Regularization

Week 04 - Day 05

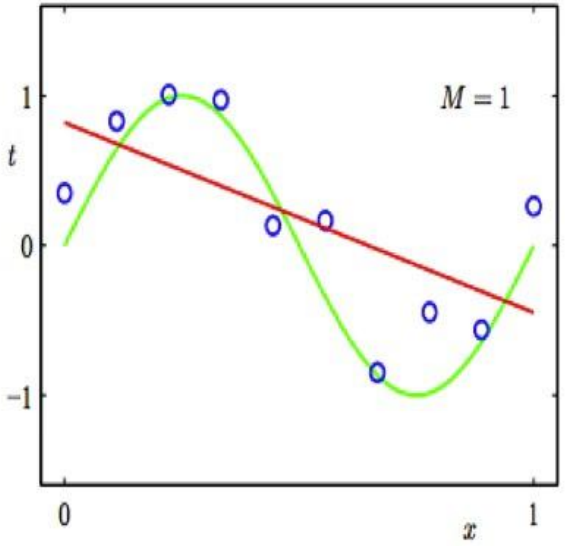
Recap on Linear Models

$$y = b0 + b1*x1 + ... + bn*xn + error$$

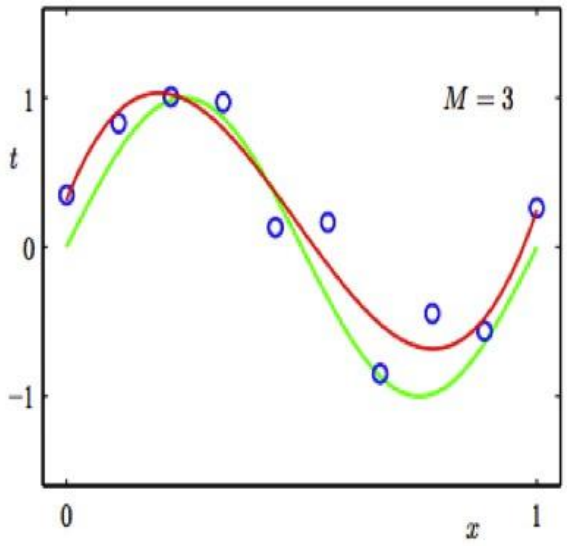
Linear Models

Overfit or underfit?

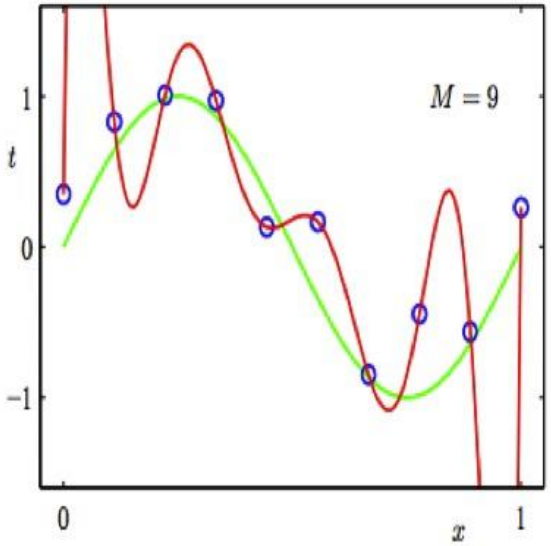
Underfitting



Good



Overfitting



Linear models = Simple models

Low variance

Risk of underfitting

Linear Models can overfit too!

Too many irrelevant features

Correlated features

Too many irrelevant features

Correlated features

What’s the name?

Too many irrelevant features

Correlated features

Multicollinearity!

Solution = Regularization!

Smallest Error

Name of the process to find the smallest error?

(The answer is simple!)

Optimization

Error = SUM((real - predicted)**2)

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Regularization

Optimize error

Optimize error + penalty

(penalty for complexity)

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$\sum_{i=1}^{n} \left(y_i - \left(\beta_0 + \sum_{j=1}^{p} \beta_j x_j \right) \right)^2$$

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$$\sum_{i=1}^{n}\left(y_i - \left(\beta_0 + \sum_{j=1}^{p}\beta_j x_j\right)\right)^2$$

$$\sum_{i=1}^{n}\left(y_i - \left(\beta_0 + \sum_{j=1}^{p}\beta_j x_j\right)\right)^2 + \lambda_2 \sum_{j=1}^{p}\beta_j^2$$

Penalty

What’s the penalty?

Sum of coefficients**2!

(beta1**2 + beta2**2 + beta3**2)

We optimize error+penalty

Everything else remains the same!

	RSS
Model 1	
Model 2	
Model 3	

	RSS
Model 1	100
Model 2	90
Model 3	80

Which one is the best?

	RSS
Model 1	100
Model 2	90
Model 3	80

Which one is the best?

