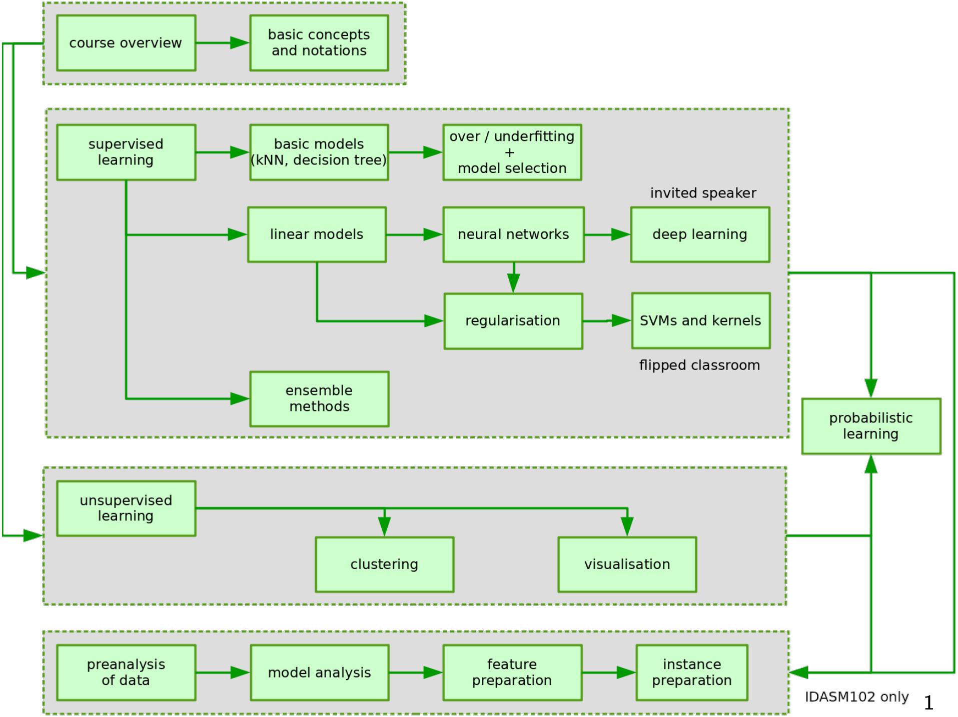
Machine Learning: Lesson 8

Regularisation

Benoît Frénay - Faculty of Computer Science

UNIVERSITÉ DE NAMUR


course overview
basic concepts and notations
supervised learning
basic models (kNN, decision tree)
over / underfitting + model selection
invited speaker
linear models
neural networks
deep learning
regularisation
SVMs and kernels
flipped classroom
ensemble methods
probabilistic learning
unsupervised learning
clustering
visualisation
preanalysis of data
model analysis
feature preparation
instance preparation
IDASM102 only

Outline of this Lesson

- motivation
- regularisation
- regularising linear models
  - L0: feature selection
  - L2: ridge regression
  - L1: LASSO / LARS
  - L1/2: elastic net

Motivation

Linear Models for Regression

Available data

a set of n training data $\mathcal{D} = \{(\mathbf{x}_i, t_i)\}$ where

- $\mathbf{x} \in \Re^d$ is a vector of d continuous features
- $\mathbf{t} \in \Re$ is a continuous target value

Linear modelling

assumption = feature values in $\mathbf{x}$ and the target value t are linearly related

$$f(x_1, \ldots, x_n) = w_1 x_1 + \cdots + w_d x_d + w_0$$

each weight w_j models the contribution of feature x_j to the target value t

Boston Housing Prices Dataset ($n = 506$ and $d = 13$)

crim	per capita crime rate by town
zn	proportion of residential land zoned for lots over 25,000 sq. ft.
indus	proportion of non-retain business acres per town
chas	Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
nox	nitrogen oxides concentration (parts per million)
rm	average number of rooms per dwelling
age	proportion of owner-occupied units built prior to 1940
dis	weighted mean of distances to five Boston employment centers
rad	index of accessibility to radial highways
tax	full-value property-tax rate per $10,000
ptratio	pupil-teacher ratio by town
black	1000(Bk - 0.63)^2, where Bk is the proportion of blacks by town
lstat	lower status of the population (percent)
medv	median value of owner-occupied homes in $1000s

