école normale supérieure paris–saclay

université PARIS-SACLAY

Introduction to Statistical Learning

Nicolas Vayatis

Lecture 4 - Regularization, stability, ML algorithms

Course overview

- Introduction

  Demystification / Learning and information / Setup

- Chapter 1 : Optimality in statistical learning

  Probabilistic view / Performance criteria / Optimal elements

- **Chapter 2 : Mathematical foundations of statistical learning**

  Concentration inequality / Complexity measures / **Regularization and stability**

- Chapter 3 : Consistency of mainstream machine learning methods

  Boosting, SVM, Neural networks / Bagging, Random forests

Chapter 2 - Mathematical tools

A. Probability inequalities
B. Complexity measures
C. Regularization and stability –> today's lecture

Regularization in linear models

Regression model
Basic theory

- Random pair : (X, Y) over $\mathbb{R}^d \times \mathbb{R}$
- Decision rules : $f : \mathbb{R}^d \to \mathbb{R}$
- Least-square prediction error : $R(f) = \mathbb{E}(Y - f(X))^2$
- Optimal predictor : $f^*(x) = \mathbb{E}(y|x)$
- ERM over a class $\mathcal{F}$ of decision rules given a sample $\{(X_i, Y_i) : i = 1, \ldots, n\}$:

$$\hat{f} = \arg\min_{f \in \mathcal{F}} \sum_{i=1}^{n} (Y_i - f(X_i))^2$$

Linear regression model

- Vector notations :

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

- Linear model with vector notations :

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

  where ε random noise vector (centered, independent of X)

- When $\mathrm{rank}(X) = d$, then the ERM is given by :

$$\hat{f}(x) = \hat{\beta}^T x \quad \text{where} \quad \hat{\beta} = (X^T X)^{-1} X^T Y = \hat{\Pi}_d Y$$

  where $\hat{\Pi}_d$ projection matrix over the columns of X in $\mathbb{R}^n$

Linear regression model
Limitations

- When $\mathrm{rank}(X) < d$ (e.g. $d > n$), then for solution $\hat{\beta}$ and $b \in \ker(X)$, we have that $\hat{\beta} + b$ is also a solution
- As a consequence :
  (a) coefficients cannot benefit of interpretation
  (b) out-of-sample predictions are not unique (while in-sample predictions are unique)
- Solution : assume that effective dimension of the linear model is smaller than d (and n !)

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 which only a subset of variables (called *support*)

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

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

Two possible formulations
Constrained vs. Penalized optimization

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

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

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

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

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

Connecting the dots
Tikhonov penalty and variance

Recall :

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

$$\min_{\beta \in \mathbb{R}^d} \left\{ \|Y - X\beta\|_2^2 + \lambda \|\beta\|_0 \right\} \tag{1}$$

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

$$\frac{1}{n}\mathbb{E}\left(\|X\beta^* - X\widehat{\beta}_n\|^2\right) = \sigma^2 \frac{d}{n} \tag{2}$$

  where d is the dimension of the data and σ^2 is the variance of the Gaussian noise

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

Model selection in linear models

- Model : $\mathsf{Y} = \mathsf{X}\beta^* + \varepsilon$
- 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 (LSE), we call call constrained LSE (except for the case $|m| = 3$ which is unconstrained).

Model selection in linear models

- Model : $\mathsf{Y} = \mathsf{X}\beta^* + \varepsilon$
- Consider the set $\mathcal{M}$ of subsets m of the variables among indices $\{1, \ldots, d\}$. There are 2^d such sets m.
- 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 $\widehat{\beta}_m$.
- The final estimator is the "best" among $\widehat{\beta}_m$ over all $m \in \mathcal{M}$

What "Best" actually means
The oracle

- In-sample error given by : $r_m = \frac{1}{n}\mathbb{E}\left(\|\mathsf{X}\beta^* - \mathsf{X}\widehat{\beta}_m\|^2\right)$
- Best theoretical estimator (called *oracle*) :

$$\widehat{\beta}_{\overline{m}} \quad \text{where} \quad \overline{m} = \underset{m\in\mathcal{M}}{\arg\min}\, r_m$$

- Example of an empirical estimator : Akaike Information Criterion (AIC penalty of Least Squares)

$$\widehat{m} = \underset{m\in\mathcal{M}}{\arg\min}\left\{\|\mathsf{Y} - \mathsf{X}\widehat{\beta}_m\|^2 + 2|m|\sigma^2\right\}$$

