

Foundations of Machine Learning and AI

Theodoros Evgeniou - Nicolas Vayatis

Sessions 3-4: From classical statistics to Machine Learning

Summary of sessions 1-2

From previous session

Main messages

- 1 Machine Learning is about choosing (= estimating = learning) a function from a set of functions (called hypothesis space)

From information theory: Number of steps to find the unknown function among K functions is of the order $\log K$

- 2 Estimation-Approximation trade-off

In the case of linear models in dimension d estimated with LS: corresponds to Bias-Variance decomposition where the variance term is proportional to d/n

- 3 Mathematical tool from probability theory

ERM/Finite case/Nonzero error: deviation inequality of the error of a function on training data from the expected out-of-sample error of the same function (Hoeffding's inequality)

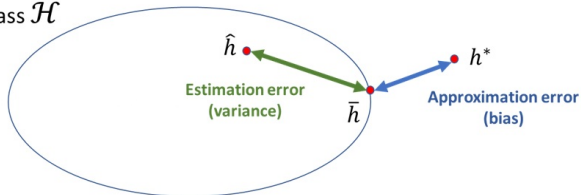
From previous session

Estimation vs. Approximation

- Error measure $L(h)$ for functions $h \in \mathcal{H}$
- Best in the class \bar{h} , world-champion h^*
- Key decomposition of error for any estimate \hat{h} :

$$L(\hat{h}) - L(h^*) = \underbrace{L(\hat{h}) - L(\bar{h})}_{\text{estimation (stochastic)}} + \underbrace{L(\bar{h}) - L(h^*)}_{\text{approximation (deterministic)}}$$

Hypothesis class \mathcal{H}



From previous session

The case of linear models

- Linear model in dimension d with vector notations (sample size n):

$$\mathbf{Y} = \mathbf{X}\beta^* + \varepsilon \in \mathbb{R}^n$$

where \mathbf{X} of size $n \times d$

- Least square estimator denoted by $\hat{\beta}_n$
- error of the LSE (expectation wrt the sample distribution)

$$\frac{1}{n} \mathbb{E}(\|\mathbf{X}\beta^* - \mathbf{X}\hat{\beta}_n\|^2) = \text{Bias} + \sigma^2 \frac{d}{n}$$

This session

- Handling models in high dimensions:

$$d \gg 1, d \gg n$$

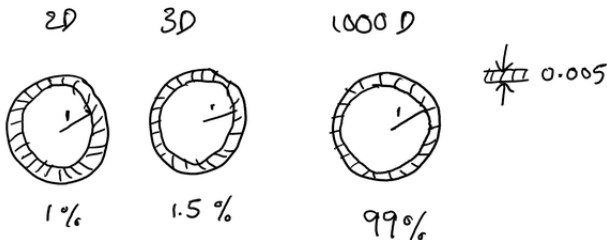
- Questioning dimensionality:

number of parameters vs. complexity of set of functions

High Dimensional Interlude

Some surprising fact

Spherical shells



- Ratio shell/volume:

$$\frac{\text{vol}(B_d(0, 1) - B_d(0, 1 - \varepsilon))}{\text{vol}(B_d(0, 1))} = 1 - (1 - \varepsilon)^d \rightarrow 1 \text{ when } d \rightarrow \infty$$

High Dimensional Interlude

Readings

- High level position paper:

"High Dimensional Data Analysis : The Curses and Blessings of Dimensionality" by D. Donoho (2000)

- Maths book:

"High-Dimensional Probability" by Roman Vershynin (2018)

From classical statistics to Machine Learning: Handling high dimensions

- A. Sparsity and linear models
- B. Estimating nonlinear functions
- C. Generalizations

A. Sparsity and linear models

Tuning the dimension of the model

Linear regression model

Notations

- Vector notations:

Response vector $\mathbf{Y} \in \mathbb{R}^n$, input data matrix \mathbf{X} (size $n \times d$)

- Linear model with vector notations:

$$\mathbf{Y} = \mathbf{X}\beta^* + \varepsilon$$

where ε random noise vector (centered, independent of \mathbf{X})