Result of linear regression (top 5 features and mean error ≈ $3200)

$$f\left(x_{\text{crim}} \ldots | w_{\text{crim}} \ldots\right) = 2.7\, x_{\text{rm}} - 3.1\, x_{\text{dis}} + 2.7\, x_{\text{rad}} - 2.1\, x_{\text{ptratio}} - 3.7\, x_{\text{lstat}} + 22.53$$

Optimising Linear Models for Regression

Criterion: mean square error

in practice, it is often impossible to exactly reproduce the target values

- the relationship between $\mathbf{x}$ and t may be partially non-linear
- t is often affected by some noise (measurement, transcription, etc.)

one solution is to minimise the mean square error

$$\mathrm{MSE} = \frac{1}{n} \sum_{i=1}^{n} \left(t_i - f(x_{i1}, \ldots, x_{id})\right)^2$$

Algorithms for MSE optimisation

linear models can be optimised for regression w.r.t. the MSE

- pseudo-inverse method: analytical, exact solution in one step
- iterative algorithms like (stochastic) gradient descent

Limitations of Models without Complexity Control

Small samples issues

- model parameters are unstable over repetitions
- overfitting occurs and cannot be easily avoided

Interpretability issues

- many (all?) model parameters are non-zero
- this issue is **not alleviated** when $n \to \infty$

Solution

we need a mechanism to control complexity $\neq$ metaparameter selection

- helps the learning procedure to avoid unreasonable parameter values (similar to bayesian priors that enforce *a priori* reasonable models)
- can rule out models with (too many) non-zero parameters

Limitations of Models without Complexity Control

Small samples issues

- model parameters are unstable over repetitions
- overfitting occurs and cannot be easily avoided

Interpretability issues

- many (all?) model parameters are non-zero
- this issue is **not alleviated** when $n \to \infty$

Solution

we need a mechanism to control complexity $\neq$ metaparameter selection

- helps the learning procedure to avoid unreasonable parameter values (similar to bayesian priors that enforce *a priori* reasonable models)
- can rule out models with (too many) non-zero parameters

Limitations of Models without Complexity Control

Small samples issues

- model parameters are unstable over repetitions
- overfitting occurs and cannot be easily avoided

Interpretability issues

- many (all?) model parameters are non-zero
- this issue is **not alleviated** when $n \to \infty$

Solution

we need a mechanism to control complexity $\neq$ metaparameter selection

- helps the learning procedure to avoid unreasonable parameter values (similar to bayesian priors that enforce *a priori* reasonable models)
- can rule out models with (too many) non-zero parameters

Regularisation

Controlling the Behaviour of Parameters

Controlling the complexity

regularisation = control the model complexity using a measure $\Omega(\boldsymbol{\theta})$

- $\Omega(\boldsymbol{\theta})$ monotonically increases with the model complexity
- $\Omega(\boldsymbol{\theta})$ = measure of complexity that fits your needs, e.g.
  - the number of non-zero parameters in the model $\Rightarrow$ sparsity
  - the number of (too) large parameters in the model $\Rightarrow$ smoothness

Example: add constraints to the mean square error

the ordinary least square (OLS) solution is obtained by

$$\mathbf{w} = \arg\min \frac{1}{n} \sum_{i=1}^{n} (t_i - f(x_{i1}, \ldots, x_{id}))^2 = \frac{1}{n} (\mathbf{Xw} - \mathbf{t})^T (\mathbf{Xw} - \mathbf{t})$$

example of $\Omega(\mathbf{w}) = \sum_j w_j^2 = \|\mathbf{w}\|_2^2$ to avoid too large weights in OLS

Controlling the Behaviour of Parameters

Controlling the complexity

regularisation = control the model complexity using a measure $\Omega(\boldsymbol{\theta})$

- $\Omega(\boldsymbol{\theta})$ monotonically increases with the model complexity
- $\Omega(\boldsymbol{\theta})$ = measure of complexity that fits your needs, e.g.
  - the number of non-zero parameters in the model $\Rightarrow$ sparsity
  - the number of (too) large parameters in the model $\Rightarrow$ smoothness