(can be computed from data as long as σ^2 is assumed to be known)

Optional material

Derivation of Akaike Information Criterion

Akaike Information Criterion (1/2)
Derivation

- Recall from least square bias-variance decomposition in linear models : error of estimator

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

with $\mathrm{X}\widehat{\beta}_m = \widehat{\Pi}_m Y$ where $\widehat{\Pi}_m$ is the orthogonal projection on the subspace S_m generated with the subset m of variables

- Similarly, we can derive :

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

- Then, we observe :

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

Akaike Information Criterion (2/2)
Empirical estimator of the error

- We have obtained that :

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

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

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

- Akaike Information Criterion

$$\widehat{m} = \underset{m \in \mathcal{M}}{\arg\min} \left\{ \|\mathsf{Y} - \mathsf{X}\widehat{\beta}_m\|^2 + 2|m|\sigma^2 \right\}$$

End of optional material

Bottom line on AIC

Is AIC an optimal penalty for model selection in linear models ?

- Tikhonov regularization for ℓ_0 norm is equivalent to AIC with $\lambda = 2\sigma^2$ in this case (λ also depends on n if we minimize the average square error on the data)
- In practice, AIC does not pick the right dimension : in high dimensions, $\widehat{r}_m$ fluctuates around r_m due to a large amount of models with same cardinality $|m|$
- The correct 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 :

$$\binom{d}{|m|} \leq \exp\left(|m|(1+\log(d/|m|))\right)$$

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 $|m| \simeq d/2$...

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 :

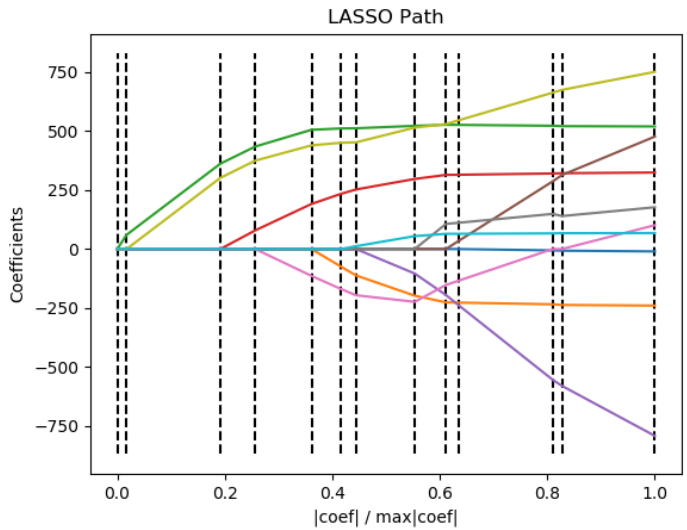
$$\|\beta\|_1 = \sum_{j=1}^{d} |\beta_j|$$

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

$$\widehat{\beta}_\lambda \in \underset{\beta \in \mathbb{R}^d}{\arg\min} \left\{ \|\mathsf{Y} - \mathsf{X}\beta\|^2 + \lambda \|\beta\|_1 \right\}$$

Blessings of the LASSO

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



- Theoretical soundness : it can be shown that : as $n, d \to \infty$, in-sample error resists to curse of dimensionality

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

(holds for the constrained formulation)

Penalized least-squares in linear regression

- LASSO

$$\hat{\beta}_\lambda \in \arg\min_{\beta \in \mathbb{R}^d} \left\{ \frac{1}{n} \|Y - X\beta\|^2 + \lambda \|\beta\|_1 \right\}$$

- Ridge regression

$$\hat{\beta}_\lambda \in \arg\min_{\beta \in \mathbb{R}^d} \left\{ \frac{1}{n} \|Y - X\beta\|^2 + \lambda \|\beta\|_2^2 \right\}$$

- Structured sparsity with $\|\beta\|_S$ being a sparsity inducing norm (group LASSO, graph LASSO, ...)

$$\hat{\beta}_\lambda \in \arg\min_{\beta \in \mathbb{R}^d} \left\{ \frac{1}{n} \|Y - X\beta\|^2 + \lambda \|\beta\|_S \right\}$$

The "mother" of shallow ML algorithms

From classical statistics to Machine Learning

"Shallow Learning"

- **Shallow Learning** are algorithms which will only depend on very few hyperparameters beyond the λ.
- **Deep Learning** relies on many architectural hyperparameters (e.g., number of layers, nodes, etc - see Sessions 9-10) whose calibration is a very complex optimization problem.
- The theory of supervised Machine Learning *should* apply to both shallow and deep learning.