The sparse linear regression model

- Intuition: what if there are uninformative variables in the model but we do not know which they are?
- Sparsity assumption: Let β^* the true parameter with support

$$m^* = \{j : \beta_j^* \neq 0\} \subset \{1, \dots, d\}$$

- ℓ_0 norm of any β : $\|\beta\|_0 = \sum_{j=1}^d \mathbb{I}\{\beta_j \neq 0\}$

Two possible formulations Constrained vs. Penalized optimization

- ① Ivanov formulation: take k between 0 and $\min\{n, d\}$

$$\min_{\beta \in \mathbb{R}^d} \|\mathbf{Y} - \mathbf{X}\beta\|_2^2 \quad \text{subject to } \|\beta\|_0 \leq k$$

- ② Tikhonov formulation: take $\lambda > 0$

$$\min_{\beta \in \mathbb{R}^d} \{ \|\mathbf{Y} - \mathbf{X}\beta\|_2^2 + \lambda \|\beta\|_0 \}$$

Comments

- Tikhonov looks as a Lagrange formulation of Ivanov
- But here the two formulations are NOT equivalent due to the lack of smoothness of the ℓ_0 norm
- Ivanov with ℓ_0 constraint is known as the Best Subset Selection problem for which there are algorithms based on heuristics (e.g. Forward Stagewise Regression) which work ok up to $k \simeq 35$. Recent advances: check Mixed Integer Optimization (MIO) formulation by Bertsimas et al. (2016).
- Focus on Tikhonov regularization from now on

Sparsity and linear models

Model selection

Connecting the dots

Tikhonov penalty and variance

Recall:

- Tikhonov formulation with ℓ_0 *penalty*: take $\lambda > 0$

$$\min_{\beta \in \mathbb{R}^d} \{ \|\mathbf{Y} - \mathbf{X}\beta\|_2^2 + \lambda \|\beta\|_0 \} \quad (1)$$

- Bias-variance decomposition of the error for the LSE $\hat{\beta}$:

$$\frac{1}{n} \mathbb{E}(\|\mathbf{X}\beta^* - \mathbf{X}\hat{\beta}_n\|^2) = \text{Bias} + \sigma^2 \frac{d}{n} \quad (2)$$

where d is the dimension of the model space and σ^2 is the variance of the noise

Questions for now: does the bias-variance decomposition (2) explains (1)? Is the penalty correct?

Model selection in linear models

- Model: $\mathbf{Y} = \mathbf{X}\beta^* + \varepsilon$
- Consider a model for β^* that is a subset m of indices of $\{1, \dots, d\}$
- Example: In dimension $d = 3$, we have:
 - 1 model of size $|m| = 0$: constant model
 - 3 models of size $|m| = 1$: $\{1\}, \{2\}, \{3\}$
 - 3 models of size $|m| = 2$: $\{1, 2\}, \{2, 3\}, \{1, 3\}$
 - 1 model of size $|m| = 3$: $\{1, 2, 3\}$

We potentially have 8 versions of Least Square Estimator, we call constrained LSE (except for the case $|m| = 3$ which is unconstrained).

Model selection in linear models

- Model: $\mathbf{Y} = \mathbf{X}\beta^* + \varepsilon$
- Consider the set \mathcal{M} of subsets m of the variables among indices $\{1, \dots, d\}$
- For every $m \in \mathcal{M}$, there is a standard linear regression model with dimension $k_m = |m|$. In other words, for those $j \notin m$, we have $\beta_j^* = 0$.
- For each model $m \in \mathcal{M}$, compute the constrained Least Square Estimator $\hat{\beta}_m$.
- The final estimator is the "best" among $\hat{\beta}_m$ over all $m \in \mathcal{M}$

What "Best" actually means

The oracle

- Error given by: $r_m = \frac{1}{n} \mathbb{E}(\|\mathbf{X}\beta^* - \mathbf{X}\hat{\beta}_m\|^2)$
- Oracle estimator:

$$\hat{\beta}_{\bar{m}} \quad \text{where} \quad \bar{m} = \arg \min_{m \in \mathcal{M}} r_m$$

- Problem: oracle estimator is NOT an empirical quantity (depends on the unknown β^*)

Akaike Information Criterion (1/3)

Derivation

- Recall from last session: error of estimator

$$r_m = \frac{1}{n} \mathbb{E}(\|\mathbf{X}\beta^* - \mathbf{X}\hat{\beta}_m\|^2) = \frac{1}{n} \mathbb{E}(\|(I_n - \hat{\Pi}_m)\mathbf{X}\beta^*\|^2) + \sigma^2 \frac{|m|}{n}$$