Example: add constraints to the mean square error

the ordinary least square (OLS) solution is obtained by

$$\mathbf{w} = \arg\min \frac{1}{n} \sum_{i=1}^{n} \left(t_i - f(x_{i1}, \ldots, x_{id})\right)^2 = \frac{1}{n} \left(\mathbf{Xw} - \mathbf{t}\right)^T \left(\mathbf{Xw} - \mathbf{t}\right)$$

example of $\Omega(\mathbf{w}) = \sum_j w_j^2 = \|\mathbf{w}\|_2^2$ to avoid too large weights in OLS

Controlling the Complexity through Regularisation

Solution 1: hard constraint

regularisation = complexity is fixed ($\Omega(\boldsymbol{\theta}) = \omega$) or constrained ($\Omega(\boldsymbol{\theta}) \leq \omega$)

$$\mathbf{w} = \arg\min \frac{1}{n} \left(\mathbf{X}\mathbf{w} - \mathbf{t}\right)^T \left(\mathbf{X}\mathbf{w} - \mathbf{t}\right) \quad \text{s.t.} \quad \|\mathbf{w}\|_0 \leq p$$

Solution 2: penalisation

regularisation = complexity is penalised proportionally to $\Omega(\boldsymbol{\theta})$

$$\mathbf{w} = \arg\min \frac{1}{n} \left(\mathbf{X}\mathbf{w} - \mathbf{t}\right)^T \left(\mathbf{X}\mathbf{w} - \mathbf{t}\right) + C\|\mathbf{w}\|_2^2$$

C = regularisation constant = metaparameter that controls the complexity

- $C = 0 \Rightarrow$ standard OLS solution with no complexity control
- $C \to \infty \Rightarrow$ overpenalised solution with almost zero weights

Controlling the Complexity through Regularisation

Solution 1: hard constraint

regularisation = complexity is fixed ($\Omega(\boldsymbol{\theta}) = \omega$) or constrained ($\Omega(\boldsymbol{\theta}) \leq \omega$)

$$\mathbf{w} = \arg\min \frac{1}{n} \left(\mathbf{X}\mathbf{w} - \mathbf{t}\right)^T \left(\mathbf{X}\mathbf{w} - \mathbf{t}\right) \quad \text{s.t.} \quad \|\mathbf{w}\|_0 \leq p$$

Solution 2: penalisation

regularisation = complexity is penalised proportionally to $\Omega(\boldsymbol{\theta})$

$$\mathbf{w} = \arg\min \frac{1}{n} \left(\mathbf{X}\mathbf{w} - \mathbf{t}\right)^T \left(\mathbf{X}\mathbf{w} - \mathbf{t}\right) + C\|\mathbf{w}\|_2^2$$

C = regularisation constant = metaparameter that controls the complexity

- $C = 0 \Rightarrow$ standard OLS solution with no complexity control
- $C \to \infty \Rightarrow$ overpenalised solution with almost zero weights

Regularising Linear Models

L0: Feature Selection

Sparse vs. Nonsparse Models

What is sparsity (an why it matters)

sparse model = model with many (most) zero parameters

- easier to interpret (fewer features to consider for model analysis)
- reduce overfitting (fewer degrees of freedom to fit model on data)
- reduce computational cost (e.g. in text processing where $d \gg 10,000$)

Feature selection (= L0 regularisation)

goal = find the best p features (that maximise the model performance)

$$\mathbf{w} = \arg\min \frac{1}{n} \left(\mathbf{X}\mathbf{w} - \mathbf{t}\right)^T \left(\mathbf{X}\mathbf{w} - \mathbf{t}\right) \quad \text{s.t.} \quad \|w\|_0 = p$$

NP hard problem: C_p^d combinations of features $\Rightarrow$ greedy approaches

- forward search: start from empty set of selected features, then grow
- backward search: start from full set of features, then eliminate

Sparse vs. Nonsparse Models

What is sparsity (an why it matters)

sparse model = model with many (most) zero parameters

- easier to interpret (fewer features to consider for model analysis)
- reduce overfitting (fewer degrees of freedom to fit model on data)
- reduce computational cost (e.g. in text processing where $d \gg 10,000$)

