Machine Learning

GT Veille OSMP - Scoring et IA pour la lutte contre la fraude

Théo Lopès-Quintas

BPCE Payment Services

24 avril 2024

Introduction

Introduction
Un peu d'histoire

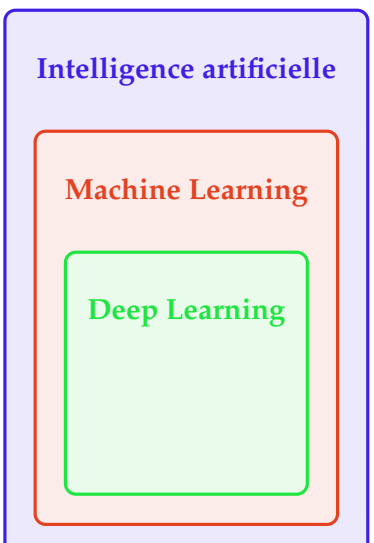
Nous ne pouvons qu'avoir un aperçu du futur, mais cela suffit pour comprendre qu'il y a beaucoup à faire.
— Alan Turing (1950)

- **Conférence de Dartmouth - 1956** : Début des travaux dans l'objectif de créer des machines intelligentes
- **Scikit-Learn - 2007** : Création d'une librairie open-source pour faciliter la modélisation en Machine Learning
- **AlexNet - 2012** : avènement du Deep Learning avec un modèle de classification d'image révolutionnaire dans la compétition ImageNet

Introduction

Qu'est-ce que *l'Intelligence Artificielle* ?

- **Algorithme** : Ensemble hiérarchisé d'opérations logiques à exécuter dans le but de résoudre un problème [a]
- **Intelligence artificielle** : Ensemble d'algorithme résolvant des problèmes sans être explicitement programmé pour le faire
- **Machine Learning** : Sous-ensemble de l'IA où les algorithmes apprennent à partir d'une base de données
- **Deep Learning** : Sous-ensemble du ML où les algorithmes sont des variantes d'un algorithme de ML nommé *réseau de neurones*



a. Aurélie Jean, De l'autre côté de la Machine

Pour chacun, on peut considérer plusieurs approches, entre autres :

- **Supervisé** : on cherche à reproduire une réponse à partir de données
- **Non supervisé** : on ne possède pas de réponse pré-définie, on peut vouloir réduire la dimension, regrouper les observations qui se ressemblent...
- **Par renforcement** : on apprend la meilleure action à réaliser dans un environnement sur lequel on agit, selon une politique fixée

Dans le cadre supervisé, nous avons accès à un dataset $\mathcal{D}$ défini comme :

$$\mathcal{D} = \left\{ (x_i, y_i) \mid \forall i \leqslant n,\ x_i \in \mathbb{R}^{d'}, y_i \in \mathcal{Y} \right\}$$

Nombre d'informations

Nombre d'observations

Avec $\mathcal{Y} \subseteq \mathbb{R}$ pour un problème de régression et $\mathcal{Y} \subset \mathbb{N}$ dans le cadre d'une classification. Dans le cadre non supervisé nous n'avons pas de y.

Les problèmes de Machine Learning supervisé peuvent souvent s'écrire sous la forme d'une optimisation d'une fonction de perte $\mathcal{L} : \mathbb{R}^d \times \mathcal{M}_{n,d'} \times \mathbb{R}^n \to \mathbb{R}_+$ comme :

Vecteur des paramètres optimaux

$$w^* = \arg\min_{w \in \mathbb{R}^d} \mathcal{L}(w, X, y)$$

Dimension du vecteur de paramètres

Dans la suite, pour simplifier les notations, nous omettrons la dépendance de $\mathcal{L}$ en X (matrice des informations) et y (vecteur réponse). Notons qu'en général, nous avons $d \neq d'$ et dans le cas du deep learning, très souvent $d >> d'$.

