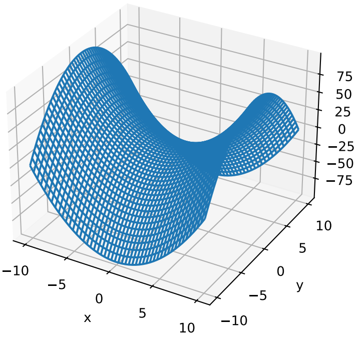
Fondamentaux théoriques du machine learning

Neither positive nor negative Hessian (saddle point)



Overview of lecture 2

Supervised learning

- Excess risk
- Bayes predictor
- Bias-variance decomposition

Supervised learning

- Excess risk
- Bayes predictor
- Bias-variance decomposition

Supervised learning

- The dataset D_n is a collection of n samples $\{(x_i, y_i)\}_{1 \leq i \leq n}$, that are **independent and identically distributed** draws of a joint random variable (X, Y).
- the law of (X, Y) is unknown, we can note it ρ. We assume there exists an unknown function f that relates X and Y (not necessary deterministic).
- we look for an estimator $\tilde{f}_n$ of f. n refers to the fact that we have n samples.

A **learning rule** $\mathcal{A}$ is a application that associates a **prediction function**, or **estimator** $\tilde{f}_n$, to D_n.

$$\mathcal{A} : \begin{cases} \cup_{n \in \mathbb{N}} (\mathcal{X} \times \mathcal{Y})^n \to \mathcal{Y}^{\mathcal{X}} \\ D_n \mapsto \tilde{f}_n \end{cases}$$

Risks

Let l be a loss.
The **risk** (or **statistical risk, generalization error, test error)** of estimator f writes

$$E_{(X,Y)\sim\rho}[l(Y, f(X))]$$

The **empirical risk (ER)** of an estimator f writes

$$R_n(f) = \frac{1}{n}\sum_{i=1}^{n} l(y_i, f(x_i))$$

The risks depend on the loss l.

Excess risk

We define the **target function** f^* by

$$f^* \in \arg\min_{f:X \to Y} R(f)$$

with $f : X \to Y$ set of measurable functions. Notation : $R(f^*) = R^*$.

Definition

Fundamental problem of Supervised Learning
Estimate f^* given only D_n and l.

$\tilde{f}_n$ is the minimizer of the empirical risk.

Definition

Excess risk

The **excess risk** $\mathcal{R}(\tilde{f}_n)$ measures how close $\tilde{f}_n$ is to the best possible f^*, in terms of expected risk (average / expected) error on new examples.

$$\mathcal{R}(\tilde{f}_n) = R(\tilde{f}_n) - R(f^*)$$

Definition

Consistency

The algorithm $\mathcal{A}$ is said to be **consistent** if

$$\lim_{n \to +\infty} E_{D_n} \mathcal{R}(\tilde{f}_n) = 0$$

Bayes predictor

Under some conditions, we can give an explicit formulation of f^*, the best predictor in $\mathcal{Y}^{\mathcal{X}}$, although we can not compute it without the knowledge of the distribution of (X, Y).
In this section we assume we have access to ρ and we approximately ignore measurability issues.
Decision theory : "if we have a perfect knowledge of the underlying probability distribution of the data, what should be done ?"

Bayes predictor

$$f^*(x) = \underset{z \in \mathcal{Y}}{\arg\min}\, E[l(Y, z)|X = x] \tag{1}$$

$E[l(Y, z)|X = x]$ denotes the **conditional expectation** of $l(Y, z)$ given that $X = x$.

$$E[l(Y, z)|X = x] = \int_{y \in \mathbb{R}} l(y, z) p_{Y|X=x}(y) dy \tag{2}$$

Bayes predictor for binary classification

- $\mathcal{Y} = \{0, 1\}$.
- $l(y, z) = 1_{y \neq z}$.

