Regularization

Johanni Brea

Einführung Machine Learning

GYMINF 2022

Table of Contents

When Linear Models Are Too Flexible

In the old days

Typically $n > p$ (much more data than predictors)

For example: predict blood pressure based on age, gender and body mass index (BMI) (e.g. $n = 200$ patients, $p = 3$).

When Linear Models Are Too Flexible

In the old days

Typically $n > p$ (much more data than predictors)

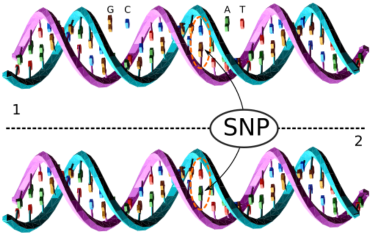
For example: predict blood pressure based on age, gender and body mass index (BMI) (e.g. $n = 200$ patients, $p = 3$).

Nowadays: Big Data

Often $n \approx p$ or $n < p$

For example: predict blood pressure based on 500 000 single nucleotide polymorphisms (SNP) ($n = 200$, $p = 500\,000$).

$\Rightarrow$ **Linear Model perfectly fits the training data.**



Making Linear Models Less Flexible

Idea 1: Fix some parameters at zero

$$\hat{y} = f(x) = f(x_1, x_2, \ldots, x_p) = \beta_0 + \cancel{\underbrace{\beta_1}_{=0} x_1} + \cancel{\underbrace{\beta_2}_{=0} x_2} + \beta_3 x_3 + \cdots + \cancel{\underbrace{\beta_{p-1}}_{=0} x_{p-1}} + \beta_p x_p$$

Making Linear Models Less Flexible

Idea 1: Fix some parameters at zero

$$\hat{y} = f(x) = f(x_1, x_2, \ldots, x_p) = \beta_0 + \underbrace{\beta_1}_{=0} \cancel{x_1} + \underbrace{\beta_2}_{=0} \cancel{x_2} + \beta_3 x_3 + \cdots + \underbrace{\beta_{p-1}}_{=0} \cancel{x_{p-1}} + \beta_p x_p$$

Problem: Many different models to fit; $\binom{p+1}{m}$ combinations of m non-fixed parameters.

Making Linear Models Less Flexible

Idea 1: Fix some parameters at zero

$$\hat{y} = f(x) = f(x_1, x_2, \ldots, x_p) = \beta_0 + \cancel{\underbrace{\beta_1}_{=0} x_1} + \cancel{\underbrace{\beta_2}_{=0} x_2} + \beta_3 x_3 + \cdots + \cancel{\underbrace{\beta_{p-1}}_{=0} x_{p-1}} + \beta_p x_p$$

Problem: Many different models to fit; $\binom{p+1}{m}$ combinations of m non-fixed parameters.

Idea 2: constrain the parameters

Minimize the original loss $L(\beta)$ under the constraint $\|\beta\|_2^2 = \sum_{i=1}^{p} \beta_i^2 \leq S$.

Making Linear Models Less Flexible

Idea 1: Fix some parameters at zero

$$\hat{y} = f(x) = f(x_1, x_2, \dots, x_p) = \beta_0 + \cancel{\underbrace{\beta_1}_{\neq 0} x_1} + \cancel{\underbrace{\beta_2}_{\neq 0} x_2} + \beta_3 x_3 + \cdots + \cancel{\underbrace{\beta_{p-1}}_{\neq 0} x_{p-1}} + \beta_p x_p$$

Problem: Many different models to fit; $\binom{p+1}{m}$ combinations of m non-fixed parameters.

Idea 2: constrain the parameters

Minimize the original loss $L(\beta)$ under the constraint $\|\beta\|_2^2 = \sum_{i=1}^{p} \beta_i^2 \leq S$.

This is equivalent to replacing the original loss $L(\beta)$ by

$$L_{\mathrm{L2}}(\beta) = L(\beta) + \lambda \|\beta\|_2^2$$

Table of Contents

Ridge Regression (L2 Regularization)

$$L_{\text{L2}}(\theta) = L(\theta) + \lambda\|\theta\|_2^2$$

with **regularization constant** λ and (squared) **L2 norm** $\|\theta\|_2^2 = \sum_{i=1}^p \theta_i^2$.