Principaux algorithmes

La régression logistique suppose un lien *linéaire* entre les features et la côte que l'observation soit de la classe d'intérêt. On modélise cela par la fonction f :

$$f(x) = \frac{1}{1 + e^{-(x_1 w_1 + \ldots + x_d w_d)}} = \frac{1}{1 + e^{-\langle x,\, w \rangle}} \qquad \text{(Régression logistique)}$$

$w = (w_1, \ldots, w_d) \in \mathbb{R}^d$

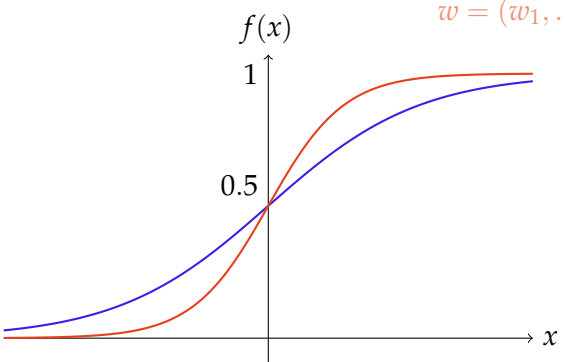


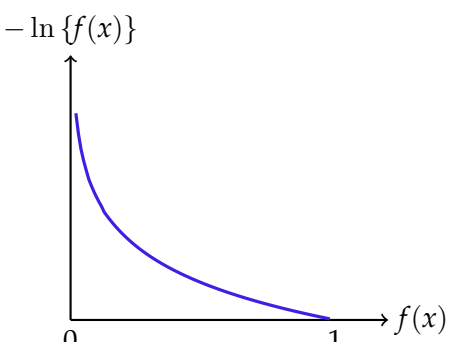
Figure – $f(x) = \frac{1}{1 + e^{-x}}$ et $f(x) = \frac{1}{1 + e^{-2x}}$

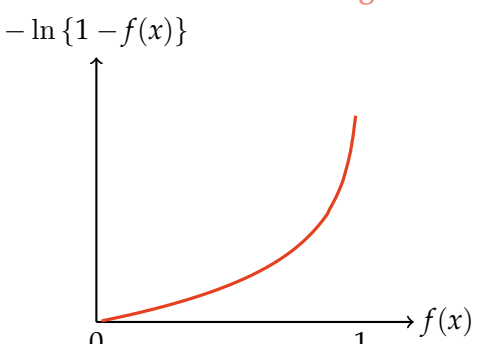
Pour apprendre $w \in \mathbb{R}^d$, on se propose la fonction de perte :

Observation positive

$$\mathcal{L}(w) = -\Big[\, y \ln\{f(x)\} + (1-y)\ln\{1-f(x)\} \,\Big]$$

Observation négative





La méthode la plus utilisée pour résoudre ce genre de problème est la descente de gradient :

$$w_{t+1} = w_t - \eta_t \nabla \mathcal{L}(w_t)$$

η_t : Learning rate

Quand on travaille avec des grands datasets, le coût de calcul/temps est grand si l'on calcule $\nabla \mathcal{L}(w_t)$ pour la totalité de la base. On peut donc considérer d'autres approches :

- **Stochastique (SGD)** : on sélectionne au hasard une observation et on met à jour w
- **Stochastique par batch** : on sélectionne aléatoirement une partie de la base (batch) et on met à jour w à chaque batch

De nombreux travaux portent sur l'accélération de cette méthode et la caractérisation de la vitesse de convergence des différents schémas.

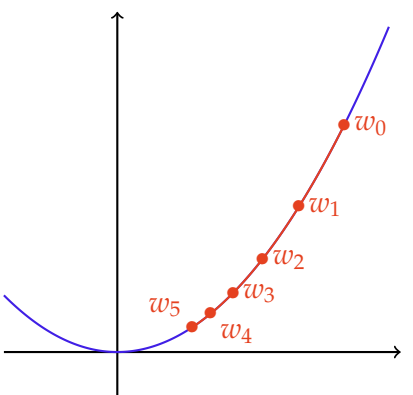


Figure – Exemple d'une descente de gradient pour $f(x) = x^2$

On cherche à présent une fonction sous la forme suivante, où l'on cherche les partitions P de l'espace :

