Machine Learning

MIRI Master

Lluís A. Belanche

belanche@cs.upc.edu

Soft Computing Research Group

Soft Computing Research Group
Departament de Ciències de la Computació (Computer Science Department)

Universitat Politècnica de Catalunya - Barcelona Tech

FACULTAT D'INFORMÀTICA DE BARCELONA

Spring Semester 2017-2018

LECTURE 3: Linear regression and beyond

Linear regression and beyond

Outline

1. Reminder of the regression framework

2. An example worked out

3. Leaping forward: basis functions and the SVD

4. Regularized least squares: ridge and the LASSO

5. Conclusions

Linear regression and beyond

Reminder of the regression framework

- The departing statistical **model** is

$$t_n = f(\boldsymbol{x}_n) + \varepsilon_n \qquad \boldsymbol{x} \in \mathbb{R}^d,\ t \in \mathbb{R}$$

where ε is a continuous r.v. such that $\mathbb{E}[\varepsilon_n] = 0$ and $\mathrm{Var}[\varepsilon_n] = \sigma^2$

- Let's assume again that we further **model** $\varepsilon_n \sim N(0, \sigma^2)$ and:

$$f(\boldsymbol{x}) \approx y(\boldsymbol{x}; \boldsymbol{\beta}) = \beta_0 + \sum_{i=1}^{d} \beta_i x_i = \boldsymbol{\beta}^\top \boldsymbol{x}$$

with $\boldsymbol{x} = (1, x_1, \ldots, x_d)^\top$ and $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_d)^\top$

Linear regression and beyond

Reminder of the regression framework

Suppose we have an i.i.d. sample of N *labelled* observations $\mathcal{D} = \{(\boldsymbol{x}_n, t_n)\}_{n=1,\ldots,N}$, where $\boldsymbol{x}_n \in \mathbb{R}^d, t_n \in \mathbb{R}$

Therefore our **statistical model** is $t_n \sim N(y(\boldsymbol{x}_n; \boldsymbol{\beta}), \sigma^2)$ or:

$$p(t_n|\boldsymbol{x}_n; \theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}\left(t_n - \boldsymbol{\beta}^\top \boldsymbol{x}_n\right)^2\right),$$

with (unknown) **parameters** $\theta := \{\beta_0, \beta_1, \ldots, \beta_d, \sigma^2\}$.

Linear regression and beyond

Reminder of the regression framework

Put $\boldsymbol{t} = (t_1, \ldots, t_N)^\top$ and $X_{N\times(d+1)}$ the matrix of the $\boldsymbol{x}_n$. Define the **likelihood** as $\mathcal{L}(\theta) := P(\boldsymbol{t}|X;\theta)$

Let us maximize the "log-likelihood":

$$
\begin{aligned}
l(\theta) &:= \ln \mathcal{L}(\theta) = \ln \prod_{n=1}^{N} p(t_n|\boldsymbol{x}_n;\theta) = \sum_{n=1}^{N} \ln p(t_n|\boldsymbol{x}_n;\theta) \\
&= -\frac{N}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{n=1}^{N}\left(t_n - \boldsymbol{\beta}^\top \boldsymbol{x}_n\right)^2 \\
&= -\frac{N}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}(\boldsymbol{t} - X\boldsymbol{\beta})^\top(\boldsymbol{t} - X\boldsymbol{\beta}) \\
&= -\frac{N}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\|\boldsymbol{t} - X\boldsymbol{\beta}\|^2
\end{aligned}
$$

Linear regression and beyond

Reminder of the regression framework

$$\frac{\partial l}{\partial \boldsymbol{\beta}} = -\frac{1}{2\sigma^2}(-2X^\top \boldsymbol{t} + 2X^\top X\boldsymbol{\beta}) = \mathbf{0}$$

$$\frac{\partial l}{\partial \sigma^2} = -\frac{N}{2\sigma^2} + \frac{1}{2\sigma^4}(\boldsymbol{t} - X\boldsymbol{\beta})^\top(\boldsymbol{t} - X\boldsymbol{\beta}) = \mathbf{0}$$

Therefore,

$$\hat{\boldsymbol{\beta}} = (X^\top X)^{-1}X^\top \boldsymbol{t}$$

$$\hat{\sigma}^2 = \frac{1}{N}(\boldsymbol{t} - X\hat{\boldsymbol{\beta}})^\top(\boldsymbol{t} - X\hat{\boldsymbol{\beta}}) = \frac{1}{N}\|\boldsymbol{t} - X\hat{\boldsymbol{\beta}}\|^2$$