Feature selection (= L0 regularisation)

goal = find the best p features (that maximise the model performance)

$$\mathbf{w} = \arg\min \frac{1}{n} \left(\mathbf{X}\mathbf{w} - \mathbf{t}\right)^T \left(\mathbf{X}\mathbf{w} - \mathbf{t}\right) \quad \text{s.t.} \quad \|w\|_0 = p$$

NP hard problem: C_p^d combinations of features $\Rightarrow$ greedy approaches

- forward search: start from empty set of selected features, then grow
- backward search: start from full set of features, then eliminate

L2: Ridge Regression

Preventing Overfitting

L2 regularisation

model with large parameter values = likely to be overfitting

- L2 penalisation = penalise large weight values
- prevent overfitting if regularisation constant is large enough

Ridge regression (= L2 regularisation)

goal = find the best model with moderate weight amplitudes

$$\mathbf{w} = \arg\min \frac{1}{n} (\mathbf{X}\mathbf{w} - \mathbf{t})^T (\mathbf{X}\mathbf{w} - \mathbf{t}) + C\|w\|_2^2 = \left(\frac{1}{n}\mathbf{X}^T\mathbf{X} + C\mathbf{I}_d\right)^{-1} \mathbf{X}^T\mathbf{t}$$

nonsparse solution: bias towards weights of similar amplitude vs. most small weights + few large weights (small weights are ≈ 0, but $\neq 0$)

- ridge regression $\approx$ adding a constant C to diagonal terms of $\frac{1}{n}\mathbf{X}^T\mathbf{X}$
- consequence: a small L2 regularisation improves numerical stability

Preventing Overfitting

L2 regularisation

model with large parameter values = likely to be overfitting

- L2 penalisation = penalise large weight values
- prevent overfitting if regularisation constant is large enough

Ridge regression (= L2 regularisation)

goal = find the best model with moderate weight amplitudes

$$\mathbf{w} = \arg\min \frac{1}{n} (\mathbf{X}\mathbf{w} - \mathbf{t})^T (\mathbf{X}\mathbf{w} - \mathbf{t}) + C\|w\|_2^2 = \left(\frac{1}{n}\mathbf{X}^T\mathbf{X} + C\mathbf{I}_d\right)^{-1} \mathbf{X}^T\mathbf{t}$$

nonsparse solution: bias towards weights of similar amplitude vs. most small weights + few large weights (small weights are ≈ 0, but $\neq 0$)

- ridge regression $\approx$ adding a constant C to diagonal terms of $\frac{1}{n}\mathbf{X}^T\mathbf{X}$
- consequence: a small L2 regularisation improves numerical stability

L1: LASSO / LARS

Relaxing the L0 Regularisation

L1 regularisation

L0 regularisation = NP hard $\Rightarrow$ obtain sparse models with L1 regularisation

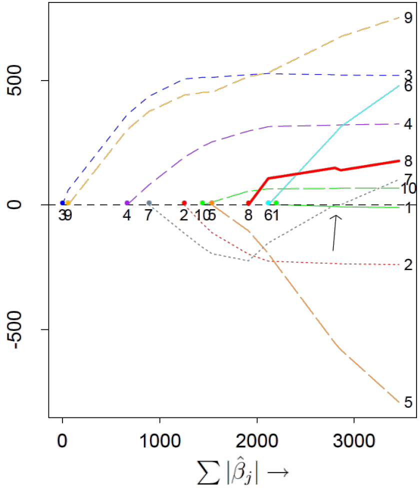
$$\mathbf{w} = \arg\min \frac{1}{n} \left(\mathbf{X}\mathbf{w} - \mathbf{t}\right)^T \left(\mathbf{X}\mathbf{w} - \mathbf{t}\right) + C \sum_{j=1}^{d} |w_j|$$

no analytical solution $\Rightarrow$ least angle regression (LARS) algorithm

The LARS procedure works roughly as follows. As with classic Forward Selection, we start with all coefficients equal to zero, and find the predictor most correlated with the response, say x_{j_1}. We take the largest step possible in the direction of this predictor until some other predictor, say x_{j_2}, has as much correlation with the current residual. At this point LARS parts company with Forward Selection. Instead of continuing along x_{j_1}, LARS proceeds in a direction equiangular between the two predictors until a third variable x_{j_3} earns its way into the "most correlated" set. LARS then proceeds equiangularly between x_{j_1}, x_{j_2} and x_{j_3}, i.e. along the "least angle direction", until a fourth variable enters, etc.