- Similarly, we can derive:

$$\frac{1}{n} \mathbb{E}(\|\mathbf{Y} - \mathbf{X}\hat{\beta}_m\|^2) = \frac{1}{n} \mathbb{E}(\|(I_n - \hat{\Pi}_m)\mathbf{X}\beta^*\|^2) + \sigma^2 \frac{(n - |m|)}{n}$$

- Then, we observe:

$$\frac{1}{n} \mathbb{E}(\|\mathbf{Y} - \mathbf{X}\hat{\beta}_m\|^2) = r_m + \sigma^2 \frac{(n - 2|m|)}{n}$$

Link with last class notations: $\mathbf{h}^* = \mathbf{X}\beta^*$, $\hat{\mathbf{h}}_m = \mathbf{X}\hat{\beta}_m = \hat{\Pi}_m \mathbf{Y}$ where $\hat{\Pi}_m$ is the orthogonal projection on S_m

Akaike Information Criterion (2/3)

Empirical estimator of the error

- We have obtained that:

$$r_m = \frac{1}{n} \mathbb{E}(\|\mathbf{Y} - \mathbf{X}\hat{\beta}_m\|^2) + \sigma^2 \frac{(2|m| - n)}{n}$$

- Unbiased estimator of the error (assuming known variance):

$$\hat{r}_m = \frac{1}{n} \|\mathbf{Y} - \mathbf{X}\hat{\beta}_m\|^2 + \sigma^2 \frac{(2|m| - n)}{n}$$

- Akaike Information Criterion

$$\hat{m} = \arg \min_{m \in \mathcal{M}} \left\{ \|\mathbf{Y} - \mathbf{X}\hat{\beta}_m\|^2 + 2|m|\sigma^2 \right\}$$

Akaike Information Criterion (3/3)

The right penalty? Discussion

- Tikhonov regularization for ℓ_0 norm is equivalent to AIC with $\lambda = 2\sigma^2$ in this case
- In practice, AIC does not pick the right dimension: in high dimensions, \hat{r}_m fluctuates around r_m due to a large amount of models with same cardinality $|m|$
- The right penalty should be of the order $2\sigma^2|m| \log(d)$

NB: The number of linear models of given size $|m|$ in dimension d is: $\binom{d}{|m|} \leq \exp(|m|(1 + \log(d/|m|)))$

AIC in large dimensions

- When d is large, is this practical ?
- There are about $e^{(d/2)}$ models to scan in the worst case where $|m| \simeq d/2 \dots$

A. Sparsity and linear models

From (mathematical) statistics to optimization

Solving the computation burden

The power of convexity

- Practical methods for model selection are essentially greedy heuristics consisting in adding and/or retrieving one variable at the time to explore part of the whole model space which is exponential in the dimension. Examples are: Forward Stagewise Regression, Forward-Backward algorithm...
- Question: would it be possible to solve the optimization wrt the unknown parameter β AND wrt to its support subset of indices jointly?
- Answer is yes at the cost of the so-called relaxation of the non-convex formulation with the ℓ_0 penalty to a convexified problem with an ℓ_1 penalty.

The LASSO for linear models

From ℓ_0 to ℓ_1

- Consider the relaxation of the previous problem replacing the ℓ_0 -norm by the ℓ_1 -norm:

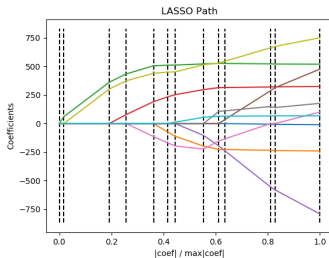
$$\|\beta\|_1 = \sum_{j=1}^d |\beta_j|$$

- The new estimator is called the LASSO: for any $\lambda > 0$,

$$\hat{\beta}_\lambda \in \arg \min_{\beta \in \mathbb{R}^d} \{ \|\mathbf{Y} - \mathbf{X}\beta\|^2 + \lambda \|\beta\|_1 \}$$

Blessings of the LASSO

- Approximate solutions via efficient algorithms building the so-called regularization paths $\lambda \rightarrow \hat{\beta}_\lambda$:



- Theoretical soundness: it can be shown that: as $n, d \rightarrow \infty$

$$\frac{1}{n} \mathbb{E}(\|\mathbf{X}\beta^* - \mathbf{X}\hat{\beta}\|^2) \leq C \|\beta^*\|_1 \sqrt{\frac{\log d}{n}}$$