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 or function norms (e.g. LASSO)
- Constant λ : smoothing parameter tuned through cross-validation procedure

How to create shallow ML algorithms ?

- Standard function classes (e.g. linear functions) and risk (e.g. least squares) and variations on the penalties

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

(e.g. in least square minimization : LASSO, Group LASSO, Elastic Net, Fused LASSO, structures penalties...)

- Playing with losses changing the training error

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

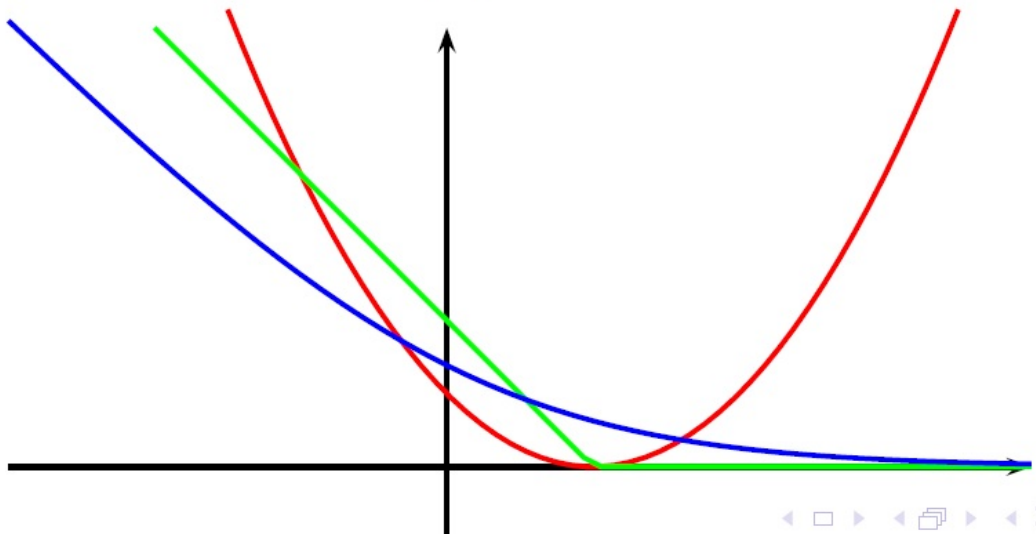
Changing loss functions

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



The principle of Structural risk minimization (SRM)

- Given a training set of size n and the corresponding empirical error $\widehat{L}_n$, consider the ERM principle over an increasing sequence of hypothesis classes $\mathcal{H}_1 \subset \ldots \mathcal{H}_j \subset \ldots$ of increasing complexity (e.g. dimension in linear models, VC dimension in nonlinear models)

SRM leads to penalized ERM

- In order to achieve the estimation-approximation ("bias-variance") trade-off, the idea is to penalize the empirical risk with a complexity term :

$$\widehat{h}_n^{\text{SRM}} = \widehat{h}_{\widehat{j},n}^{\text{ERM}}$$

where :

$$\widehat{h}_{j,n}^{\text{ERM}} = \underset{h \in \mathcal{H}_j}{\arg\min}\, \widehat{L}_n(h)$$

and

$$\widehat{j} = \underset{j \geq 1}{\arg\min} \left\{ \widehat{L}_n(\widehat{h}_{j,n}^{\text{ERM}}) + \lambda\, \text{pen}(n, \text{complexity}(\mathcal{H}_j)) \right\}$$

where $\lambda = \lambda_n$ is called the regularization parameter, or smoothing parameter.