Linear regression and beyond

Reminder of the regression framework

- Note $\widehat{\sigma}^2 = R_{\text{emp}}(y_{\mathcal{D}})$, which turns out to be a biased estimator of σ^2; an unbiased estimator is:

$$\bar{\sigma}^2 = \frac{N}{N-d}\,\widehat{\sigma}^2$$

- It is also known that $\hat{\boldsymbol{\beta}}$ is an unbiased estimator of $\boldsymbol{\beta}$ and

$$\text{Var}[\hat{\boldsymbol{\beta}}] = (X^\top X)^{-1}\,\sigma^2$$

Which implies that $\hat{\boldsymbol{\beta}} \sim N\Big(\boldsymbol{\beta}, (X^\top X)^{-1}\,\sigma^2\Big)$

Linear regression and beyond

Reminder of the regression framework

- The matrix $X^\dagger := (X^\top X)^{-1} X^\top$ is known as the Moore-Penrose **pseudo-inverse** of X

- It is the generalization of the notion of an inverse matrix to non-square matrices

- It has the property that $X^\dagger X = I$ (although in general $XX^\dagger \neq I$) (note however that both $X^\dagger X$ and $XX^\dagger$ are symmetric)

Linear regression and beyond

Reminder of the regression framework

Theorem. Let $X_{N\times M}$ with $N > M$. If the column vectors of X are linearly independent, *i.e.*, if rank$(X) = M$, then

1. the matrix $X^\top X$ is symmetric and positive definite (p.d.) –in particular, it is non-singular

2. the least-squares problem

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^M} \|\boldsymbol{t} - X\boldsymbol{\beta}\|^2, \qquad \text{has a unique solution}$$

3. this solution can be found solving the so-called Gauss' normal equations $(X^\top X)\boldsymbol{\beta} = X^\top \boldsymbol{t}$ for $\boldsymbol{\beta}$

Linear regression and beyond

An example

The comet Tentax was discovered in 1968 and follows a quadratic orbit (elliptic, parabolic or hyperbolic) according to Kepler's laws.

The orbit has equation:

$$r = \frac{p}{1 - e\,\mathsf{cos}\,\varphi}$$

where p is a comet-specific coefficient, e is the comet's eccentricity (both unknown) and the (r, φ) pairs indicate measured positions (in polar coordinates centered at the Sun).

Astronomers have gathered a set of coordinates:

$\{(2{,}70, 48^\circ), (2{,}00, 67^\circ), (1{,}61, 83^\circ), (1{,}20, 108^\circ), (1{,}02, 126^\circ)\}$

Goal: estimate the two constants p, e based on measured data

Linear regression and beyond

An example

First we write the relation in linear form: $r - (r \cos\varphi)e = p$ from which we arrive at the following system:

$$\left.\begin{array}{lrcl} p & +1{,}806e & = & 2,70 \\ p & +0{,}782e & = & 2,00 \\ p & +0{,}196e & = & 1,61 \\ p & -0{,}371e & = & 1,20 \\ p & -0{,}600e & = & 1,02 \end{array}\right\}$$

which we express as:

$$X \cdot \boldsymbol{\beta} = \boldsymbol{t} \qquad \text{or} \qquad \begin{pmatrix} 1 & 1{,}806 \\ 1 & 0{,}782 \\ 1 & 0{,}196 \\ 1 & -0{,}371 \\ 1 & -0{,}600 \end{pmatrix} \cdot \begin{pmatrix} p \\ e \end{pmatrix} = \begin{pmatrix} 2,70 \\ 2,00 \\ 1,61 \\ 1,20 \\ 1,02 \end{pmatrix}$$

Linear regression and beyond

An example

Solving the normal equations $(X^\top X)\boldsymbol{\beta} = X^\top \boldsymbol{t}$

as $\widehat{\boldsymbol{\beta}} = (X^\top X)^{-1} X^\top \boldsymbol{t}$

we obtain:

$$\widehat{\boldsymbol{\beta}} = \begin{pmatrix} p \\ e \end{pmatrix} \approx \begin{pmatrix} 1,454 \\ 0,694 \end{pmatrix}$$

Linear regression and beyond

Quality of the fit

- In statistics, $-2l = -2\ln\mathcal{L}$ is called the **deviance**

- In ML, this is the **square error**:

  $N\ln(2\pi\sigma^2) + \frac{1}{\sigma^2}\|\boldsymbol{t} - X\widehat{\boldsymbol{\beta}}\|^2$, that is to say $\|\boldsymbol{t} - X\widehat{\boldsymbol{\beta}}\|^2$

- A much better quantity to report is the **NRMSE**:

$$\mathrm{NRMSE}(\widehat{\boldsymbol{\beta}}) = \sqrt{\frac{\|\boldsymbol{t} - X\widehat{\boldsymbol{\beta}}\|^2}{(N-1)\mathrm{Var}[\boldsymbol{t}]}}$$

In statistics, $R^2 = 1 - \mathrm{NRMSE}^2$ is the proportion of the (target) variability *explained* by the model

Linear regression and beyond

Leaping forward

We say that a model is **linear** if its parameters play a linear role in the model.

Example

$$y(x;\boldsymbol{\beta}) = \beta_0 + \sum_{j=1}^{d} \beta_j x^j = \beta_0 + \beta_1 x + \ldots + \beta_d x^d, \qquad x \in \mathbb{R}$$

is a polynomial on x but a linear model!

Linear regression and beyond

Leaping forward

A simple but powerful idea is the introduction of **basis functions**:

$$y(\boldsymbol{x};\boldsymbol{w}) = w_0 + \sum_{j=1}^{M} w_j \phi_j(\boldsymbol{x}) = \sum_{j=0}^{M} w_j \phi_j(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x})$$

where $\phi_0(\boldsymbol{x}) = 1$.

$\boldsymbol{\phi}(\boldsymbol{x}) = (1, \phi_1(\boldsymbol{x}), \ldots, \phi_M(\boldsymbol{x}))^\top$ and $\boldsymbol{w} = (w_0, w_1, \ldots, w_M)^\top$.

The basis function expansion above is still a **linear model**.

Linear regression and beyond

Leaping forward

Example:

$$y(x;\boldsymbol{w}) = w_0 + \sum_{j=1}^{M} w_j\phi_j(x) = \sum_{j=0}^{M} w_j\phi_j(x) = \boldsymbol{w}^\top\boldsymbol{\phi}(x)$$

where $\phi_0(x) = \sqrt{T^{-1}}$ and M is even

$$\phi_j(x) = \begin{cases} \sqrt{2T^{-1}}\,\mathsf{sin}(a_j x), \quad a_j = (j+1)\pi T^{-1} & \textbf{if } j \text{ is odd} \\ \sqrt{2T^{-1}}\,\mathsf{cos}(a_j x), \quad a_j = j\pi T^{-1} & \textbf{if } j \text{ is even} \end{cases}$$

This y is a truncated Fourier series in $[0, T]$. If T is known, it is a linear model for x; what if T is unknown (then a parameter)?

Linear regression and beyond

Leaping forward

Define $\boldsymbol{t} = (t_1, \ldots, t_N)^\top$ the vector of targets
$\Phi_{N\times(M+1)}$ the matrix of the $\boldsymbol{\phi}(\boldsymbol{x}_n)$

where $\Phi_{ij} = \phi_j(\boldsymbol{x}_i), \qquad i = 1, \ldots, N, j = 0, \ldots, M.$

$$\Phi = \begin{pmatrix} 1 & \phi_1(\boldsymbol{x}_1) & \phi_2(\boldsymbol{x}_1) & \ldots & \phi_M(\boldsymbol{x}_1) \\ 1 & \phi_1(\boldsymbol{x}_2) & \phi_2(\boldsymbol{x}_2) & \ldots & \phi_M(\boldsymbol{x}_2) \\ \ldots & \ldots & \ldots & \ddots & \vdots \\ 1 & \phi_1(\boldsymbol{x}_N) & \phi_2(\boldsymbol{x}_N) & \ldots & \phi_M(\boldsymbol{x}_N) \end{pmatrix}$$

Linear regression and beyond

Leaping forward

Let us maximize the new log-likelihood:

1. The Gauss' normal equations are $(\Phi^\top \Phi)\boldsymbol{w} = \Phi^\top \boldsymbol{t}$

2. The solution of which is

$$\hat{\boldsymbol{w}} = (\Phi^\top \Phi)^{-1} \Phi^\top \boldsymbol{t} = \Phi^\dagger \boldsymbol{t}$$

   and

$$\widehat{\sigma}^2 = \frac{1}{N} (\boldsymbol{t} - \Phi \hat{\boldsymbol{w}})^\top (\boldsymbol{t} - \Phi \hat{\boldsymbol{w}}) = \frac{1}{N} \|\boldsymbol{t} - \Phi \hat{\boldsymbol{w}}\|^2$$