Exercice 1 : **What is the Bayes predictor ?**

Bayes predictor for binary classification

- $\mathcal{Y} = \{0, 1\}$.
- $l(y, z) = 1_{y \neq z}$.
- If $\eta(x) = P(Y = 1 | X = x)$, then

$$R^* = E[min(\eta(x), 1 - \eta(x))] \tag{3}$$

Bayes predictor for binary classification

- $\mathcal{Y} = \{0, 1\}$.
- $l(y, z) = 1_{y \neq z}$.
- If $\eta(x) = P(Y = 1 | X = x)$, then

$$R^* = E[min(\eta(x), 1 - \eta(x))] \tag{4}$$

Exercice 2 : **What is the meaning of having $R^* = 0$ in that context ?**

Bayes predictor for regression, squared loss

- $\mathcal{Y} = \mathbb{R}$, $\mathcal{X} = \mathbb{R}$.
- $l(y, z) = (y - z)^2$

Exercice 3 : **What is the Bayes predictor ?**

Conditional expectation

Definition

Conditional expectation

$$f^*(x) = E[Y|X = x] \tag{5}$$

Risk decomposition

We will introduce the concept of risk decomposition.

- f^* : Bayes predictor
- F : Hypothesis space
- $\tilde{f}_n$: estimated predictor (hence in F).

$$R(\tilde{f}_n) - R^* = \Big(R(\tilde{f}_n) - \inf_{f \in F} R(f)\Big) + \Big(\inf_{f \in F} R(f) - R^*\Big) \qquad (6)$$

Risk decomposition

We will introduce the concept of risk decomposition.

- f^* : Bayes predictor
- F : Hypothesis space
- $\tilde{f}_n$: estimated predictor ($\in F$).

$$R(\tilde{f}_n) - R^* = \left(R(\tilde{f}_n) - \inf_{f \in F} R(f) \right) + \left(\inf_{f \in F} R(f) - R^* \right) \tag{7}$$

However : $\tilde{f}_n$ is a **random variable**, and so is $R(\tilde{f}_n)$. We can also consider the expected value of this quantity.

Risk decomposition

- f^* : Bayes predictor
- F : Hypothesis space
- $\tilde{f}_n$: estimated predictor ($\in F$).

$$E[R(\tilde{f}_n)] - R^* = \Big(E[R(\tilde{f}_n)] - \inf_{f \in F} R(f)\Big) + \Big(\inf_{f \in F} R(f) - R^*\Big) \quad (8)$$

Risk decomposition : bias term

- f^* : Bayes predictor
- F : Hypothesis space
- $\tilde{f}_n$: estimated predictor ($\in F$).

$$E[R(\tilde{f}_n)] - R^* = \left(E[R(\tilde{f}_n)] - \inf_{f \in F} R(f)\right) + \left(\inf_{f \in F} R(f) - R^*\right) \quad (9)$$

Approximation error (bias term) : depends on f^* and F, not on $\tilde{f}_n$, D_n.

$$\inf_{f \in F} R(f) - R^* \geq 0$$

Risk decomposition : bias term

- f^* : Bayes predictor
- F : Hypothesis space
- $\tilde{f}_n$: estimated predictor ($\in F$).

$$E[R(\tilde{f}_n)] - R^* = \Big(E[R(\tilde{f}_n)] - \inf_{f \in F} R(f)\Big) + \Big(\inf_{f \in F} R(f) - R^*\Big) \quad (10)$$

Estimation error (variance term, fluctuation error, stochastic error) : depends on D_n, F, $\tilde{f}_n$.

$$E(R(\tilde{f}_n)) - \inf_{f \in F} R(f) \geq 0$$

Underfitting and overfitting

Approximation error (bias term) : depends on f^* and F, not on $\tilde{f}_n$, D_n.

$$\inf_{f \in F} R(f) - R^* \geq 0$$

Estimation error (variance term, fluctuation error, stochastic error) : depends on D_n, F, $\tilde{f}_n$.

$$E(R(\tilde{f}_n)) - \inf_{f \in F} R(f) \geq 0$$

- too small F : underfitting (large bias, small variance)
- too large F : overfitting (small bias, large variance)

Expected value of empirical risk

If $h \in F$ is fixed (not $\tilde{f}_n$), then $R_n(h)$ is an **unbiased estimator** of the generalization error $R(h)$.

$$E[R_n(h)] = R(h) \tag{11}$$

But

$$E[R_n(\tilde{f}_n)] \neq R(\tilde{f}_n) \tag{12}$$

References I