Efron, B., Hastie, T., Johnstone, I., Tishirani, R. Least Angle Regression. Annals of Statistics 32 p. 407–499, 2004.

Application: Diabetes Progression



What are the **best features**?

3 body mass index (BMI)

9 serum measurement #5

4 blood pressure (BP)

7 serum measurement #3

2 sex

10 serum measurement #6

5 serum measurement #1

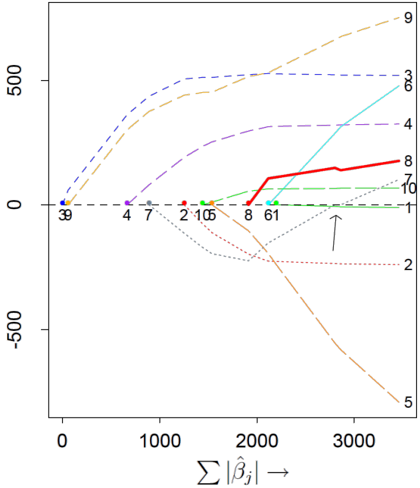
8 serum measurement #4

6 serum measurement #2

1 age

Efron, B., Hastie, T., Johnstone, I., Tishirani, R. Least Angle Regression. Annals of Statistics 32 p. 407–499, 2004.

Application: Diabetes Progression



What are the **1 best features**?

3 body mass index (BMI)

9 serum measurement #5

4 blood pressure (BP)

7 serum measurement #3

2 sex

10 serum measurement #6

5 serum measurement #1

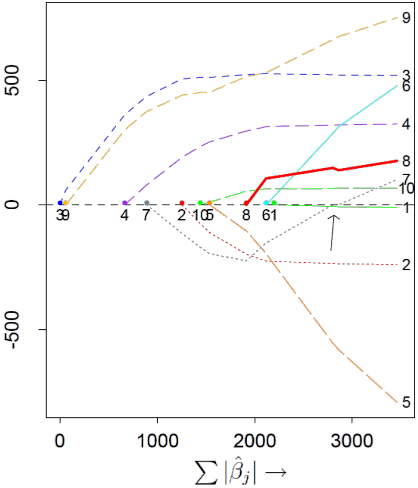
8 serum measurement #4

6 serum measurement #2

1 age

Efron, B., Hastie, T., Johnstone, I., Tishirani, R. Least Angle Regression. Annals of Statistics 32 p. 407–499, 2004.

Application: Diabetes Progression



What are the **2 best features**?

3 body mass index (BMI)

9 serum measurement #5

4 blood pressure (BP)

7 serum measurement #3

2 sex

10 serum measurement #6

5 serum measurement #1

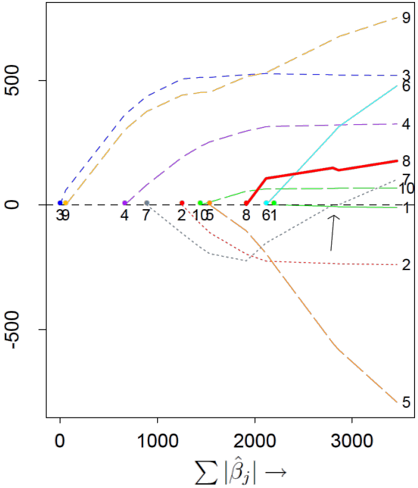
8 serum measurement #4

6 serum measurement #2

1 age

Efron, B., Hastie, T., Johnstone, I., Tishirani, R. Least Angle Regression. Annals of Statistics 32 p. 407–499, 2004.

Application: Diabetes Progression



What are the **3 best features**?