Probabilité de la classe d'intérêt dans la partition P

$$f_\theta(x) = \sum_{P \in \theta} \mu_P \mathbb{1}_{\{x \in P\}}$$

Partition de l'espace

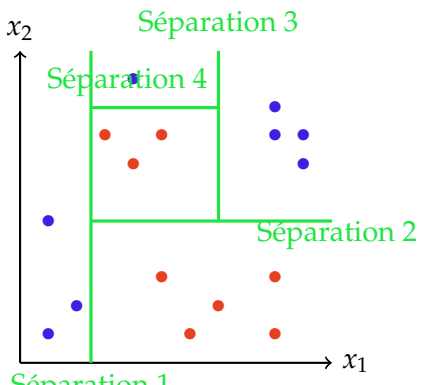


Figure – Exemple de partitionnement de l'espace

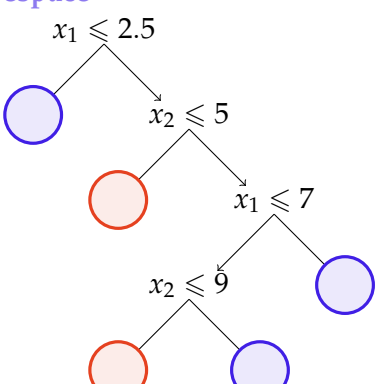
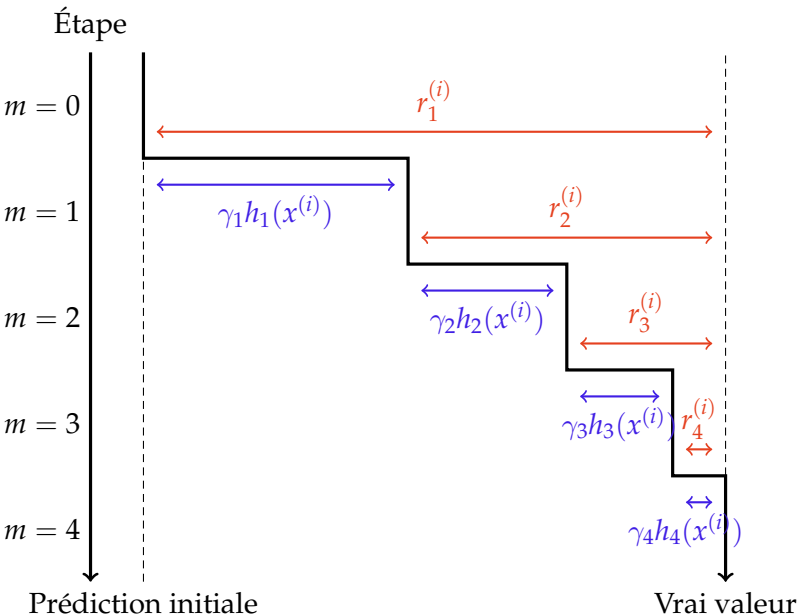


Figure – Arbre de décision associé au partitionnement

Le principe du Boosting est de combiner plusieurs algorithmes les uns après les autres pour que chacun corrige les erreurs du précédent.
Les fonctions h sont ce qu'on cherche à apprendre, et on prend très souvent des arbres.
Les scalaires γ sont appris pendant l'entraînement



Quels challenges dans la lutte contre la fraude ?

Quels challenges dans la lutte contre la fraude ?

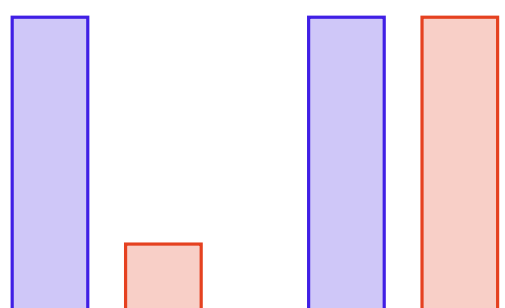
Déséquilibre de classe : Sur-échantillonnage

Une première méthode pour équilibrer le dataset est de dupliquer aléatoirement des observations de la classe minoritaire, jusqu'à ce qu'on atteigne la proportion de la classe minoritaire voulue p_f.