	RSS	Penalty for complexity	Total loss
Model 1	100		
Model 2	90		
Model 3	80		

	RSS	Penalty for complexity	Total loss
Model 1	100	5	105
Model 2	90		
Model 3	80		

	RSS	Penalty for complexity	Total loss
Model 1	100	5	105
Model 2	90	10	100
Model 3	80		

	RSS	Penalty for complexity	Total loss
Model 1	100	5	105
Model 2	90	10	100
Model 3	80	50	130

	RSS	Penalty for complexity	Total loss
Model 1	100	5	105
Model 2	90	10	100
Model 3	80	50	130

Results of Regularization

Smaller coefficients (smaller variance)

Or

Coefficients = 0 (less variables)

Different types of penalty

$$\text{beta1}^{**}2 + \text{beta2}^{**}2 + \text{beta3}^{**}2$$

Vs.

$$|\text{beta1}| + |\text{beta2}| + |\text{beta3}|$$

Ridge

beta1**2 + beta2**2 + beta3**2

Vs.

|beta1| + |beta2| + |beta3|

Lasso

Ridge = smaller coefficients

Lasso = zeroed coefficients

Parameters Tuning

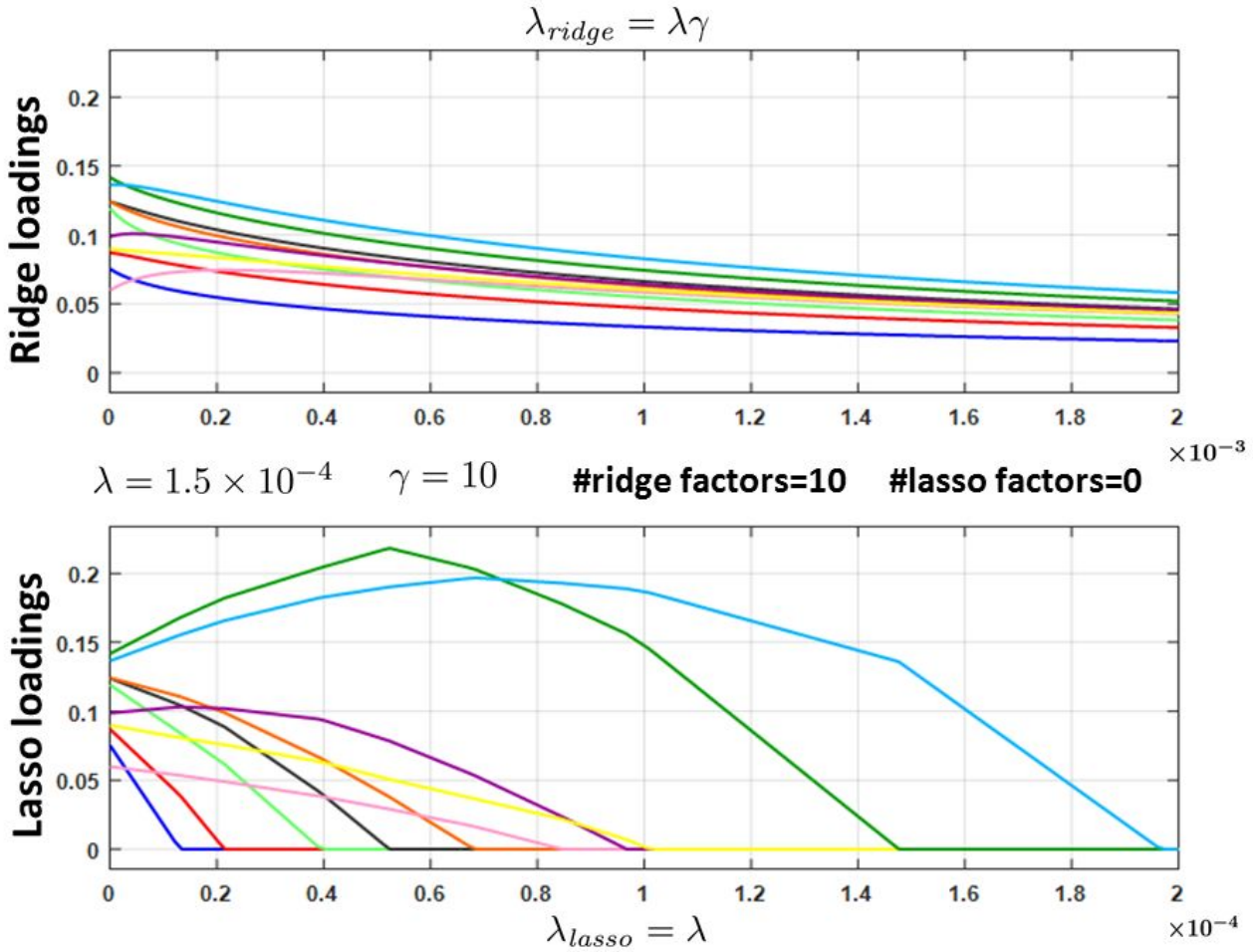
$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$$\sum_{i=1}^{n}\left(y_i - \left(\beta_0 + \sum_{j=1}^{p}\beta_j x_j\right)\right)^2$$

$$\sum_{i=1}^{n}\left(y_i - \left(\beta_0 + \sum_{j=1}^{p}\beta_j x_j\right)\right)^2 + \lambda_2 \sum_{j=1}^{p}\beta_j^2$$

Parameter to tune

Penalty


$\lambda_{ridge} = \lambda\gamma$
Ridge loadings
0.2
0.15
0.1
0.05
0
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
2
$\times 10^{-3}$
$\lambda = 1.5 \times 10^{-4}$
$\gamma = 10$
#ridge factors=10
#lasso factors=0
Lasso loadings
0.2
0.15
0.1
0.05
0
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
2
$\times 10^{-4}$
$\lambda_{lasso} = \lambda$

Coefficients: Normalized or not?

Yes!

We don’t want to penalise the “scale”

Elastic Net

Elastic net = lasso + ridge

Parameter 1 = strength of the penalty

Parameter 2 = more lasso vs. more ridge

Practical advices

Try all models!

Play with parameters tuning!

Sklearn has all these models

Summary

- 3 new models (lasso, ridge, elastic net)
- Fight overfitting + multicollinearity
- Optimize error + penalty
- We need to tune the (hyper)parameters