The "mother" of ML algorithms

Penalized optimization

- Learning process as the optimization of a data-dependent criterion:

$$\text{Criterion}(h) = \text{Training error}(h) + \lambda \text{ Penalty}(h)$$

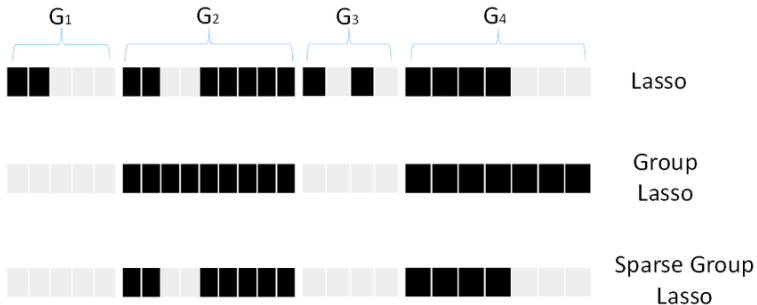
- Training error: data-fitting term related to a loss function
- Penalty: complexity of the decision function
- Constant λ : smoothing parameter tuned through cross-validation procedure

A. Sparsity and linear models

Structured sparsity

Putting human priors in penalties

Sparsity patterns



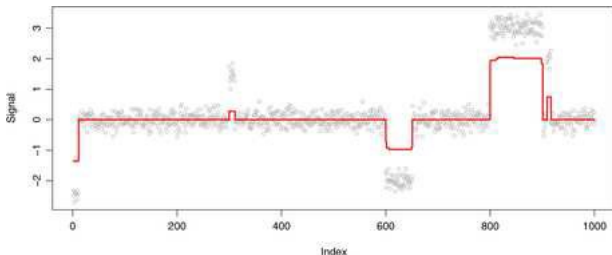
The simplest structured penalty

Group LASSO

- Group structure on the parameter β^* : let G the number of groups of subsets of indices in $\{1, \dots, d\}$ and, for $g = 1, \dots, G$, we denote by $\mathbf{X}^{(g)}$ the submatrix of \mathbf{X} with variables in group g and by $\beta^{(g)}$ the coefficient vector applied to variables in group g and d_g is the size of group g .
- Group LASSO formulation:

$$\hat{\beta}_\lambda \in \arg \min_{\beta \in \mathbb{R}^d} \left\{ \|\mathbf{Y} - \mathbf{X}\beta\|^2 + \lambda \sum_{g=1}^G \sqrt{d_g} \|\beta^{(g)}\| \right\}$$

Case of temporal patterns Fused LASSO



- Enforcing temporal coherence leads to adding a penalty term:

$$\hat{\beta}_{\lambda} \in \arg \min_{\beta \in \mathbb{R}^d} \left\{ \|\mathbf{Y} - \mathbf{X}\beta\|^2 + \lambda \|\beta\|_1 + \mu \sum_{j=2}^d |\beta_j - \beta_{j-1}| \right\}$$

A. Sparsity and linear models

Ridge regression

Penalized optimization

Other penalties?

- Until now: hypothesis class with linear functions $h \in \mathcal{H}$ and variations on sparsity-inducing penalties

$$\text{Criterion}(h) = \text{Training error}(h) + \lambda \text{Penalty}(h)$$

- This idea goes back to the 60s (Ivanov, John, Lavrent'ev, Tikhonov) where the penalty operated as a regularizer of solutions for ill-posed problems.

Ill-posed problem in statistics

High dimensional least square regression

- Assume d larger than n
- Then when solving the gradient cancellation equations in the least square optimization problem, we observe that we have less equations than variables: this is the case of an *underdetermined* linear system.
- Another way to put this is to observe that $\mathbf{X}^T \mathbf{X}$ is not full rank, hence it is not invertible and there is an infinity of solutions.