Quels challenges dans la lutte contre la fraude ?

Déséquilibre de classe : Sur-échantillonnage

Une première méthode pour équilibrer le dataset est de dupliquer aléatoirement des observations de la classe minoritaire, jusqu'à ce qu'on atteigne la proportion de la classe minoritaire voulue p_f.



(a) Situation initiale **(b)** Situation finale

Figure – Illustration du sur-échantillonnage

Une première méthode pour équilibrer le dataset est de dupliquer aléatoirement des observations de la classe minoritaire, jusqu'à ce qu'on atteigne la proportion de la classe minoritaire voulue p_f.

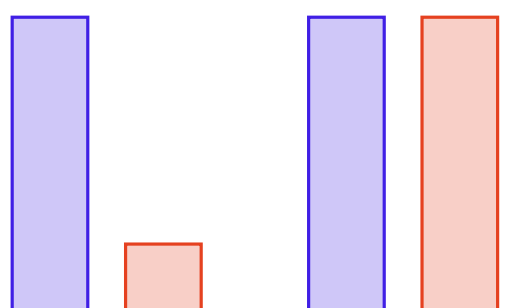
Proportion de la classe minoritaire initiale

Taille initiale du dataset

$$n_f = n_i \frac{1 - p_i}{1 - p_f}$$

Taille finale du dataset

(Over sampling)



(a) Situation initiale **(b)** Situation finale

Figure – Illustration du sur-échantillonnage

Quels challenges dans la lutte contre la fraude ?

Déséquilibre de classe : Sur-échantillonnage

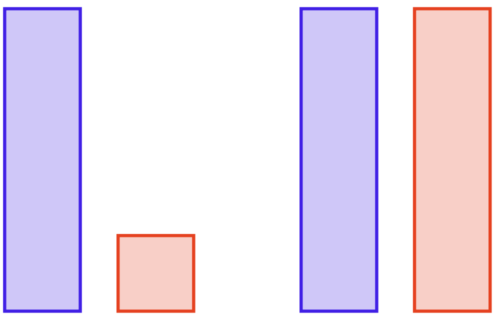
Une première méthode pour équilibrer le dataset est de dupliquer aléatoirement des observations de la classe minoritaire, jusqu'à ce qu'on atteigne la proportion de la classe minoritaire voulue p_f.

Proportion de la classe minoritaire initiale

Taille initiale du dataset

$$n_f = n_i \frac{1 - p_i}{1 - p_f} \qquad \text{(Over sampling)}$$

Taille finale du dataset



(a) Situation initiale **(b)** Situation finale

Figure – Illustration du sur-échantillonnage

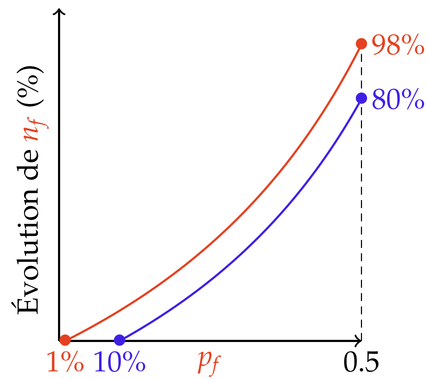


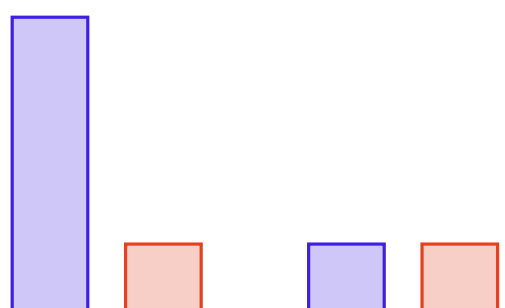
Figure – Évolution de la taille du dataset final en fonction de la proportion finale de la classe minoritaire

Quels challenges dans la lutte contre la fraude ?