SRM example : Regularization in decision trees

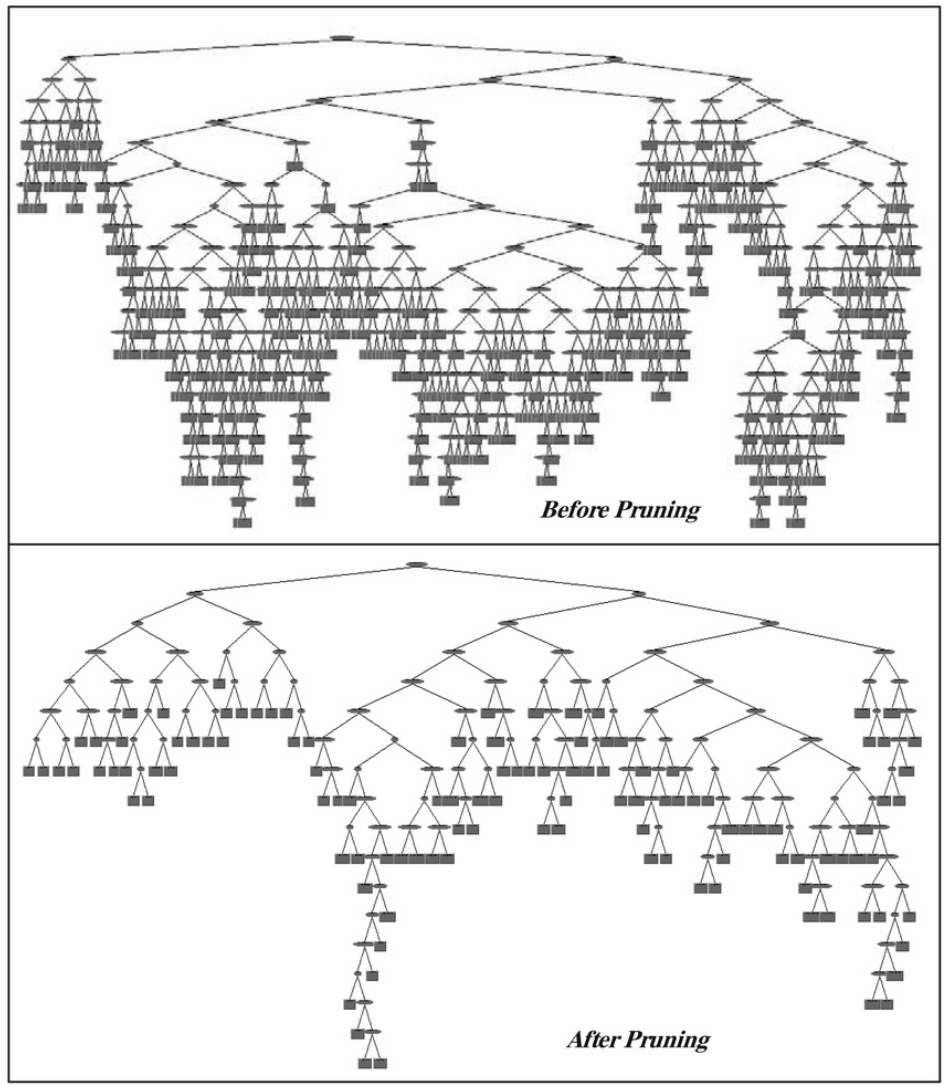
Building a complexity calibrated decision tree involves two steps :

1. Growing a decision tree - output : a tree classifier $\hat{h}_\pi$ with data-dependent partition $\widehat{\pi}$ (which may overfit the data !)
2. Pruning the tree - optimization over all subpartitions (subtrees) a penalized criterion of the form

$$\arg\min_{\pi \subset \widehat{\pi}} \hat{L}_n(\hat{h}_\pi) + \lambda|\pi|$$

where $\widehat{h}_\pi$ is a tree classifier obtained with the training data based on a partition π which is a subpartition of $\widehat{\pi}$

Example of original and pruned decision tree



Other examples of regularized formulations

- Linear (soft) SVM (hinge loss, L2 penalty)

$$\hat{\beta}_\lambda \in \underset{\beta \in \mathbb{R}^d}{\arg\min} \left\{ \frac{1}{n} \sum_{i=1}^{n} (1 - Y_i \cdot \beta^T X_i)_+ + \lambda \|\beta\|_2^2 \right\}$$

- Kernel ridge regression : K kernel and its parameter

$$\hat{\alpha}_\lambda = \min_{\alpha \in \mathbb{R}^n} \{\alpha^T \mathsf{K} \alpha - 2\alpha^T \mathsf{Y} + \lambda \alpha^T \alpha\}$$

