Machine Learning

1 Introduction

1.1 Binary Classification problem

Find a binary classifier : $\begin{aligned} h : \mathbb{R}^n &\to \{-1, 1\} \\ x &\mapsto h(x) \end{aligned}$

So that $\mathbb{P}_{(x,y)\sim D}[h(x) \neq y]$ is small.
With :

- X : space of input data (image, text, sound, etc.)
- Y : space of label (e.g. $Y = \{-1, 1\}$)
- D : joint probability distribution of $(x, y) \in X \times Y$

$\boxed{\to \textbf{Objectif of ML :} \text{ Risk Minimization for } h \in H}$

Définition - Test error of h

Let $h \in H$, the test error of h is defined as :

$$\begin{aligned} R_D(h) &= \mathbb{E}_{(x,y)\sim D}[\mathbb{1}_{h(x)\neq y}] \\ &= \int_{X\times Y} \mathbb{1}_{h(x)\neq y}\, D(dx, y) \end{aligned}$$

Propriété - Bayes Classifier

The minimal risk is given by the Bayes Classifier :

$$h_{\text{Bayes}} = \underset{y\in\{-1,1\}}{\text{argmax}}\ \mathbb{P}_{(x,y)\sim D}[y|x] \in \{-1, 1\}$$

Example : $\mathbb{P}(y|x)$ with gaussians mixitures

$$\text{Let } \mathbb{P}(y = -1) = \pi_0, \qquad \mathbb{P}(y = 1) = 1 - \pi_0 \qquad \text{with } \pi_0 \in [0, 1]$$
$$\mathbb{P}(x|y = -1) = \mathcal{N}(\mu_1, \Sigma_1), \qquad \mathbb{P}(x|y = 1) = \mathcal{N}(\mu_2, \Sigma_2)$$

$$
\begin{aligned}
\mathbb{P}(y=-1|x) &= \frac{\mathbb{P}(x|y=-1)\mathbb{P}(y=-1)}{\mathbb{P}(x)} \qquad \text{(Bayes' rule)} \\
&= \frac{\mathbb{P}(x|y=-1)\mathbb{P}(y=-1)}{\mathbb{P}(x|y=-1)\mathbb{P}(y=-1)+\mathbb{P}(x|y=1)\mathbb{P}(y=1)}
\end{aligned}
$$

$$
\mathbb{P}(y=1|x) = 1-\mathbb{P}(y=-1|x)
$$

Therefore,

$$
\begin{aligned}
& h_{\text{Bayes}}(x) = \begin{cases} 1 & \text{if } \mathbb{P}(y=1|x) > \mathbb{P}(y=-1|x) \\ -1 & \text{if } \mathbb{P}(y=1|x) < \mathbb{P}(y=-1|x) \\ \pm 1 & \text{if } \mathbb{P}(y=1|x) = \mathbb{P}(y=-1|x) \end{cases} \\
\Leftrightarrow \quad & h_{\text{Bayes}}(x) = \text{sign}(\mathbb{P}(x|y=-1)\pi_0 - \mathbb{P}(x|y=1)(1-\pi_0))
\end{aligned}
$$

Théorème

Let H be all measurable functions from X to $\{-1,1\}$.
Then, $R_D(h) \geq R_D(h_{\text{Bayes}})$ for all $h \in H$.

- Assume $D(dx,y) = \mathbb{P}(x|y)dx \cdot \mathbb{P}(y) = \mathbb{P}(y|x)\mathbb{P}(x)dx$

Then :

$$
\begin{aligned}
R_D(h) &= \mathbb{E}_{(x,y)\sim D}[\mathbb{1}_{h(x)\neq y}] \\
&= \sum_{y\in\{-1,1\}} \int_X \mathbb{1}_{h(x)\neq y} D(dx,y) \\
&= \sum_{y\in\{-1,1\}} \int_X \mathbb{1}_{h(x)\neq y} \mathbb{P}(y|x)\mathbb{P}(x)dx \\
&= \int_X \mathbb{P}(y=1|x)\mathbb{1}_{h(x)\neq 1}\mathbb{P}(x)dx + \int_X \mathbb{P}(y=-1|x)\mathbb{1}_{h(x)\neq -1}\mathbb{P}(x)dx
\end{aligned}
$$

Texte manquant

1.2 Linear Classification problem

In general, D is unknown and $\mathbb{P}(x|y)$ is hard to model, $\mathbb{P}(y)$ prior to choose.
Start from "simple" H : linear classifiers on $x \in \mathbb{R}^n$.