Déséquilibre de classe : Sous-échantillonnage

Une deuxième méthode pour équilibrer le dataset est de supprimer aléatoirement des observations de la classe majoritaire, jusqu'à ce qu'on atteigne la proportion de la classe minoritaire voulue p_f.

Une deuxième méthode pour équilibrer le dataset est de supprimer aléatoirement des observations de la classe majoritaire, jusqu'à ce qu'on atteigne la proportion de la classe minoritaire voulue p_f.



(a) Situation initiale **(b)** Situation finale

Figure – Illustration du sous-échantillonnage

Une deuxième méthode pour équilibrer le dataset est de supprimer aléatoirement des observations de la classe majoritaire, jusqu'à ce qu'on atteigne la proportion de la classe minoritaire voulue p_f.

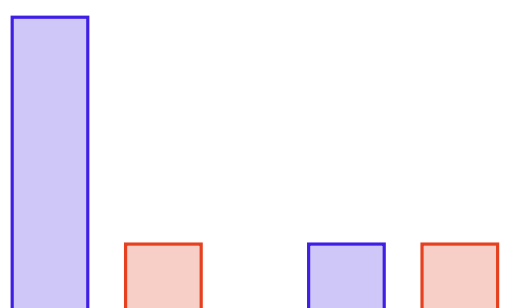
Proportion de la classe minoritaire initiale

Taille initiale du dataset

$$n_f = n_i \frac{p_i}{p_f}$$

(Under sampling)

Taille finale du dataset



(a) Situation initiale **(b)** Situation finale

Figure – Illustration du sous-échantillonnage

Quels challenges dans la lutte contre la fraude ?

Déséquilibre de classe : Sous-échantillonnage

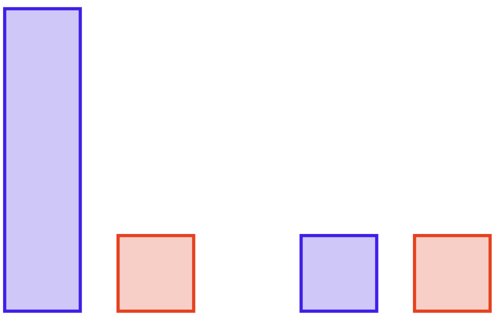
Une deuxième méthode pour équilibrer le dataset est de supprimer aléatoirement des observations de la classe majoritaire, jusqu'à ce qu'on atteigne la proportion de la classe minoritaire voulue p_f.

Proportion de la classe minoritaire initiale

Taille initiale du dataset

$$n_f = n_i \frac{p_i}{p_f} \qquad \text{(Under sampling)}$$

Taille finale du dataset



(a) Situation initiale **(b)** Situation finale

Figure – Illustration du sous-échantillonnage

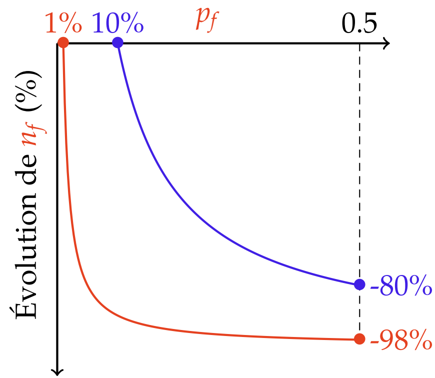


Figure – Évolution de la taille du dataset final en fonction de la proportion finale de la classe minoritaire

Quels challenges dans la lutte contre la fraude ?