1. The regularization constant λ is a hyper-parameter.

Ridge Regression (L2 Regularization)

$$L_{L2}(\theta) = L(\theta) + \lambda \|\theta\|_2^2$$

with **regularization constant** λ and (squared) **L2 norm** $\|\theta\|_2^2 = \sum_{i=1}^{p} \theta_i^2$.

1. The regularization constant λ is a hyper-parameter.
2. Often the intercept θ_0 is not regularized.

Ridge Regression (L2 Regularization)

$$L_{\mathrm{L2}}(\theta) = L(\theta) + \lambda \|\theta\|_2^2$$

with **regularization constant** λ and (squared) **L2 norm** $\|\theta\|_2^2 = \sum_{i=1}^{p} \theta_i^2$.

1. The regularization constant λ is a hyper-parameter.
2. Often the intercept θ_0 is not regularized.
3. If $\lambda = 0$: original loss (no penalty)

Ridge Regression (L2 Regularization)

$$L_{\mathrm{L2}}(\theta) = L(\theta) + \lambda\|\theta\|_2^2$$

with **regularization constant** λ and (squared) **L2 norm** $\|\theta\|_2^2 = \sum_{i=1}^{p} \theta_i^2$.

1. The regularization constant λ is a hyper-parameter.
2. Often the intercept θ_0 is not regularized.
3. If $\lambda = 0$: original loss (no penalty)
4. The larger λ, the stronger the impact of the penalty on the result.

Ridge Regression (L2 Regularization)

$$L_{\mathrm{L2}}(\theta) = L(\theta) + \lambda \|\theta\|_2^2$$

with **regularization constant** λ and (squared) **L2 norm** $\|\theta\|_2^2 = \sum_{i=1}^{p} \theta_i^2$.

1. The regularization constant λ is a hyper-parameter.
2. Often the intercept θ_0 is not regularized.
3. If $\lambda = 0$: original loss (no penalty)
4. The larger λ, the stronger the impact of the penalty on the result.
5. With increasing λ the model becomes less flexible.

Ridge Regression (L2 Regularization)

$$L_{\mathrm{L2}}(\theta) = L(\theta) + \lambda\|\theta\|_2^2$$

with **regularization constant** λ and (squared) **L2 norm** $\|\theta\|_2^2 = \sum_{i=1}^{p} \theta_i^2$.

1. The regularization constant λ is a hyper-parameter.
2. Often the intercept θ_0 is not regularized.
3. If $\lambda = 0$: original loss (no penalty)
4. The larger λ, the stronger the impact of the penalty on the result.
5. With increasing λ the model becomes less flexible.
6. With increasing λ all parameters tend to zero; it happens rarely that one is exactly zero.

Lasso (L1 Regularization)

$$L_{\mathrm{L1}}(\theta) = L(\theta) + \lambda\|\theta\|_1$$

with **regularization constant** λ and **L1 norm** $\|\theta\|_1 = \sum_{i=1}^{p} |\theta_i|$.

Points 1-5 from ridge regression are also valid for the Lasso. However:

6. With large λ some parameters are exactly zero (in contrast to ridge regression).

An Alternative Formulation of Regularization

Thanks to a result from constraint optimization (see Karush-Kuhn-Tucker conditions, a generalization of Lagrange multipliers) the above formulations of Ridge Regression and the Lasso are equivalent to a constraint optimization problem:

An Alternative Formulation of Regularization

Thanks to a result from constraint optimization (see Karush-Kuhn-Tucker conditions, a generalization of Lagrange multipliers) the above formulations of Ridge Regression and the Lasso are equivalent to a constraint optimization problem:

Ridge Regression

minimize $L(\theta)$ under the constraint that $\|\theta\|_2^2 \leq S$.

The parameters are confined to a p-ball of radius S with center at the origin.

An Alternative Formulation of Regularization

Thanks to a result from constraint optimization (see Karush-Kuhn-Tucker conditions, a generalization of Lagrange multipliers) the above formulations of Ridge Regression and the Lasso are equivalent to a constraint optimization problem:

Ridge Regression

minimize $L(\theta)$ under the constraint that $\|\theta\|_2^2 \leq S$.

The parameters are confined to a p-ball of radius S with center at the origin.

Lasso

minimize $L(\theta)$ under the constraint that $\|\theta\|_1 \leq S$.

The parameters are confined to a hypercube with edge length S, center at the origin and corners on the axes.