The oldest regularizer in statistics

Ridge regression

- The Ridge estimator is the solution of the following penalized optimization problem: for any $\lambda > 0$,

$$\hat{\beta}_{\lambda} \in \arg \min_{\beta \in \mathbb{R}^d} \{ \|\mathbf{Y} - \mathbf{X}\beta\|^2 + \lambda \|\beta\|_2^2 \}$$

Derivation of ridge regression estimator

- We denote the objective function:

$$F(\beta) = (\mathbf{Y} - \mathbf{X}\beta)^T (\mathbf{Y} - \mathbf{X}\beta) + \lambda\beta^T \beta$$

- Thanks to convexity and differentiability of F , we obtain the solution by solving

$$\nabla F(\beta) = 2\mathbf{X}^T(\mathbf{X}\beta - \mathbf{Y}) + 2\lambda\beta = 0$$

- Solution:

$$\hat{\beta}_\lambda = (\mathbf{X}^T\mathbf{X} + \lambda I_d)^{-1} \mathbf{X}^T\mathbf{Y}$$

because $\mathbf{X}^T\mathbf{X} + \lambda I_d$ always invertible.

- Computation still painful for d very large...

Dual optimization problem

Formulation and KKT conditions

- Equivalent formulation of ridge regression optimization:

$$\min_{\beta \in \mathbb{R}^d, r \in \mathbb{R}^n} \left\{ \frac{1}{2} \|r\|^2 + \frac{\lambda}{2} \|\beta\|^2 \right\} \quad \text{subject to } r = \mathbf{X}\beta - \mathbf{Y}$$

- Lagrange formulation with multiplier vector α

$$\mathcal{L}(\beta, r, \alpha) = \frac{1}{2} \|r\|^2 + \frac{\lambda}{2} \|\beta\|^2 + \alpha^T (r - \mathbf{X}\beta + \mathbf{Y})$$

- Karush-Kuhn-Tucker conditions: zeroing gradient wrt primal variables β, r , leads to:

$$\beta(\alpha) = \frac{1}{\lambda} \mathbf{X}^T \alpha \quad \text{and} \quad r(\alpha) = -\alpha$$

Dual optimization problem Resolution

- Then, an equivalent formulation of ridge regression optimization is given by:

$$\mathcal{L}(\beta(\alpha), r(\alpha), \alpha) = \frac{1}{2}\|\alpha\|^2 + \frac{1}{2\lambda}\|\mathbf{X}^T\alpha\| + \alpha^T \left(-\alpha - \frac{1}{\lambda}\mathbf{X}\mathbf{X}^T\alpha + \mathbf{Y} \right)$$

- Solution:

$$\hat{\alpha}_n = \lambda \left(\mathbf{X}\mathbf{X}^T + \lambda I_n \right)^{-1} \mathbf{Y} \quad \text{and} \quad \hat{\beta}_n = \frac{1}{\lambda} \mathbf{X}^T \hat{\alpha}_n$$

- Then use the identity:
 $\mathbf{X}^T (\mathbf{X}\mathbf{X}^T + \lambda I_n)^{-1} = (\mathbf{X}^T \mathbf{X} + \lambda I_d)^{-1} \mathbf{X}^T$ to check the solutions are the same.
- Important observation! Optimization and function evaluation only require the pairwise scalar product between x 's and data points X_i 's

Elastic Net

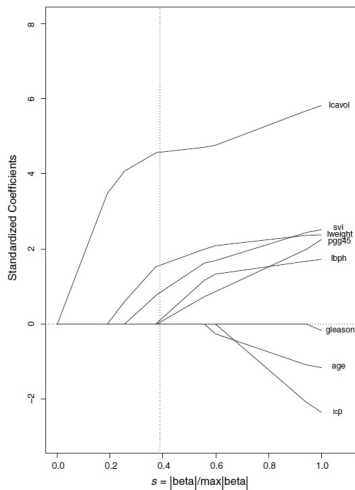
The best of LASSO and Ridge?

- Rationale (from [Zou and Hastie, 2005])
 - (a) In the $p > n$ case, the lasso selects at most n variables before it saturates, because of the nature of the convex optimization problem. This seems to be a limiting feature for a variable selection method. Moreover, the lasso is not well defined unless the bound on the L_1 -norm of the coefficients is smaller than a certain value.
 - (b) If there is a group of variables among which the pairwise correlations are very high, then the lasso tends to select only one variable from the group and does not care which one is selected. See Section 2.3.
 - (c) For usual $n > p$ situations, if there are high correlations between predictors, it has been empirically observed that the prediction performance of the lasso is dominated by ridge regression (Tibshirani, 1996).
- Combination of ℓ_1 and ℓ_2 penalties

