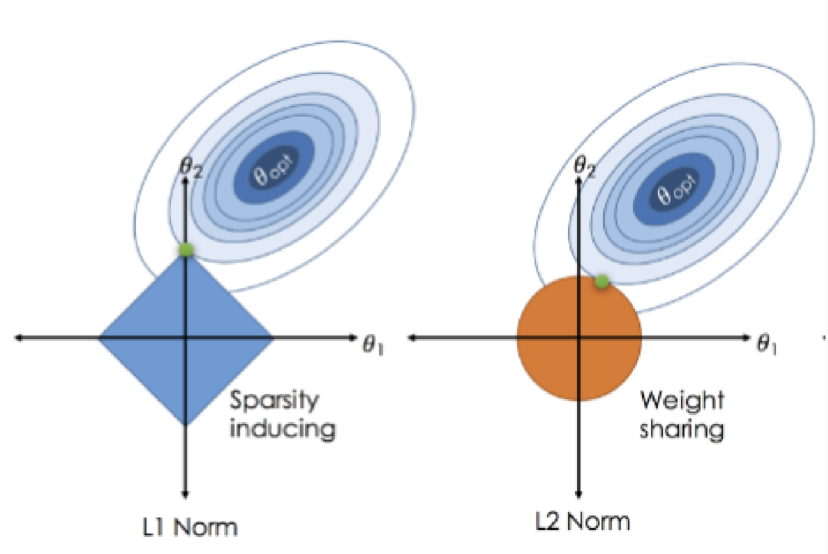
$$(\mathbf{y} - \mathbf{X}\mathbf{w})^T (\mathbf{y} - \mathbf{X}\mathbf{w}) + \lambda \|\mathbf{w}\|_1$$

can be reformulated as

$$\begin{aligned} &\arg\min_{\mathbf{w}} \quad && (\mathbf{y} - \mathbf{X}\mathbf{w})^T (\mathbf{y} - \mathbf{X}\mathbf{w}) \\ &\text{subject to} && \|\mathbf{w}\|_1 \le K \end{aligned}$$

Lasso Regression with Constraint Definition

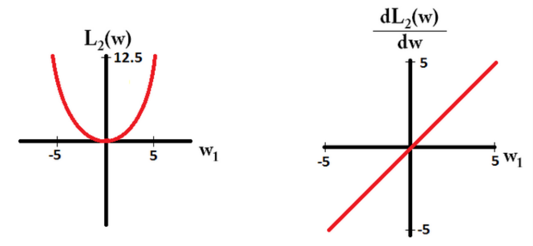
Interpretation using the Constrained Optimization Formulation



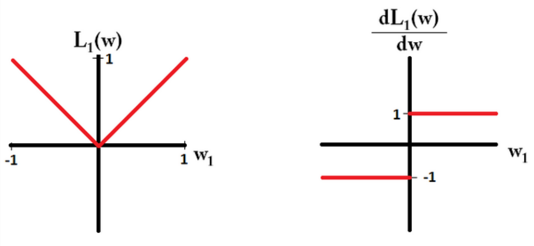
Source: `https://towardsdatascience.com/from-linear-regression-to-ridge-regression-the-lasso-and-the-elastic-net-4eaecaf5f7e6`

Gradient Descent Behavior

L2-Regularization



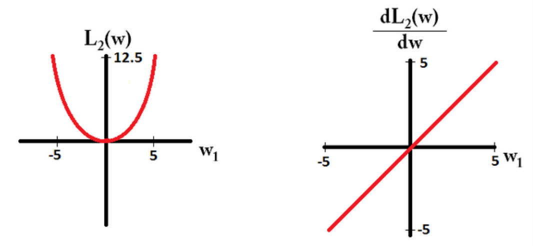
L1-Regularization



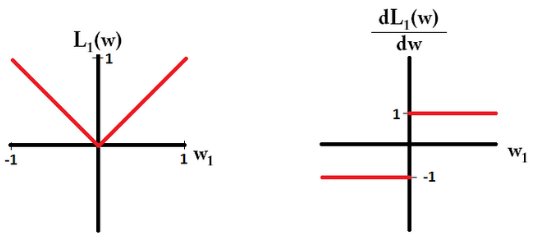
Source: `https://stats.stackexchange.com/questions/45643/why-l1-norm-for-sparse-models`