An Alternative Formulation of Regularization

Thanks to a result from constraint optimization (see Karush-Kuhn-Tucker conditions, a generalization of Lagrange multipliers) the above formulations of Ridge Regression and the Lasso are equivalent to a constraint optimization problem:

Ridge Regression

minimize $L(\theta)$ under the constraint that $\|\theta\|_2^2 \leq S$.

The parameters are confined to a p-ball of radius S with center at the origin.

Lasso

minimize $L(\theta)$ under the constraint that $\|\theta\|_1 \leq S$.

The parameters are confined to a hypercube with edge length S, center at the origin and corners on the axes.

S is a (complicated) function of λ and the original loss $L(\theta)$.
With increasing S the model becomes more flexible.

Analytical Solutions for Simple Linear Regression

Notation: $\langle x \rangle = \frac{1}{n} \sum_{i=1}^{n} x_i$

Ridge Regression

$$L(\theta, \lambda) = \langle (y - \theta_0 - \theta_1 x)^2 \rangle + \lambda \theta_1^2$$

$$\theta_1 = \frac{\langle xy \rangle - \langle x \rangle \langle y \rangle}{\langle x \rangle^2 - \langle x^2 \rangle + \lambda}, \qquad \theta_0 = \langle y \rangle - \theta_1 \langle x \rangle$$

Lasso

$$L(\theta, \lambda) = \frac{1}{2} \langle (y - \theta_0 - \theta_1 x)^2 \rangle + \lambda |\theta_1|$$

$$\theta_1 = \frac{\langle xy \rangle - \langle x \rangle \langle y \rangle - \mathrm{sign}(\theta_1)\lambda}{\langle x \rangle^2 - \langle x^2 \rangle} \text{ or } 0 \text{ if } |\langle xy \rangle - \langle x \rangle \langle y \rangle| < \lambda$$

Standardized Inputs for Regularization

Problem

Assume we find in multiple linear regression on the weather data the following parameters

X_1	LUZ_pressure	[hPa]	$\theta_1 = -1$	[km/h/hPa]
X_2	LUZ_temperature	[°C]	$\theta_2 = 0.5$	[km/h/°C]

Standardized Inputs for Regularization

Problem

Assume we find in multiple linear regression on the weather data the following parameters

X_1	LUZ_pressure	[hPa]	$\theta_1 = -1$	[km/h/hPa]
X_2	LUZ_temperature	[°C]	$\theta_2 = 0.5$	[km/h/°C]

We could have measured the pressure in Pa and get the equivalent result

X_1	LUZ_pressure	[Pa]	$\theta_1 = -1/100$	[km/h/Pa]
X_2	LUZ_temperature	[°C]	$\theta_2 = 0.5$	[km/h/°C]

Standardized Inputs for Regularization

Problem

Assume we find in multiple linear regression on the weather data the following parameters

X_1	LUZ_pressure	[hPa]	$\theta_1 = -1$	[km/h/hPa]
X_2	LUZ_temperature	[°C]	$\theta_2 = 0.5$	[km/h/°C]

We could have measured the pressure in Pa and get the equivalent result

X_1	LUZ_pressure	[Pa]	$\theta_1 = -1/100$	[km/h/Pa]
X_2	LUZ_temperature	[°C]	$\theta_2 = 0.5$	[km/h/°C]

With regularization $\lambda(\theta_1^2 + \theta_2^2)$ we would get different results for measurements in hPa and in Pa, because θ_1 contributes less to the penalty in the latter case.

Standardized Inputs for Regularization

Problem

Assume we find in multiple linear regression on the weather data the following parameters

X_1	LUZ_pressure	[hPa]	$\theta_1 = -1$	[km/h/hPa]
X_2	LUZ_temperature	[°C]	$\theta_2 = 0.5$	[km/h/°C]

We could have measured the pressure in Pa and get the equivalent result

X_1	LUZ_pressure	[Pa]	$\theta_1 = -1/100$	[km/h/Pa]
X_2	LUZ_temperature	[°C]	$\theta_2 = 0.5$	[km/h/°C]

With regularization $\lambda(\theta_1^2 + \theta_2^2)$ we would get different results for measurements in hPa and in Pa, because θ_1 contributes less to the penalty in the latter case.

Solution

Standardize all predictors, such that they have variance 1:

$$\tilde{X}_i = X_i / \sqrt{\mathrm{Var}(X_i)}$$

Scaling of the Regularization Constant with n

With loss $L(\theta) = \sum_{i=1}^{n} \ell(y_i, f(x_i)) + \lambda \|\theta\|_2^2$
the effective regularization depends on the size of the data set.