Linear regression and beyond

Singular Value Decomposition

The direct computation of the pseudo-inverse of Φ has two major drawbacks:

1. When M is large, $(\Phi^\top\Phi)$ is a large $(M+1) \times (M+1)$ matrix; then the required inverse $(\Phi^\top\Phi)^{-1}$ can be costly

2. If $(\Phi^\top\Phi)$ is singular –or close to– then the required inverse $(\Phi^\top\Phi)^{-1}$ can be impossible –or numerically delicate

Linear regression and beyond

Singular Value Decomposition

Theorem. Every matrix $X_{N\times M}$ can be expressed as:

$$X = U\Delta V^\top$$

$U_{N\times N}, V_{M\times M}$ are **orthogonal** matrices ($U^\top U = I_N, V^\top V = I_M$)
$\Delta_{N\times M}$ is a **diagonal** matrix

The columns of U are the **eigenvectors** of $XX^\top$
The columns of V are the **eigenvectors** of $X^\top X$

Linear regression and beyond

Singular Value Decomposition

1. Let rank$(X) = r \leq$ mín(N, M). Then exactly r elements λ_k in the diagonal of Δ are strictly positive; the remaining elements are 0

2. These $\lambda_k > 0$ are called the **singular values** and correspond to the square roots of the positive **eigenvalues** of $XX^\top$ (same as $X^\top X$)

3. Sometimes an "economy size" decomposition is delivered:

   If X is $N \times M$ with $N > M$, then only the first M columns of U are given and Δ is $M \times M$

Linear regression and beyond

SVD example

$$X = \begin{pmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \\ 7 & 8 \end{pmatrix}$$

$$U = \begin{pmatrix} -0{,}1525 & -0{,}8226 & -0{,}3945 & -0{,}3800 \\ -0{,}3499 & -0{,}4214 & 0{,}2428 & 0{,}8007 \\ -0{,}5474 & -0{,}0201 & 0{,}6979 & -0{,}4614 \\ -0{,}7448 & 0{,}3812 & -0{,}5462 & 0{,}0407 \end{pmatrix}$$

$$\Delta = \begin{pmatrix} 14{,}2691 & 0 \\ 0 & 0{,}6268 \\ 0 & 0 \\ 0 & 0 \end{pmatrix}$$

$$V = \begin{pmatrix} -0{,}6414 & 0{,}7672 \\ -0{,}7672 & -0{,}6414 \end{pmatrix}$$

SVD economy size example

$$X = \begin{pmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \\ 7 & 8 \end{pmatrix}$$

$$U = \begin{pmatrix} -0{,}1525 & -0{,}8226 \\ -0{,}3499 & -0{,}4214 \\ -0{,}5474 & -0{,}0201 \\ -0{,}7448 & 0{,}3812 \end{pmatrix}$$

$$\Delta = \begin{pmatrix} 14{,}2691 & 0 \\ 0 & 0{,}6268 \end{pmatrix}$$

$$V = \begin{pmatrix} -0{,}6414 & 0{,}7672 \\ -0{,}7672 & -0{,}6414 \end{pmatrix}$$

Linear regression and beyond

The SVD for least squares

Given the least-squares problem

$$\min_{\boldsymbol{w}\in\mathbb{R}^M} \|\boldsymbol{t} - X\boldsymbol{w}\|^2$$

the solution can be obtained with the SVD as:

1. Compute the economy size SVD of $X = U\Delta V^\top$
2. Solve for $\boldsymbol{w}$ as $\hat{\boldsymbol{w}} = V\text{diag}\left(\lambda_k^{-1}\right)U^\top\boldsymbol{t}$, where only the $\lambda_k > 0$ are considered

Linear regression and beyond

Regularized least squares

The maximum likelihood framework can yield unstable parameter estimates, especially when:

1. The explanatory variables are highly correlated

2. There is an insufficient number of observations (N) relative to the number of predictors (basis functions $M + 1$ or dimensions $d + 1$)

Linear regression and beyond

Regularized least squares

In the context of regression with Gaussian noise (square error), it is quite common to penalize the parameter vector:

1. Define the **penalized empirical error** as:

$$R_{\text{emp}}(y(\cdot;\boldsymbol{w})) := \|\boldsymbol{t} - \Phi\boldsymbol{w}\|^2 + \lambda\|\boldsymbol{w}\|^2, \qquad \lambda > 0$$

2. Set the derivative w.r.t. $\boldsymbol{w}$ to $\mathbf{0}$: $(-2\Phi^\top\boldsymbol{t} + 2\Phi^\top\Phi\boldsymbol{w}) + 2\lambda\boldsymbol{w} = \mathbf{0}$

3. Therefore, $\hat{\boldsymbol{w}} = (\Phi^\top\Phi + \lambda I)^{-1}\Phi^\top\boldsymbol{t}$

Linear regression and beyond

Regularized least squares

- This is known as Tikhonov or L_2 **regularization** in ML
- Perhaps it is best known as **ridge regression** in statistics, where it is usually explained as a "penalized log-likelihood"
- It can also be derived from Bayesian statistics arguments
- Advantages:

  1. Pushing the length of the parameter vector $\|\boldsymbol{w}\|$ to 0 allows the fit to be under explicit control with the regularization parameter λ
  2. The matrix $\Phi^\top \Phi$ is positive semi-definite (p.s.d.); therefore $\Phi^\top \Phi + \lambda I$ is guaranteed to be p.d. (hence non-singular), $\forall \lambda > 0$

Linear regression and beyond

Regularized least squares

Yes, nice, but ...

- How to do the **explicit control** on the fit?
  - regularization permits the specification of models that are more complex than needed because it limits the effective complexity
  - instead of trial-and-error on complexity, we can set a large complexity and adjust the λ

Linear regression and beyond

Regularized least squares

Yes, **very** nice, but ...

- How to set the value of λ? Using LOOCV, because
  - in this case λ is a very forgiving parameter (we usually perform a log search)
  - there is a closed efficient formula for the LOOCV (for **linear models**)

Linear regression and beyond

Regularized least squares

1. Choose a (large) set of values Λ

2. For every $\lambda \in \Lambda$,

   a) Solve for $\hat{\boldsymbol{w}} = (\Phi^\top \Phi + \lambda I)^{-1} \Phi^\top \boldsymbol{t}$

   b) Compute the "hat matrix" $H := \Phi\Phi^\dagger = \Phi(\Phi^\top \Phi + \lambda I)^{-1} \Phi^\top$

   c) Compute the LOOCV of $y(\cdot) = \widehat{\boldsymbol{w}}^\top \phi(\cdot)$ in $\mathcal{D}$ as

$$\mathrm{LOOCV}(y) = \frac{1}{N} \sum_{n=1}^{N} \left(\frac{t_n - \widehat{\boldsymbol{w}}^\top \boldsymbol{\phi}(\boldsymbol{x}_n)}{1 - h_{nn}} \right)^2$$

3. Choose the model with the lowest LOOCV

Linear regression and beyond

Regularized least squares

A very popular method is GCV (**generalized cross-validation**):

$$\mathrm{GCV}(y) = \frac{1}{N} \frac{\sum_{n=1}^{N} \left(t_n - \widehat{\boldsymbol{w}}^\top \phi(\boldsymbol{x}_n)\right)^2}{\left(1 - \frac{\mathrm{Tr}(H)}{N}\right)^2}$$

which is a more stable computation for the LOOCV

Note that λ is needed to compute both $\widehat{\boldsymbol{w}}$ and H

Linear regression and beyond

LASSO regression

The LASSO (Least Absolute Shrinkage and Selection Operator) regression is L_1-regularized Linear regression

The choice for the regularizer is $\|\boldsymbol{w}\|_1$ and we get:

$$R_{\text{emp}}(y(\cdot;\boldsymbol{w})) = \|\boldsymbol{t} - \Phi\boldsymbol{w}\|^2 + \tau\|\boldsymbol{w}\|_1, \qquad \tau > 0$$