Gradient Descent Behavior

L2-Regularization



L1-Regularization

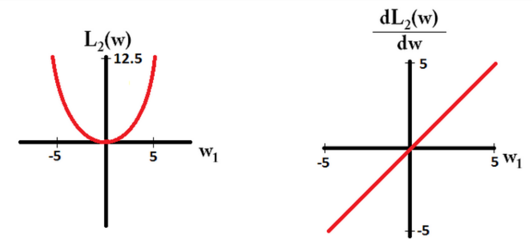


Source: `https://stats.stackexchange.com/questions/45643/why-l1-norm-for-sparse-models`

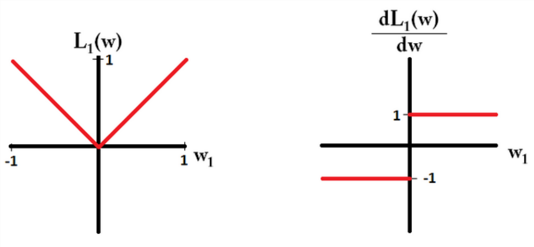
- L2-regularization will also move any weight towards 0, but it will take smaller and smaller steps as a weight approaches 0

Gradient Descent Behavior

L2-Regularization



L1-Regularization



Source: `https://stats.stackexchange.com/questions/45643/why-l1-norm-for-sparse-models`

- L2-regularization will also move any weight towards 0, but it will take smaller and smaller steps as a weight approaches 0
- L1-regularization will move any weight towards 0 with the same step size, regardless the weight's value. his means that Lasso leads to sparse solutions (many zeros)
- The Lasso can also be used for feature selection.

When to Use Regularization?

- **If $D > N$:**
  Not possible to invert $\hat{\mathbf{w}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$
  The constraints added via regularization allow to solve ill-posed problems
- **To reduce variance:**
  Ridge regression shrinks towards zero the size of the coefficients
- **To perform feature selection:**
  By introducing sparsity (variables go to zero), Lasso reduces variance and does variable selection

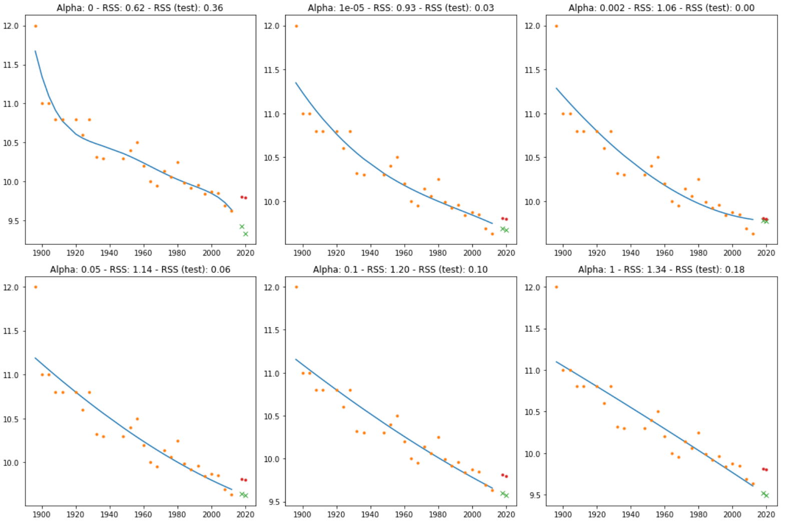
Example: Revisiting Polynomial Features

- Let us recall the use of polynomial features
- nth order model:

$$\hat{y} = \hat{w}_0 + \hat{w}_1 x + \hat{w}_2 x^2 + \ldots + \hat{w}_n x^n$$

- Could we use regularization as a way to determine the good order?