Définition - Linear classifier

A linear classifier is a function $h : \mathbb{R}^n \to \{-1,1\}$ of the form :

$$
\begin{aligned}
h(x) &= \text{sign}(\langle w, x\rangle + b) \\
&= \text{sign}(\sum_{i=1}^{n} w_i x_i + b)
\end{aligned}
$$

with $w \in \mathbb{R}^n$ and $b \in \mathbb{R}$.

Remarque : Labels :
$+1$ if $w^T x + b > 0$
-1 if $w^T x + b < 0$
± 1 if $w^T x + b = 0$

Given a set of training samples iid from D :
$S = \{(x_1, y_1), \dots, (x_m, y_m)\} \in (X \times Y)^m$
Find a classifier $h_S \in H$ such that the generalization error $R_D(h_S)$ is small.

Algorithm 1: Perceptron

Initialize $k = 0$ and $w_0 \in \mathbb{R}^n$
repeat
    for $i = 1, \dots, m$ **do**
        if $sign(w_k^T x_i) = y_i$ **then**
            exit if k big
        else
            if $y_i = 1$ **then**
                $w_{k+1} = w_k + x_i$
            else
                $w_{k+1} = w_k - x_i$
    $k = k + 1$
until;

Remarque : k is the number of iterations or the number of errors made by the algorithm.

Remarque : S can be separated by some $h \in H$.
i.e. $\exists w^* \in \mathbb{R}^n$ so that $||w^*|| = 1$ and $\forall i \in \{1, \dots, m\}, y_i(w^{*T} x_i) > 0$.

Théorème

On linear separable data S and $w_0 = 0$, the Perceptron algorithm generates $(w_k)_{k \geq 0}$ which converges in finite number of error corrections.

- Let $\forall i \leq m, y_i = \text{sign}(\langle w^*, x_i \rangle)$
  Let $R = \max\limits_{i \leq m} ||x_i|| < \infty$ and $M = \min\limits_{i \leq m} y_i \langle w^*, x_i \rangle > 0$

  We have to show that $\langle w_{k+1}, w^* \rangle \geq \langle w_k, w^* \rangle + M$
  Indeed, if $y_i = 1$ and $\text{sign}(\langle w_k, x_i \rangle) = -1$ then :

$$\begin{aligned} & w_{k+1} = w_k + x_i \text{ and } \langle w^*, x_i \rangle \geq M \\ \Rightarrow & \langle w_{k+1}, w^* \rangle = \langle w_k, w^* \rangle + \langle x_i, w^* \rangle \geq \langle w_k, w^* \rangle + M \end{aligned}$$

  Similarly, if $y_i = -1$.

Therefore, $\langle w_k, w^* \rangle \geq kM$ and $||w_k|| \sim \mathcal{O}(\sqrt{k})$.

Then, $\frac{\langle w_k, w^* \rangle}{||w_k||} \geq \frac{kM}{\sqrt{k}R} \geq \frac{M}{R}\sqrt{k}$ and $\langle w_k, w^* \rangle \leq ||w_k|| \cdot ||w^*|| \leq ||w_k||$.

So, $k \leq \left(\frac{R}{M}\right)^2$.

Remarque : M is the margin of the data.
$M = \min\limits_{i \leq m} y_i \langle w^*, x_i \rangle \qquad M \nearrow \Rightarrow k_{\max} \searrow$

Remarque : Unclear if the the Perceptron algorithm finds h_{Bayes} which minimize the test error.
Unclear if S non linear separable ($M \leq 0$).

Extend algo to $H = \{\bar{x} \mapsto \text{sign}(\bar{w}^T \bar{x}) + \bar{b} | \bar{w} \in \mathbb{R}^n, \bar{b} \in \mathbb{R}\}$.
Consider $\bar{x} = (x, 1) \in \mathbb{R}^{n+1}$ and $\bar{w} = (w, b) \in \mathbb{R}^{n+1}$.

2 Support Vector Machine

Find a linear classification which has a maximal margin.
$\Rightarrow$ Smallest test error.