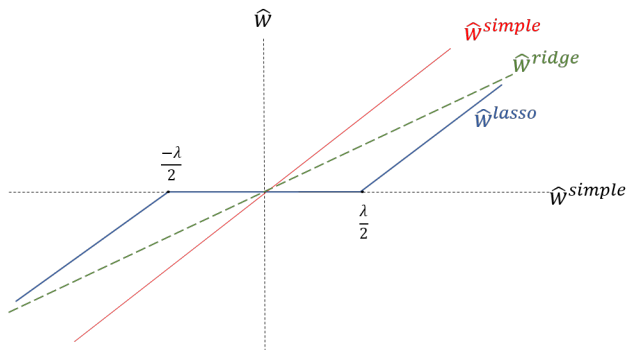
where $\mathsf{K} = (K(X_i, X_j))_{i,j}$

Example of kernel K with one parameter μ :

$$K(x, x') = \exp\left(-\frac{\|x - x'\|_2^2}{\mu}\right) \quad , \quad \mu > 0$$

The four faces of regularization

- Penalized optimization : Tikhonov regularization
- Bayesian priors :
  - LS with gaussian prior <–> Ridge regression
  - LS with Laplace prior <–> LASSO
- Soft order constraint : minimize least squares subject to $\|\beta\|^2 \leq C$ (budget C is the twin sister of smoothing parameter λ)
- Weight decay (also known as 'shrinkage' in mathematical statistics)



Stability of ML algorithms

The principle of stability

- Builds on sensitivity analysis approach applied to machine learning algorithms with respect to changes in the training set
- Stability is a property of the algorithm (e.g. ERM, KRR...) and depends on the loss function
- It builds upon the good old concept of robustness in statistics, revisited with modern tools from probability (concentration inequalities) and applied to the analysis of learning algorithms
- Key references : Bousquet and Elisseeff (2002) and Mukherjee, Niyogi, Poggio, and Rifkin (2006)

Definition of (uniform) stability

- Consider an algorithm which provides an estimator $\tilde{h}_n$ on a sample of size n and we denote $\tilde{h}'_n$ the estimator resulting from the same sample where one observation was changed.
- We say that the algorithm is stable if there exists a constant γ for which we have : for any training sample, and for any pair (x, y),

$$|\ell(y, \tilde{h}_n(x)) - \ell(y, \tilde{h}'_n(x))| \leq \gamma$$

Error bound based on stability

- Consider a cost function L which is uniformly bounded by $M > 0$ and $\tilde{h}_n$ is the output of a γ-uniformly stable learning algorithm.
- We have, with probability at least $1 - \delta$:

$$L(\tilde{h}_n) \leq \hat{L}_n(\tilde{h}_n) + \gamma + (2n\gamma + M)\sqrt{\frac{\log(1/\delta)}{2n}}$$

- Proof based on McDiarmid's concentration inequality

  Hint : under γ-uniform stability assumption, the function $L(\tilde{h}_n) - \hat{L}_n(\tilde{h}_n)$ satisfies the bounded difference assumption with $c/n = 2\gamma + M/n$

  Consequence : the upper bound converges to zero when $\gamma = \gamma_n \to 0$ and $\gamma_n\sqrt{n} \to 0$

Classification case

- Consider a soft classification algorithm (outputs real-valued functions) which is γ-uniformly stable, and a margin loss function ℓ such that : for any y, z

$$\ell_\mu(y,z) = \begin{cases} 1 & \text{if } yz \leq 0 \\ 1 - yz/\mu & \text{if } 0 < yz \leq \mu \\ 0 & \text{if } yz \geq \mu \end{cases}$$

  and the loss is given by $L(h) = \mathbb{E}(\ell_\mu(Y, h(X)))$ for classification data (labels $Y \in \{0,1\}$)

- It can be shown that the previous bound holds with $M = 1$, *and* γ/μ instead of γ

Stability of soft margin SVM

- Assume now that classification data are in $\{-1, +1\}$ and the loss function is :

$$\ell(y, z) = \begin{cases} 1 - yz & \text{if } 1 - yz \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

- We consider the hypothesis space $\mathcal{H}_K$ which is a reproducing kernel Hilbert space with kernel K such that, for any x, $K(x, x) \leq M^2$ for some $M > 0$, with norm denoted by $\|h\|_K$, and the soft margin SVM algorithm which provides the following output : for any $\lambda > 0$

$$\widehat{h}_n^K(\lambda) = \arg\min_{\mathcal{H}_K} \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell(Y_i, h(X_i)) + \lambda \|h\|_K^2 \right\}$$

- It can be shown that this algorithm is stable with parameter γ such that :

$$\gamma \leq \frac{M^2}{2n\lambda}$$

End of Chapter 2

Coming next : analysis of mainstream ML algorithms