Drifts : changements de distributions

Le monde des paiements change, et les fraudeurs aussi. On peut noter deux manière principales dont les mouvements se font :

- **Covariate shift** : La distribution des features change
- **Concept drift** : La relation entre la cible et les features change

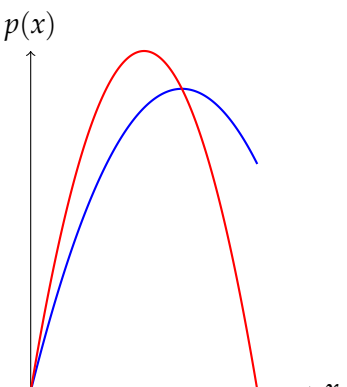


Figure – Covariate shift : changement de distribution d'une feature

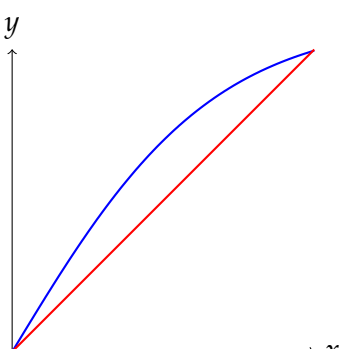


Figure – Concept drift : changement de relation entre x et y

Quels challenges dans la lutte contre la fraude ?

En résumé

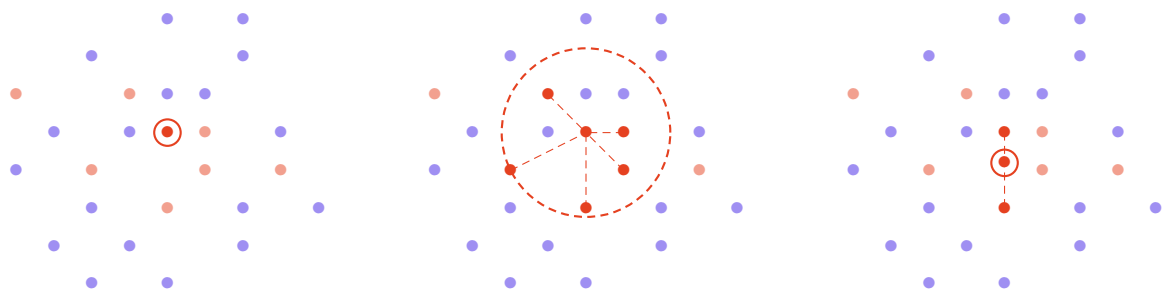
- **L'hyper déséquilibre** de classe induit des difficultés d'entraînement et nécessite des réponses mesurées
- **Les drifts** inhérent à l'activité de production accentuent le premier point et nécessite une création et un suivi minutieux des algorithmes
- **Le fléau de la dimension** peut survenir plus tôt que prévu, une réponse est de mettre l'accent sur le travail métier pour créer de la valeur

Annexe : SMOTE

Annexe : SMOTE

Fonctionnement

Les deux précédentes approches dupliquent ou suppriment des observations du dataset. On peut explorer la possibilité de *créer* des observations synthétiques : c'est l'objet de SMOTE.



(a) Sélection aléatoire d'une observation de la classe minoritaire

(b) Identification des k=5 voisins les plus proches

(c) Sélection aléatoire d'un voisin et création de l'observation

Figure – Fonctionnement de SMOTE

Annexe : SMOTE

Observations synthétiques ambiguës

Soit $\mathcal{D}$ un dataset comme défini dans l'introduction dont on reprend les notations, et $\mathcal{D}^-$ le dataset contenant uniquement les observations de la classe majoritaire. On note $n^- = \#\mathcal{D}^-$.
Soit $\widetilde{x}$ un point synthétique généré par l'algorithme SMOTE. On dit que $\widetilde{x}$ est un **point ambigu** si $\min_{x \in \mathcal{D}^-} \|\widetilde{x} - x\| \leqslant \delta$ pour $\delta \geqslant 0$ un paramètre fixé.
On peut montrer, sous l'hypothèse que $\mathcal{D} \sim \mathcal{N}(0_d, I_d)$, que :

$$\mathbb{P}\left(\min_{x \in \mathcal{D}^-} \|\widetilde{x} - x\| \leqslant \delta\right) = 1 - \left(1 - \int_0^{\frac{\delta^2}{2}} \frac{t^{\frac{d}{2}-1} e^{-\frac{t}{2}}}{2^{\frac{d}{2}} \Gamma\left(\frac{d}{2}\right)} \mathrm{d}t\right)^{n^-}$$

Ainsi, plus n^- est grand, plus la probabilité de créer des observations synthétiques ambiguës est importante. Il faut également noter que quand d tend vers l'infini, la probabilité tend vers 0 : mais ce régime n'est pas souhaitable avec le fléau de la dimension.