Définition - Margin

Let $w \in \mathbb{R}^n$ and $b \in \mathbb{R}$, the margin of (w, b) is :

$$\phi_h = \min_{i \leq m} \frac{||w^T x_i + b||}{||w||}$$

2.1 Problem formulation

2.1.1 Linearly separable case

$$\max_{w \in \mathbb{R}^n, b \in \mathbb{R}} \phi_h \qquad \text{so that} \qquad \forall i \leq m, y_i(w^T x_i + b) > 0$$

Feasible solution exists : $\exists w \in \mathbb{R}^n, b \in \mathbb{R}$ so that $\forall i \leq m, y_i(w^T x_i + b) > 0$.

Reformulation of SVM :

$$\max_{w \in \mathbb{R}^n, b \in \mathbb{R}} \min_{i \leq m} \frac{y_i(w^T x_i + b)}{||w||}$$

Remarque : Invariance by scaling : $\forall \lambda > 0, (w, b)$ solution $\Rightarrow (\lambda w, \lambda b)$ solution.
$\Rightarrow$ Set $\min\limits_{i \leq m} y_i(w^T x_i + b) = 1$.

Formulation of SVM :

$$(P) \qquad \max_{w \in \mathbb{R}^n, b \in \mathbb{R}} \frac{1}{||w||} \quad \text{so that} \quad \min_{i \leq m} y_i(w^T x_i + b) = 1 \tag{1}$$

$$(P') \qquad \max_{w \in \mathbb{R}^n, b \in \mathbb{R}} \frac{1}{||w||} \quad \text{so that} \quad \forall i \leq m, y_i(w^T x_i + b) \geq 1 \tag{2}$$

Propriété

(P) and (P') are equivalent.

▶
- If $(\bar{w}, \bar{b})$ is a solution of (P'), then it is a feasible solution of (P).
  - If $\min\limits_{i \leq m} y_i(\bar{w}^T x_i + \bar{b}) = 1$, then $(P') \Rightarrow (P)$.
  - If $\min\limits_{i \leq m} y_i(\bar{w}^T x_i + \bar{b}) > 1$:
    Let $\phi_{\bar{w},\bar{b}} = \min\limits_{i \leq m} \frac{y_i(\bar{w}^T x_i + \bar{b})}{||\bar{w}||} > \frac{1}{||\bar{w}||}$.
    Let $\hat{w} = \frac{\bar{w}}{||\bar{w}||} \frac{1}{\phi_{\bar{w},\bar{b}}}$ and $\hat{b} = \frac{\bar{b}}{||\bar{w}||} \frac{1}{\phi_{\bar{w},\bar{b}}}$
    Then, $\min\limits_{i \leq m} y_i(\hat{w}^T x_i + \hat{b}) = 1$ and $\frac{1}{||\hat{w}||} < \frac{1}{||\bar{w}||}$. So $(\bar{w}, \bar{b})$ is not optimal for (P') : absurd.
- $\forall (w, b)$ so that $\min\limits_{i \leq m} y_i(w^T x_i + b) = 1$ (solution of (P)), we have :
  $\frac{1}{||\bar{w}||} \geq \frac{1}{||w||}$ (by optimality of (P')).
  So $(\bar{w}, \bar{b})$ is a solution of (P').

Primal problem :

$$(P'') \qquad \min_{w \in \mathbb{R}^n, b \in \mathbb{R}} \frac{1}{2}||w||^2 \quad \text{so that} \quad \forall i \leq m, y_i(w^T x_i + b) \geq 1 \tag{3}$$

Remarque : $(P'') \Leftrightarrow (P')$
(P'') is a quadratic programming with linear constraints.

Remarque : We can deduce the dual problem of (P'') with KKT.

2.1.2 Extension to non-linearly separable data

$$\min_{w\in\mathbb{R}^n, b\in\mathbb{R}, \xi\in\mathbb{R}^m} \frac{1}{2}||w||^2 + C\sum_{i=1}^{m}\xi_i^p \quad \text{so that} \quad \forall i \leq m, y_i(w^T x_i + b) \geq 1 - \xi_i \tag{4}$$

With :

- $C > 0$: regularization parameter
- $\xi_i \geq 0$: slack variable
- $p \geq 1$: norm of the slack variable

3 Generalisation theory in binary classification

Find $h \in H$ so that $R_D(h) = \mathbb{P}_{(x,y)\sim D}[h(x) \neq y]$ is small.

