Modern Regression

Jean-Luc Bouchot

School of Mathematics and Statistics
Beijing Institute of Technology
jlbouchot@bit.edu.cn

Spring 2021

1. Regularization and model selection

Outline

Remark 2.1

As seen from the height/weight dataset, it is, if one wishes, possible to fit perfectly the data.

Theorem 2.1 (Lagrange Interpolation - Admitted)

*Let $\{(x_i, y_i)\}_{1 \leq i \leq n}$ be n samples of a given phenomenon.
Assuming $x_i \neq x_j$ for all $i \neq j$, then there exists a degree $n-1$ polynomial P_n such that the approximation error is 0: $P_n(x_i) = y_i$ for all $1 \leq i \leq n$.*

Remark 2.2

Assume the underlying model is indeed a polynomial one: what happens if the samples are noisy?

Example 2.1

Assume the following data are given

Target	Predictor	Noisy target
-0.5	-2.5	-0.492
1	-1	0.936
2.5	0.5	2.542
4	2	4.011

1. Compute the estimations using polynomial features of degree 0 up to 3 (included)
2. Compute the approximation errors for each of the polynomial features.

Remark 2.3

This gives the following results:

	Degree features	Coef 0	Coef 1	Coef 2	Coef 3	Error with true
Noiseless	d = 0	1.75	0	0	0	3.354
	d = 1	2	1	0	0	0
	d = 2	2	1	0	0	0
	d = 3	2	1	0	0	0
Noisy	d = 0	1.749	0	0	0	3.354
	d = 1	2.001	1.008	0	0	0.026
	d = 2	1.989	1.001	0.005	0	0.033
	d = 3	2.006	1.079	-0.007	-0.016	0.078

Example 2.2

We reiterate the same idea, with the following data:

Predictor	Noisier target	Noisiest target
-2.5	-0.502	0.013
-1	0.922	1.204
0.5	2.608	2.473
2	3.896	4.220

Remark 2.4

We obtain the following results:

	Degree features	Coef 0	Coef 1	Coef 2	Coef 3	Error with true
Noisy	d = 0	1.749	0	0	0	3.354
	d = 1	2.001	1.008	0	0	0.026
	d = 2	1.989	1.001	0.005	0	0.033
	d = 3	2.006	1.079	-0.007	-0.016	0.078
Noisier	d = 0	1.731	0	0	0	3.354
	d = 1	1.979	0.992	0	0	0.047
	d = 2	2.021	0.984	-0.015	0	0.082
	d = 3	2.05	1.129	-0.040	-0.033	0.169
Noisiest	d = 0	1.978	0	0	0	3.385
	d = 1	2.209	0.926	0	0	0.518
	d = 2	2.039	0.957	0.062	0	0.588
	d = 3	2.017	0.870	0.077	0.020	0.595

Example 2.3

Looking back at the solution we have obtained, we notice the following: let $\beta(d)$ denotes the $(d+1)$-dimensional vector of coefficients obtained in the regression, its norm is

	$\|\beta(0)\|$	$\|\beta(1)\|$	$\|\beta(2)\|$	$\|\beta(3)\|$
Noiseless	1.75	2.236	2.236	2.236
Noisy	1.749	2.241	2.230	2.278
Noisier	1.731	2.214	2.248	2.347
Noisiest	1.977	2.395	2.253	2.196

Example 2.4

Let us try on a bigger training set: 20 sampling points uniformly spaced, the target values are computed from a noisy linear model. We let d vary from 0 to 25.

Definition 2.1

The **ridge regression** *is a regression problem which penalizes heavy coefficients. It is expressed as*

$$\widehat{\beta} := \operatorname*{argmin}_{\beta \in \mathbb{R}^D} \|X\beta - \mathbf{y}\|_2^2 + \lambda \|\beta\|_2^2,$$

where $X \in \mathbb{R}^{n \times D}$ *denotes the data matrix and* $\mathbf{y} \in \mathbb{R}^n$ *the target (dependent) variables.*

Proposition 2.1

The Ridge Regression approach is equivalent to the following constrained optimization problem

$$\widehat{\beta} = \operatorname*{argmin}_{\beta \in \mathbb{R}^D} \|X\beta - \mathbf{y}\|_2^2$$

$$\textit{subject to } \|\beta\|_2^2 \leq \tau,$$

for a certain value of τ which depends on λ.