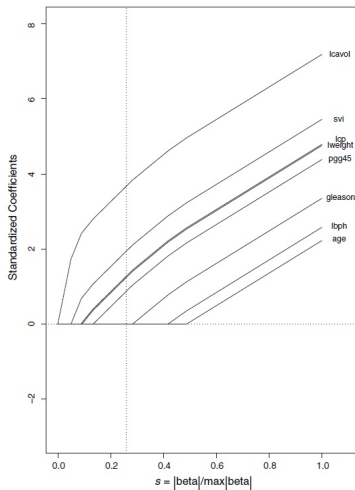
$$\hat{\beta}_\lambda \in \arg \min_{\beta \in \mathbb{R}^d} \{ \|\mathbf{Y} - \mathbf{X}\beta\|^2 + \lambda \|\beta\|_1 + \mu \|\beta\|_2^2 \}$$

LASSO vs. Elastic Net

Comparison of regularization paths



(a)



(b)

Tuning the hyperparameters

Cross-validation

- How do we select the parameters λ and μ ? These are called hyperparameters or smoothing parameters or regularization parameters.
- This is a universal problem in monitoring the overfitting effect of Machine Learning methods.
- The procedure of cross-validation will be developed later in the course.

Exercise

Comparison of the three penalties

- Consider the following toy problem: $Y \sim \mathcal{N}_1(\beta^*, 1)$ where β is a real-valued parameter ($d = 1$).
- Find the three estimators when minimizing the following three functions:

$$(i) \frac{1}{2}(Y - \beta)^2 + \lambda, \quad (ii) \frac{1}{2}(Y - \beta)^2 + \lambda|\beta|, \quad (iii) \frac{1}{2}(Y - \beta)^2 + \lambda\beta^2$$

- Show a plot of the estimators as functions of the unconstrained LSE and explain the use of the following terminology for the penalized procedures: hard thresholding, soft thresholding, shrinkage.

B. Estimating nonlinear functions

From nonlinear to linear

Polynomial regression example

- Consider a polynomial regression in dimension $d = 2$: this corresponds to a linear model of dimension $d' = 7$ with feature vector:

$$\Phi(x_1, x_2) = (1, \sqrt{2}x_1, \sqrt{2}x_2, \sqrt{2}x_1x_2, x_1^2, x_2^2)^T$$

- Note that:

$$\Phi(x)^T \Phi(x') = (x^T x' + 1)^2$$

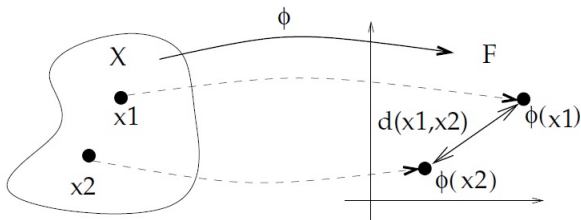
- We call $K(x, x') = (x^T x' + 1)^2$ a polynomial kernel. A kernel has the property to be represented as a scalar product in a high dimensional feature space. The feature space is the image of the original input space of dimension d through Φ . The feature space can be of huge dimension.

Examples of basic kernels

- Linear: $K(x, z) = x \cdot z$
- Polynomial: $K(x, z) = (x \cdot z)^d$ or $K(x, z) = (1 + x \cdot z)^d$
- Gaussian: $K(x, z) = \exp \left[-\frac{\|x-z\|^2}{2 \sigma^2} \right]$
- Laplace Kernel: $K(x, z) = \exp \left[-\frac{\|x-z\|}{2 \sigma^2} \right]$

How a kernel defines a metric

Definition



$$\begin{aligned}d_K(x_1, x_2)^2 &= \|\phi(x_1) - \phi(x_2)\|_{\mathcal{H}}^2 \\&= \langle \phi(x_1) - \phi(x_2), \phi(x_1) - \phi(x_2) \rangle_{\mathcal{H}} \\&= \langle \phi(x_1), \phi(x_1) \rangle_{\mathcal{H}} + \langle \phi(x_2), \phi(x_2) \rangle_{\mathcal{H}} - 2 \langle \phi(x_1), \phi(x_2) \rangle_{\mathcal{H}} \\d_K(x_1, x_2)^2 &= K(x_1, x_1) + K(x_2, x_2) - 2K(x_1, x_2)\end{aligned}$$

Kernel engineering

- Specific kernels have been design to process structured data such as strings (text, DNA sequence...)
- Example of spectrum kernel used for DNA sequences:

Kernel definition

- The 3-spectrum of

$\mathbf{x} = \text{CGGS LIAMMWFGV}$

is:

$(\text{CGG}, \text{GGS}, \text{GSL}, \text{SLI}, \text{LIA}, \text{IAM}, \text{AMM}, \text{MMW}, \text{MWF}, \text{WFG}, \text{FGV})$.