Annexe : SMOTE

Observations synthétiques ambiguës - observation sur un dataset réel

Nous souhaitons appliquer cela à un dataset de 20 millions de lignes ayant 1% de déséquilibre.

Déséquilibre final (%)	Observation synthétique ambiguë (%)	Taille du dataset final (en millions)
1	0	20.000
2	39.29	20.204
5	62.86	20.842
10	70.72	22.000
20	74.65	24.750
30	75.96	28.285
40	76.61	33.000
50	77.01	39.600

Table – Évolution de la proportion d'observation synthétique ambiguë et de la taille du dataset final en fonction de la proportion de déséquilibre souhaité

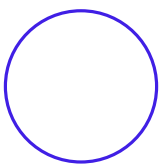
Nous avons fixé la valeur de δ comme le quantile à 0.5% de la loi du χ^2 induite par le dataset.

Annexe : Fléau de la dimension

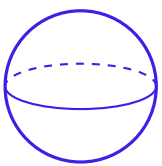
Annexe : Fléau de la dimension

Volume d'une hypersphère

(a) En dimension 1, volume = 2

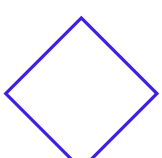


(b) En dimension 2, volume = π

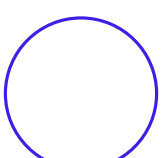


(c) En dimension 3, volume = $\frac{4}{3}\pi$

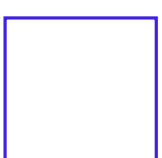
Figure – Représentation et volume d'une hypersphère de rayon 1 dans 3 espaces de dimensions différentes



(a) Avec la norme 1



(b) Avec la norme 2



(c) Avec la norme infinie

Figure – Représentation d'une hypersphère de rayon 1 en dimension 2 pour 3 normes différentes

Annexe : Fléau de la dimension

Volume d'une hypersphère

On appelle *boule* ou hypersphère l'objet défini par :

$$B_n^p(R) = \{u \in \mathbb{R}^n, \|u\|_p^p \leqslant R^p\}$$

Et son volume par :

$$V_n^p(R) = \int_{B_n^p(R)} \bigotimes_{i=1}^{n} \mathrm{d}x_i$$

Proposition 1 (Volume d'une hypersphere)

Avec les notations précédentes, on a :

$$\begin{aligned} \forall R > 0, \forall n \geqslant 2, \forall p \geqslant 1, \quad V_n^p(R) &= \frac{\left(2R\Gamma\left(\frac{1}{p}+1\right)\right)^n}{\Gamma\left(\frac{n}{p}+1\right)} \\ &\sim \sqrt{\frac{p}{2\pi n}}\left[2R\Gamma\left(\frac{1}{p}+1\right)\left(\frac{pe}{n}\right)^{\frac{1}{p}}\right]^n \end{aligned}$$

Avec la fonction Γ *définie comme :*

$$\Gamma(x) = \int_0^{+\infty} e^{-t} t^{x-1} \mathrm{d}t$$

Annexe : Fléau de la dimension

Concentration dans une hypersphère

On rappelle que :

$$\forall R > 0, \forall n \geqslant 2, \forall p \geqslant 1, \quad V_n^p(R) = \frac{\left(2R\Gamma\left(\frac{1}{p}+1\right)\right)^n}{\Gamma\left(\frac{n}{p}+1\right)}$$

Exercice 1 (Concentration dans l'hypersphère)

Soit $\varepsilon > 0$. On considère une hypersphère de rayon R. Montrer que :

$$\frac{V_n^p(R-\varepsilon)}{V_n^p(R)} = \left(1 - \frac{\varepsilon}{R}\right)^n$$