Introduction to Statistical Learning

Nicolas Vayatis

Lecture # 5 - Statistical analysis of mainstream ML algorithms

Part I - Margin bounds and application to SVM

Machine Learning Methods
Optimization is central

Some popular examples :

- Sparse linear models $\longrightarrow$ convex optimization (gradient methods)
- Kernel ridge regression $\longrightarrow$ convex optimization (quadratic optimization)
- Deep learning $\longrightarrow$ nonconvex optimization (stochastic gradient descent) + implicit regularization (tricks)

At the end of the day :

loss+training data+functional class+optimization$\rightarrow$random rule $\widehat{f}_n$

Main theoretical objectives of the course

- Take a well-known ML algorithm which operates in $\mathcal{F}$: it produces a (random) sequence of decision rules $(\widehat{f}_n)_{n\geq 1}$ in $\mathcal{F}$. Then show :

- **Convergence of estimation error** :

$$L(\widehat{f}_n) \to \inf_{\mathcal{F}} L \text{ almost surely as } n \to \infty \ ,$$

- **Upper bounds :** with probability at least $1-\delta$, there exists some constant c such that :

$$L(\widehat{f}_n) - \inf_{\mathcal{F}} L \leq C(\mathcal{F}, n) + c\sqrt{\frac{\log(1/\delta)}{n}} \ ,$$

where $C(\mathcal{F}, n)) = O(1/\sqrt{n})$ after processing some complexity/stability measure

Key principle to lower bias : Regularized optimization

- Objective : aim at consistency $L(\widehat{f_n}) \to L^*$ almost surely as $n \to \infty$.
- Take $\mathcal{F}$ a *very large* space and define a proper penalty term :

$$C_n(f) = \underbrace{\hat{L}_n(f)}_{\text{Training error}} + \lambda \underbrace{\mathrm{pen}(f, n)}_{\text{Regularization}}$$

- Example : ridge regression where $f(x) = \theta^T x$: $\hat{L}_n(f) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \theta^T X_i)^2$ and $\mathrm{pen}(f, n) = \frac{1}{n}\|\theta\|_2^2$
- The penalty grows with the complexity of f and vanishes when $n \to \infty$

Overview of Chapter 3

1. Consistency of local methods :
   a. k-Nearest Neighbors
   b. (decision trees)
   c. (local averaging)
2. Consistency of global methods
   a. Support Vector Machines
   b. Boosting
   c. Neural networks
3. Consistency of ensemble methods
   + Bagging, Random Forests

1. Local methods

The example of k-Nearest neighbors (k-NN)

Problem considered
(Multiclass) Classification

- Given :
  - Consider a sample of classification data

$$(X_1, Y_1)...(X_n, Y_n)$$

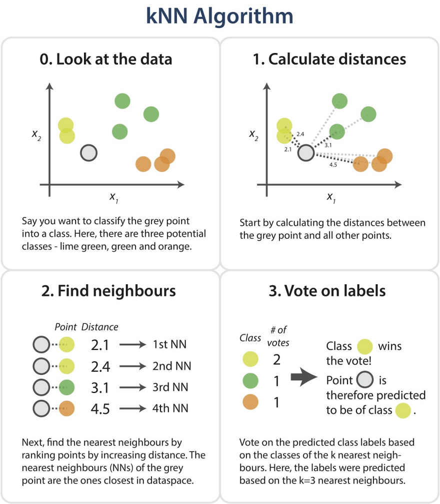
  where $X_i \in \mathbb{R}^d$ vector of independent variables,
  $Y_i \in \{1, \dots, C\}$ the label