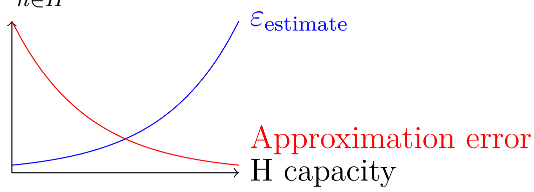
Given a set of training samples iid from D : $S = \{(x_1, y_1), \ldots, (x_m, y_m)\} \in (X \times Y)^m$
$h_s \in H$ is the classifier learned from an algorithm (Perceptron, SVM, etc.)

Propriété

$$\min_{h\in H} \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}_{h(x_i)\neq y_i} \xrightarrow[m\to\infty]{\text{Loi des Grands Nombres}} R_D(h) \tag{5}$$

Classical picture of ML theory :

$R_D(h_S) = \min_{h\in H} R_D(h)$



Overfitting :
$R_D(h_S)$ big but $\frac{1}{m}\sum_{i=1}^{m} \mathbb{1}_{h(x_i)\neq y_i} \approx 0$

Texte manquant
How to analyse $R_D(h_S)$?
S iid from $D \Rightarrow h_S$ is a random variable $\Rightarrow R_D(h_S)$ is a random variable.

1. Control of $\mathbb{E}_{S\sim D^m}[R_D(h_S)]$
2. Confidence interval for $R_D(h_S)$

3.1 Control of $\mathbb{E}_{S\sim D^m}[R_D(h_S)]$

Leave-one-out cross-validation analysis for linear separable data :
$\mathbb{P}_{S\sim D^m}[S \text{ linear separable}] = 1$.

Leave-one-out algo A : $h_S = A(S)$

$\hat{R}_{\text{LOO}}(A) = \frac{1}{m}\sum_{i=1}^m \mathbb{1}_{h_{S\setminus\{x_i\}}(x_i)\neq y_i}$

Propriété

If $m \geq 2$, then $\mathbb{E}_{S\sim D^m}[\hat{R}_{\text{LOO}}(A)] = \mathbb{E}_{S'\sim D^{m-1}}[R_D(h'_S)]$

▶
$$\begin{aligned}
\mathbb{E}_{S\sim D^m}[\hat{R}_{\text{LOO}}(A)] &= \mathbb{E}_{S\sim D^m}\left[\frac{1}{m}\sum_{i=1}^m \mathbb{1}_{h_{S\setminus\{x_i\}}(x_i)\neq y_i}\right] \\
&= \frac{1}{m}\sum_{i=1}^m \mathbb{E}_{S\sim D^m}[\mathbb{1}_{h_{S\setminus\{x_i\}}(x_i)\neq y_i}] \\
&= \mathbb{E}_{S\sim D^m}[\mathbb{1}_{h_{S\setminus\{x_1\}}(x_1)\neq y_1}] \qquad \text{(by independence)} \\
&= \mathbb{E}_{S'\sim D^{m-1}}[\mathbb{E}_{(x,y)\sim D}[\mathbb{1}_{h_{S'}(x)\neq y}]] \qquad (S' = S\setminus\{x_1\}) \\
&= \mathbb{E}_{S'\sim D^{m-1}}[R_D(h_{S'})]
\end{aligned}$$

Théorème

Assume S is linearly separable (almost surely).
Let $N_{\text{sn}}(S) = |\{x_i | y_i(w^T x_i + b) = 1, i \leq m\}|$ (number of support vectors).
Then, $\mathbb{E}_{S\sim D^m}[\hat{R}_{\text{LOO}}(A)] \leq \mathbb{E}_{S\sim D^m}[\frac{N_{\text{sn}}(S)}{m}]$

▶ Let $(x, y) \in S$. If x is not a support vector of h_S : $h_{S\setminus\{x\}} = h_S$.
Therefore, if $h_{S\setminus\{x\}}(x) \neq y$ then x is a support vector of h_S.
So, $\hat{R}_{\text{LOO}}(h_S) \leq \frac{N_{\text{sn}}(S)}{m}$.