Machine Learning pour la lutte contre la fraude

Voir grand et tout petit

Théo Lopès-Quintas

BPCE Payment Services,
Université Paris Dauphine

9 octobre 2023

Contexte

Formalisation du problème Machine Learning

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

$$\mathcal{D} = \left\{ (x_i, y_i) \mid \forall i \leqslant \underbrace{n}_{\text{Nombre d'observations}},\ x_i \in \mathbb{R}^{\overbrace{d'}^{\text{Nombre d'informations}}}, y_i \in \mathcal{Y} \right\} \tag{1}$$

Avec $\mathcal{Y} \subseteq \mathbb{R}$ pour un problème de régression et $\mathcal{Y} \subset \mathbb{N}$ dans le cadre d'une classification. Les problèmes de Machine Learning supervisés 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 :

$$\overbrace{\theta^*}^{\text{Vecteur des paramètres optimaux}} = \arg\min_{\theta \in \mathbb{R}^{\underbrace{d}_{\text{Dimension du vecteur de paramètres}}}} \mathcal{L}(\theta; X, y)$$

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

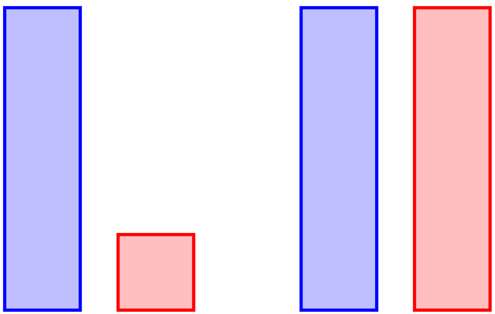
Hyper déséquilibre de classe

Sur-échantillonnage aléatoire

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.

Hyper déséquilibre de classe
Sur-échantillonnage aléatoire

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

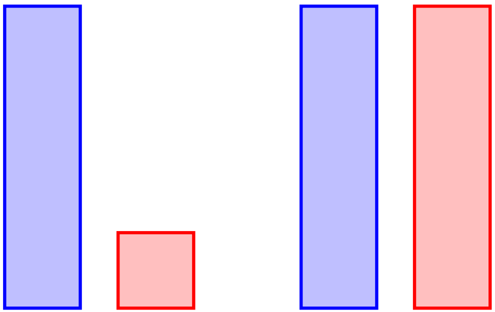
Figure – Illustration du sur-échantillonnage

Hyper déséquilibre de classe

Sur-échantillonnage aléatoire

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.

$$\underbrace{n_f}_{\text{Taille finale du dataset}} = \overbrace{n_i}^{\text{Taille initiale du dataset}} \frac{1 - \overbrace{p_i}^{\text{Proportion de la classe minoritaire initiale}}}{1 - p_f} \qquad \text{(Over sampling)}$$



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

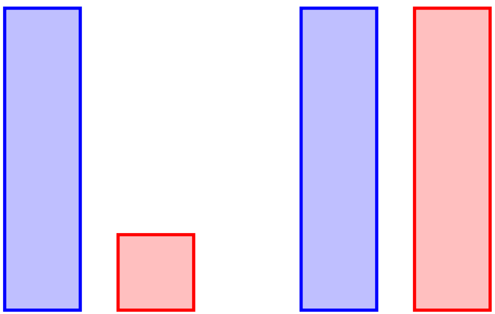
Figure – Illustration du sur-échantillonnage

Hyper déséquilibre de classe

Sur-échantillonnage aléatoire

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.

$$\underbrace{n_f}_{\text{Taille finale du dataset}} = \overbrace{n_i}^{\text{Taille initiale du dataset}} \frac{1 - \overbrace{p_i}^{\text{Proportion de la classe minoritaire initiale}}}{1 - p_f} \qquad \text{(Over sampling)}$$



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

Figure – Illustration du sur-échantillonnage

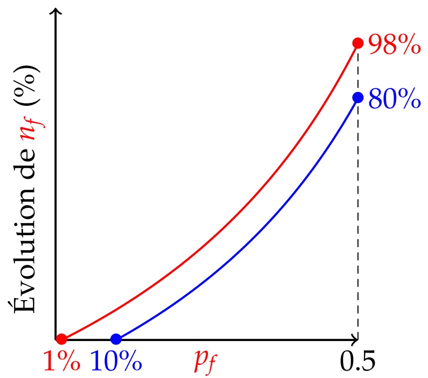


Figure – Évolution de la taille du dataset final en fonction de la proportion finale de la classe minoritaire

Hyper déséquilibre de classe

Sur-échantillonnage aléatoire

Proportion de la classe minoritaire initiale

Taille initiale du dataset

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

Taille finale du dataset

Proportion de la classe minoritaire finale

Exercice 1 (Paramétrer un arbre)

Nous avons accès à un dataset d'un million de lignes avec une proportion initiale de la classe minoritaire de 1%. Après avoir équilibré le dataset avec du sur-échantillonnage, vous utilisez un arbre ou une méthode à base d'arbres.

1. *Combien y avait-il d'observations de la classe minoritaire au début ? A la fin ?*
2. *En moyenne, combien de fois une même observation de la classe minoritaire est présente dans le dataset final ?*
3. *Comment ajuster en conséquence les paramètres d'un arbre ?*

Hyper déséquilibre de classe

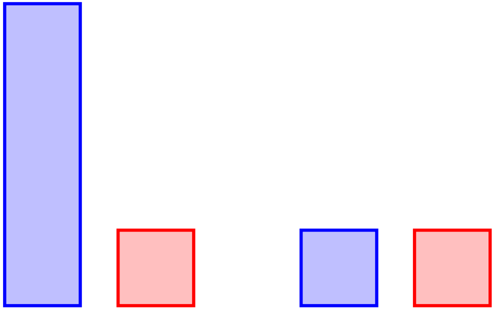
Sous-échantillonnage aléatoire

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.