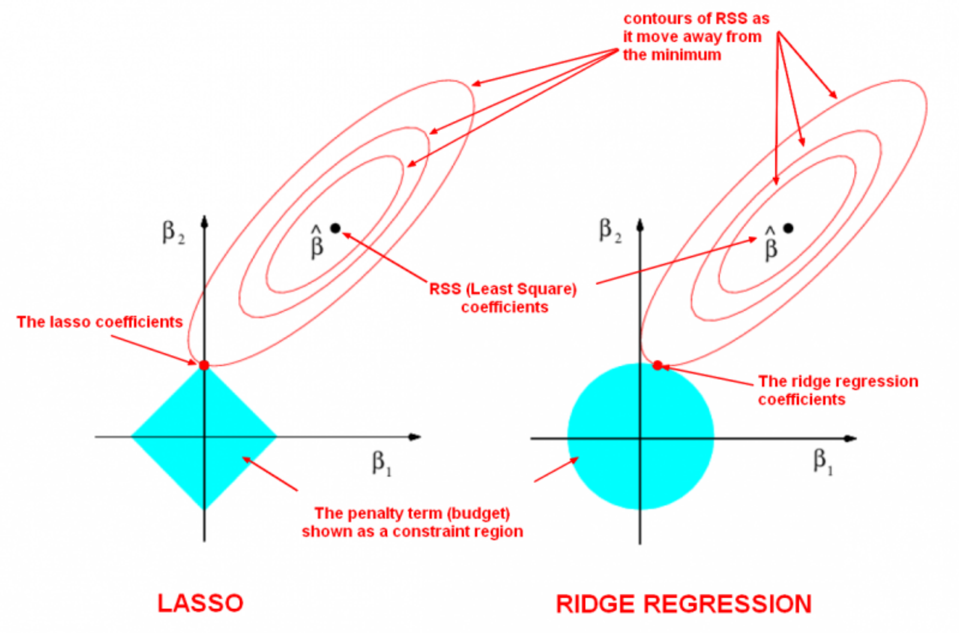
which turns out to be equivalent to

$$R_{\text{emp}}(y(\cdot;\boldsymbol{w})) = \|\boldsymbol{t} - \Phi\boldsymbol{w}\|^2, \text{ subject to } \|\boldsymbol{w}\|_1 \leq \tau$$

Linear regression and beyond

LASSO regression

In ridge regression, as the penalty λ is increased, all coefficients are reduced while still remaining non-zero; in the LASSO, increasing the τ penalty causes more and more of the coefficients to be driven to zero



As d increases, the multidimensional diamond has an increasing number of corners, and so it is highly likely that some coefficients will be set equal to zero. Hence, the LASSO performs **shrinkage** and therefore **feature selection**

Linear regression and beyond

LASSO regression

- The LASSO loss function is no longer quadratic, but is still convex
- The LASSO is a special **quadratic programming** (QP) problem, for which the **Least Angle Regression** (LARS) procedure is used
- It exploits the special structure of the problem, and provides an efficient way to compute the solutions for all possible values of $\tau > 0$ (**regularization path**)

Linear regression and beyond

Conclusions

We have introduced **linear models** as linear combinations of non-linear **basis functions** (BF)

ADVANTAGES:

1. We can represent non-linear functions of the data using linear fitting techniques; we have the freedom to choose the form of the BFs

2. The fit can be under tight explicit control by regularization

3. The computations can be very efficient, no need to refit for LOOCV

4. Interpretability of the model is rather high

Linear regression and beyond

Conclusions

LIMITATIONS:

The most important weak point is in the basis functions!

1. Many interesting basis functions scale very poorly with dimension (polynomials, Fourier series, splines, ...)

2. Our BFs are not flexible (they are independent of the data)

3. As a consequence, their number may be very high, which in turn leads to unstability (because of low significance of the coefficients)

Linear regression and beyond

Conclusions

The solution is to develop **basis functions with parameters** that

1. ... scale well with dimension (inner products, distances, ...)

2. ... are data dependent (because of the parameters)

3. As a consequence, their number may be much lower (and the coefficients be significant)

4. Unfortunately, the new parameters play a non-linear role in the model: their optimization is plagued with local optima

Machine Learning

Syllabus

1. Introduction to Machine Learning

2. Theoretical issues (I): regression

3. Linear regression and beyond

4. Theoretical issues (II): classification

5. Generative classifiers

6. Discriminative classifiers

7. Clustering: k-means and E-M

8. Learning with kernels (I): The SVM

9. Learning with kernels (II): Kernel functions

10. Artificial neural networks (I): Delta rule, MLP-1

11. Artificial neural networks (II): MLP-2, RBF

12. Artificial neural networks (III): DL and CNNs

13. Ensemble methods: Random Forests

14. Advanced topics and frontiers