- Want :
  - to predict the label y at any position x

k-Nearest Neighbor (1/4)
Principle of the k-NN algorithm

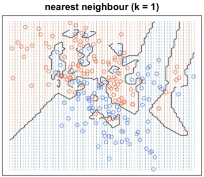
1. Compute distances
   - Compute pairwise distances $d(x, X_i)$ for all $i = 1, \ldots, n$
2. Sort training data
   - Sort the data points from the closest $X_{(1)}$ to the farthest $X_{(n)}$ (i.e. $d(x, X_{(1)}) \leq \ldots \leq d(x, X_{(n)})$)
3. Prediction $\hat{h}(x, k)$ = Majority vote of the k-NN
   - Consider the labels $Y_{(1)}, \ldots, Y_{(k)}$ of the k closest points to x and take the majority vote $\hat{h}(x, k) = \arg\max_c \{\sum_{l=1}^{k} \mathbb{I}\{Y_{(l)} = c\}\}$

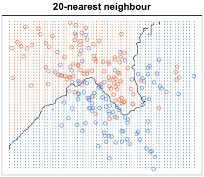
k-Nearest Neighbor (2/4)
Principle of the *k*-NN algorithm

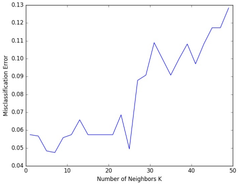


Nearest Neighbors (3/4)
Hyperparameters

- Choice of a distance d between points of $\mathbb{R}^d$
- Number k of Nearest Neighbors, estimated by cross-validation :







k-Nearest Neighbor (4/4) Theory

- Recall : classification error $L(h) = \mathbb{P}(Y \neq h(X))$ and $L^* = \inf L$
- Consistency result :

$$\mathbb{E}L(\hat{h}(\cdot, k_n)) \to L^*$$

under the condition : $k_n \to \infty$ and $k_n/n \to 0$ when $n \to \infty$

- No closed-form solution for optimal k_n (in practice, we use cross-validation)
- No theoretical clue on the choice of the distance (related to data representation and the physics of the problem)

Interlude - Some tools

Definition of Margin Loss, Contraction Principle, Concentration Inequality

Margin loss

- Fix $\rho > 0$
- The *margin loss* is defined, for any $u, v \in \mathbb{R}$, as : $\ell(u, v) = m_\rho(uv)$ where

$$m_\rho(t) = \begin{cases} 0 & \text{if } \rho \le t \\ 1 - \dfrac{t}{\rho} & \text{if } 0 \le t \le \rho \\ 1 & \text{if } t \le 0 \end{cases}$$

- Empirical margin error on a sample D_n :

$$\widehat{L}_{n,\rho}(f) = \frac{1}{n} \sum_{i=1}^{n} m_\rho(Y_i f(X_i))$$

Contraction principle

Theorem. (Ledoux, Talagrand (1991))

Consider $\psi : \mathbb{R} \to \mathbb{R}$ *a Lipschitz function with constant* κ

Then, for any class $\mathcal{F}$ *of real-valued functions, we have :*

$$\widehat{R}_n(\psi \circ \mathcal{F}) \leq \kappa \widehat{R}_n(\mathcal{F})$$

Reminder from Chapter 2
Uniform bound with Rademacher average

Proposition.

Consider $\mathcal{F}$ a class of functions from $\mathcal{Z}$ to $[0,1]$

Then, with probability at least $1-\delta$:

$$\sup_{f\in\mathcal{F}}\left(\mathbb{E}(f(Z_1))-\frac{1}{n}\sum_{i=1}^{n}f(Z_i)\right)\leq 2R_n(\mathcal{F})+\sqrt{\frac{\log(1/\delta)}{2n}}$$

and

$$\sup_{f\in\mathcal{F}}\left(\mathbb{E}(f(Z_1))-\frac{1}{n}\sum_{i=1}^{n}f(Z_i)\right)\leq 2\widehat{R}_n(\mathcal{F})+3\sqrt{\frac{\log(2/\delta)}{2n}}$$