One can use instead an average loss $L(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f(x_i)) + \lambda \|\theta\|_2^2$ or (equivalently) scale the regularization term $L(\theta) = \sum_{i=1}^{n} \ell(y_i, f(x_i)) + n \cdot \lambda \|\theta\|_2^2$

Quiz

- L1 Regularisierung führt zu grösserer Varianz und kleinerem Bias verglichen mit unregularisierter Regression.
- Was stimmt: Der Trainingsfehler als Funktion von S in L2 regularisierter Regression
  1. hat umgekehrte U-Form
  2. hat U-Form
  3. nimmt kontinuierlich zu
  4. nimmt kontinuierlich ab
  5. ist konstant.

Quiz

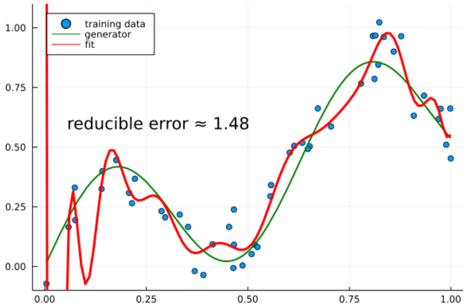
- L1 Regularisierung führt zu grösserer Varianz und kleinerem Bias verglichen mit unregularisierter Regression.
- Was stimmt: Der Trainingsfehler als Funktion von S in L2 regularisierter Regression
  1. hat umgekehrte U-Form
  2. hat U-Form
  3. nimmt kontinuierlich zu
  4. nimmt kontinuierlich ab
  5. ist konstant.

  Was stimmt: Der Testfehler als Funktion von S in L2 regularisierter Regression
  1. hat umgekehrte U-Form
  2. hat U-Form
  3. nimmt kontinuierlich zu
  4. nimmt kontinuierlich ab
  5. ist konstant.

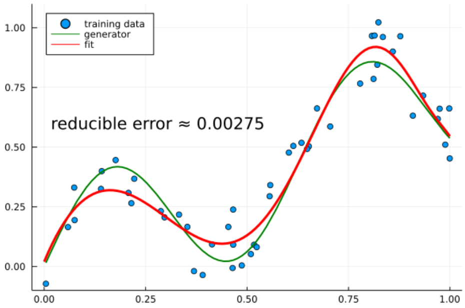
Table of Contents

Polynomial Ridge Regression

$d = 20, \lambda = 0$



$d = 20, \lambda = 10^{-4}$



With a little bit of L2 regularization ($\lambda = 10^{-4}$)
one can prevent overfitting of polynomials with high degrees.

Multiple Logistic Ridge Regression on the Spam Data

$n = 2000$ emails, $p = 801$ features (size of the lexicon)

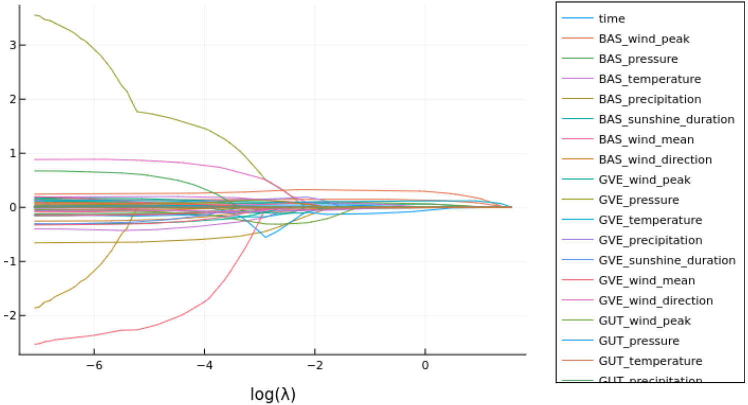
Without regularization

training misclassification rate: 0.0015
test misclassification rate: 0.048

With L2 regularization

training misclassification rate: 0.013
test misclassification rate: 0.041

The Lasso Path for the Weather Data



As we lower λ, `BER_wind_peak` is the first non-zero factor, `BAS_wind_peak` the second and `LUZ_wind_mean` the third.

Summary

- Regularization allows to lower the flexibility of a model by restricting the parameters to certain areas of the parameter space.
- L1 regularization leads to sparse models with some parameters exactly zero $\Rightarrow$ great for interpretability.