- Let $\Phi_u(\mathbf{x})$ denote the number of occurrences of u in \mathbf{x} . The k -spectrum kernel is:

$$K(\mathbf{x}, \mathbf{x}') := \sum_{u \in \mathcal{A}^k} \Phi_u(\mathbf{x}) \Phi_u(\mathbf{x}') .$$

Kernel machine learning estimation

Is it doable?

- Nice modeling properties of kernel functions
- The question is whether the penalized optimization in the sense of least squares is feasible?

The magic of kernels

Kernel ridge regression

- In the linear case of ridge regression, we have seen that the only data-dependent quantities that matter in both problem formulation and evaluation of predictions are the pairwise scalar products of $X_i^T X_j$ and $x^T X_i$.
- We can basically replace any scalar product by the kernel evaluation of the considered pair without changing at all the algorithmic complexity of resolution. We are then able to estimate the parameters α_i of nonlinear functions of the form:

$$f(x) = \sum_{i=1}^n \alpha_i K(x, X_i)$$

C. Generalizations

Application to other estimation problems

Penalized optimization

What about other variations?

- Until now: hypothesis class with linear functions $h \in \mathcal{H}$ and variations on the penalties

$$\text{Criterion}(h) = \text{Training error}(h) + \lambda \text{Penalty}(h)$$

- From now on: play with other losses affects the Training error

$$\text{Criterion}(h) = \text{Training error}(h) + \lambda \text{Penalty}(h)$$

Using other loss functions

A few examples:

Ridge regression:

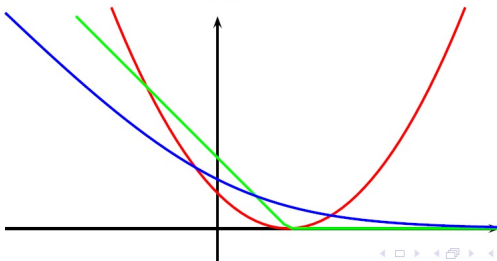
$$\min_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^n \frac{1}{2} (y_i - \beta^\top \mathbf{x}_i)^2 + \lambda \|\beta\|_2^2.$$

Linear SVM:

$$\min_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^n \max(0, 1 - y_i \beta^\top \mathbf{x}_i) + \lambda \|\beta\|_2^2.$$

Logistic regression:

$$\min_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^n \log(1 + e^{-y_i \beta^\top \mathbf{x}_i}) + \lambda \|\beta\|_2^2.$$



C. Generalizations

From vectors to matrices

Motivation

Recommender systems

- One million big ones!
- Given 100 million ratings on a scale of 1 to 5, predict 3 million ratings to highest accuracy



- 17770 total movies
- 480189 total users
- Over 8 billion total ratings

Matrix completion

Problem formulation

	Item 1	Item 2	...	Item 99	Item 100
Customer 1	5	NA	...	NA	3
Customer 2	NA	2	...	3	NA
...
Customer 49	2	3	...	NA	4
Customer 50	1	NA	...	NA	NA

- **PROBLEM:** Find the matrix of lowest rank has the specified entries

$$\begin{aligned} & \text{minimize} && \text{rank}(\mathbf{X}) \\ & \text{subject to} && X_{ij} = M_{ij} \quad \forall (i, j) \in \Omega \end{aligned}$$

- **When is this problem easy?**
 - Which algorithms?
 - Which sampling sets?
 - Which low-rank matrices?

Matrix completion

Discussion

- Can be used for other matrix estimation problems such as: community detection on social networks, covariance matrix estimation in portfolio management, ...
- Efficient solutions have been developed through relaxation of the initial problem
- Relies on linear algebra techniques (matrix norms, SVD...)
- Interesting to consider structural assumptions on matrices, tensors, depending on the physics of the problem

Next sessions

- Back to theory!
- Back to sparsity!
- Other methods