2. Consistency of global methods

a. Support Vector Machines

Principle of Support Vector Machines

- Kernel $k \ : \ \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ symmetric and positive
- Reproducing Kernel Hilbert Space $(\mathcal{H}_k, \langle \cdot, \cdot \rangle)$ corresponding to kernel k.
- Class of functions/classifiers : $g = \mathrm{sgn}(h)$ where

$$h \in \mathcal{H}(X) \mathrel{\hat{=}} \left\{ h = \sum_{i=1}^{n} \alpha_i k(X_i, \cdot) \ : \ \alpha_1, \ldots, \alpha_n \in \mathbb{R} \right\} \subset \mathcal{H}_k$$

- Optimization problem : set $\lambda > 0$

$$\hat{h}_\lambda = \arg\min_{\mathcal{H}_k} \left\{ \sum_{i=1}^{n} (1 - Y_i h(X_i))_+ + \lambda \|h\|_k \right\}$$

RKHS theory in a nutshell

Theorem.
Let $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ *a kernel that is symmetric and positive.*

Then, there exists :

- *a Hilbert space* $(\mathcal{H}_k, \langle \cdot, \cdot \rangle)$*, called the Reproducing Kernel Hilbert Space*
- *a mapping* $\Phi : \mathbb{R}^d \to \mathcal{H}_k$ *such that :*

$$\forall u, v \in \mathbb{R}^d \ , \quad k(u, v) = \langle \Phi(u), \Phi(v) \rangle$$

Plus, we have the reproducing property :

$$\forall h \in \mathcal{H}_k \ , \ \ \forall u \in \mathbb{R}^d \ , \quad h(u) = \langle h, k(u, \cdot) \rangle$$

and $\|h\|_k = \sqrt{\langle h, h \rangle}$

Key property of SVM

- By the representer's theorem (admitted), it suffices to minimize over $\mathcal{H}(X)$ instead of $\mathcal{H}_k$
- Note that, if $h \in \mathcal{H}(X)$:

$$\|h\|_k^2 = \sum_{i,j} \alpha_i \alpha_j k(X_i, X_j)$$

Global methods (e.g. CRM)

- Based on empirical minimization of error functionals
- Example in the case of *soft* classifiers $h \; : \; \mathbb{R}^d \to \mathbb{R}$
- Convex risk minimization, with φ positive convex cost function :

$$\widehat{A}(h) = \frac{1}{n} \sum_{i=1}^{n} \varphi(-Y_i h(X_i))$$

- Note that if $h \in \mathrm{span}(\mathcal{H})$ with $\mathcal{H}$ some class of classifiers, then the minimization problem is convex.
- Main issue : complexity of the class $\mathcal{H}$ of candidate decision rules

Rademacher complexity of SVM

Proposition.

Let $X_1, \ldots, X_n$ be an n-sample in $\mathbb{R}^d$, and denote by K the Gram matrix with coefficients $k(X_i, X_j)$, $1 \leq i, j \leq n$.

Introduce the subspace of functions with bounded RKHS norm :

$$\mathcal{F}_M = \{h \in \mathcal{H}_k \ : \ \|h\|_k \leq M\}$$

We then have :

$$\widehat{R}_n(\mathcal{F}_M) \leq \frac{M\sqrt{\mathrm{trace}\,(K)}}{n}$$

In addition, if we have : $k(X_i, X_i) \leq R^2$ for $1 \leq i \leq n$, then

$$\widehat{R}_n(\mathcal{F}_M) \leq \frac{MR}{\sqrt{n}}$$

Margin bounds for SVM classification

Theorem. (Fixed margin)

Let $\mathcal{H}_k$ the RKHS with kernel k.

Fix $\rho \in (0,1)$, and $\delta > 0$. Then with probability at least $1-\delta$, we have, for any SVM classifier g :

$$L(g) \leq \widehat{L}_{n,\rho}(g) + 2\left(\frac{MR}{\rho\sqrt{n}}\right) + \sqrt{\frac{\log(1/\delta)}{2n}}$$

and

$$L(g) \leq \widehat{L}_{n,\rho}(g) + 2\left(\frac{M\sqrt{\mathrm{trace}\,(K)}}{\rho n}\right) + 3\sqrt{\frac{\log(2/\delta)}{2n}}$$