- 3 body mass index (BMI)
- 9 serum measurement #5
- 4 blood pressure (BP)
- 7 serum measurement #3
- 2 sex
- 10 serum measurement #6
- 5 serum measurement #1
- 8 serum measurement #4
- 6 serum measurement #2
- 1 age

Efron, B., Hastie, T., Johnstone, I., Tishirani, R. Least Angle Regression. Annals of Statistics 32 p. 407–499, 2004.

L1/2: Elastic Net

Combining the Strengths of L1 and L2 regularisation

L1 vs. L2 regularisation

- L2 penalisation prevents large weight to occur in OLS solution
- but ridge regression does not obtain sparse vector of weights
- L1 penalisation enforces sparse weights in linear regression
- but groups of colinear features are not correctly handled
- and LARS cannot use more than n features (biomedical applications)

Elastic net (= L1/2 regularisation)

goal = compromise between L1 (LARS) and L2 (ridge) regularisation

$$\mathbf{w} = \arg\min \frac{1}{n} (\mathbf{X}\mathbf{w} - \mathbf{t})^T (\mathbf{X}\mathbf{w} - \mathbf{t}) + C \left(\alpha \|w\|_1 + (1 - \alpha) \|w\|_2^2 \right)$$

Combining the Strenghts of L1 and L2 regularisation

L1 vs. L2 regularisation

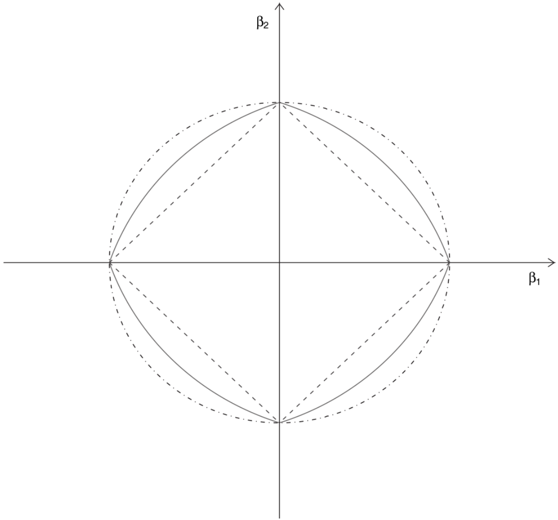
- L2 penalisation prevents large weight to occur in OLS solution
- but ridge regression does not obtain sparse vector of weights
- L1 penalisation enforces sparse weights in linear regression
- but groups of colinear features are not correctly handled
- and LARS cannot use more than n features (biomedical applications)

Elastic net (= L1/2 regularisation)

goal = compromise between L1 (LARS) and L2 (ridge) regularisation

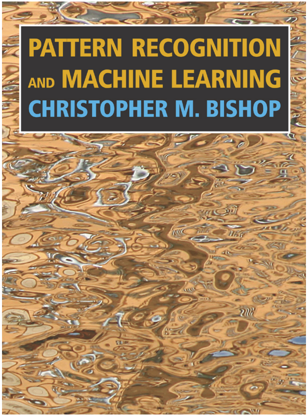
$$\mathbf{w} = \arg\min \frac{1}{n} \left(\mathbf{X}\mathbf{w} - \mathbf{t}\right)^T \left(\mathbf{X}\mathbf{w} - \mathbf{t}\right) + C \left(\alpha \|w\|_1 + (1-\alpha)\|w\|_2^2\right)$$

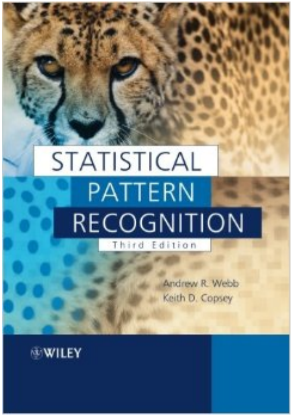
Geometric Interpretation of L1, L2 and L1/2 Regularisation



Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67(2), 301-320.

References


PATTERN RECOGNITION
AND MACHINE LEARNING
CHRISTOPHER M. BISHOP


STATISTICAL
PATTERN
RECOGNITION
Third Edition
Andrew R. Webb
Keith D. Copsey
WILEY