Example: 100m Olympic Games Revisited



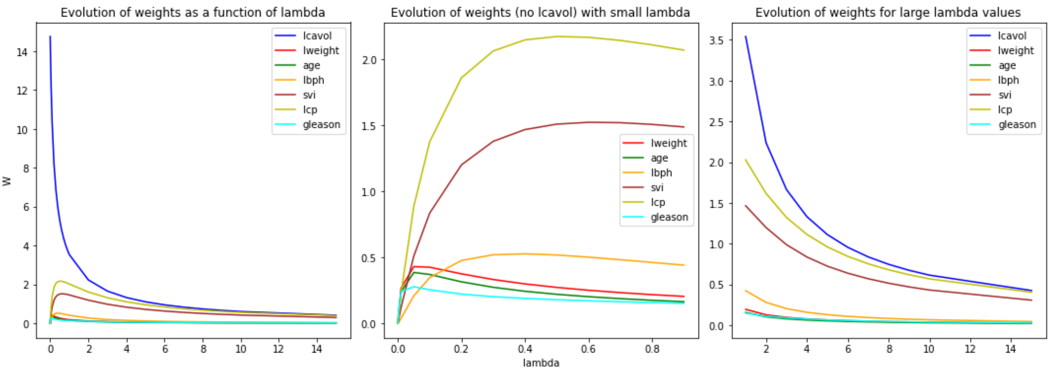
Notebook: See `05_regularization.ipynb`

Example: 100m Olympic Games Revisited

```
alpha: 0 gives W: [-6.23269427e-09 -8.13638626e+04 -1.23154391e+06  3.99864925e+06
 -2.10982203e+06 -4.13071960e+06  5.24217644e+06 -1.68737681e+06]
alpha: 1e-05 gives W: [ 4.50905945e-16 -4.11498118e+01 -8.99905545e-01  2.32795512e+01
  3.15181252e+01  2.39574392e+01  7.50942192e-01 -3.79363466e+01]
alpha: 0.002 gives W: [-4.66498851e-16 -1.73507591e+00 -9.97927927e-01 -3.69725207e-01
  1.50016308e-01  5.61866710e-01  8.66481144e-01  1.06459864e+00]
alpha: 0.05 gives W: [-4.51061544e-16 -7.79251911e-01 -5.34277344e-01 -2.93511369e-01
 -5.70108294e-02  1.75170822e-01  4.02983604e-01  6.26381070e-01]
alpha: 0.1 gives W: [-4.43138216e-16 -5.33876150e-01 -3.74265676e-01 -2.16731203e-01
 -6.13142840e-02  9.19452170e-02  2.43009178e-01  3.91841276e-01]
alpha: 1 gives W: [-4.16716051e-16 -1.30621767e-01 -1.08510109e-01 -8.65980794e-02
 -6.48920071e-02 -4.33980570e-02 -2.21222209e-02 -1.07030941e-03]
```

Notebook: See `05_regularization.ipynb`

Example: Prostate Cancer Dataset



Notebook: See `05_regularization.ipynb`

Open points

- How we choose the good value for λ? The order of the polynomial?

Open points

- How we choose the good value for λ? The order of the polynomial?
- In regression, we want to keep control of the size of $\mathbf{w}_{1:D}$. How we implement this if w_0 should not be affected?

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

Open points

- How we choose the good value for λ? The order of the polynomial?
- In regression, we want to keep control of the size of $\mathbf{w}_{1:D}$. How we implement this if w_0 should not be affected?

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

- A better understanding of the variance and its consequences in a model's performance

Wrap-up

Wrap-up

- We formalized the concept of regularization
- We saw how to estimate the parameters of the regularized quadratic function using L2-regularization
- We studied some properties of both L2- and L1-regularization in the context of linear regression (ridge and lasso)
- The question on how to estimate the regularization parameter λ remains open

Key Concepts

- Regularization
- Ridge regression
- Lasso regression
- Penalty term
- L2-regularization, L1-regularization

References

Further Reading and Useful Material

Source	Notes
The Elements of Statistical Learning	Ch 3