Hyper déséquilibre de classe

Sous-échantillonnage aléatoire

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

Hyper déséquilibre de classe

Sous-échantillonnage aléatoire

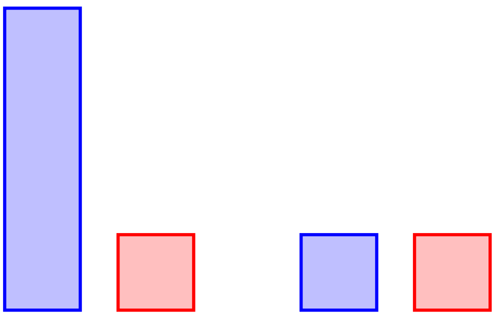
Une deuxième méthode pour équilibrer le dataset est de supprimer aléatoirement des observations de la classe majoritaire, jusqu'à ce qu'on atteigne la proportion de la classe minoritaire voulue p_f.

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

Hyper déséquilibre de classe

Sous-échantillonnage aléatoire

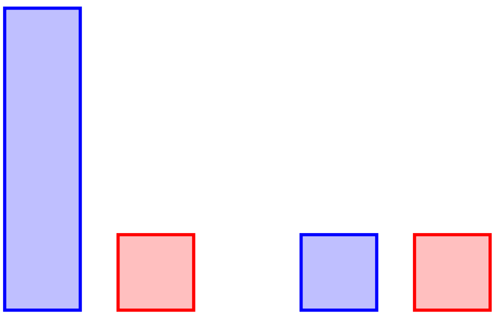
Une deuxième méthode pour équilibrer le dataset est de supprimer aléatoirement des observations de la classe majoritaire, jusqu'à ce qu'on atteigne la proportion de la classe minoritaire voulue p_f.

Proportion de la classe minoritaire initiale

Taille initiale du dataset

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

Taille finale du dataset



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

Figure – Illustration du sous-échantillonnage

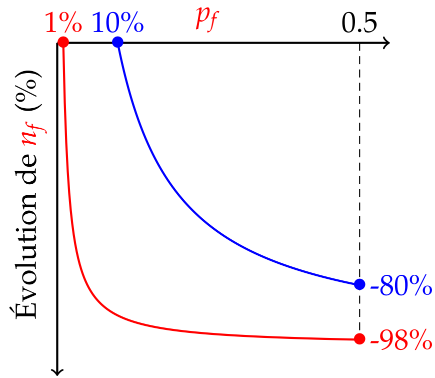


Figure – Évolution de la taille du dataset final en fonction de la proportion finale de la classe minoritaire

HYPER DÉSÉQUILIBRE DE CLASSE

SOUS-ÉCHANTILLONNAGE ALÉATOIRE

$$\underbrace{n_f}_{\text{Taille finale du dataset}} = \overbrace{n_i}^{\text{Taille initiale du dataset}} \frac{\overbrace{p_i}^{\text{Proportion de la classe minoritaire initiale}}}{\underbrace{p_f}_{\text{Proportion de la classe minoritaire finale}}} \qquad \text{(Under sampling)}$$

Exercice 2 (Conserver la représentativité)

Nous avons accès à un dataset d'un million de lignes avec une proportion initiale de la classe minoritaire de 1%. Après avoir équilibré le dataset avec du sous-échantillonnage, vous explorez les données.

1. *Combien y avait-il d'observations de la classe majoritaire au début ? A la fin ?*
2. *Quelle est la probabilité qu'une observation de la classe majoritaire soit présente dans le dataset final ?*
3. *Supposons qu'un groupe d'observations (par exemple les transactions par téléphone ou mail) représente 5% des observations de la classe majoritaire. Quelle est la probabilité qu'il ne soit plus présent dans le dataset final ?*
4. *Quel problème cela induit-il sur le modèle à construire en production ? Comment s'en prémunir ?*

Hyper déséquilibre de classe

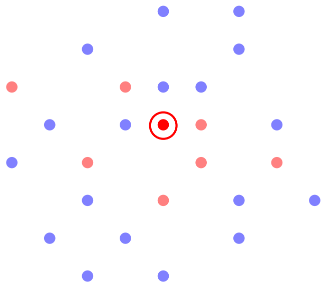
SMOTE : une approche prometteuse

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.

Hyper déséquilibre de classe

SMOTE : une approche prometteuse

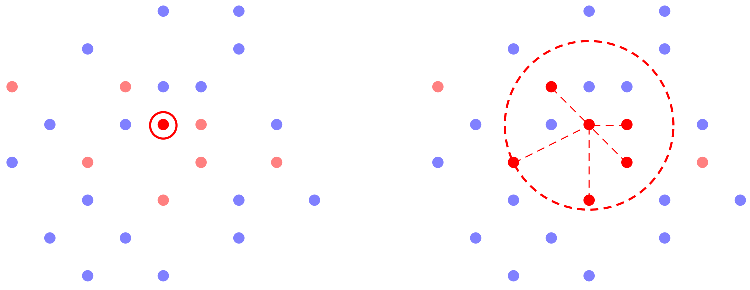
Les deux précédentes approches dupliquent ou suppriment des observations du dataset. On peut explorer la possibilité de *créer* des observations synthétiques : c'est l'objet de SMOTE.



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

Figure – Fonctionnement de SMOTE

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

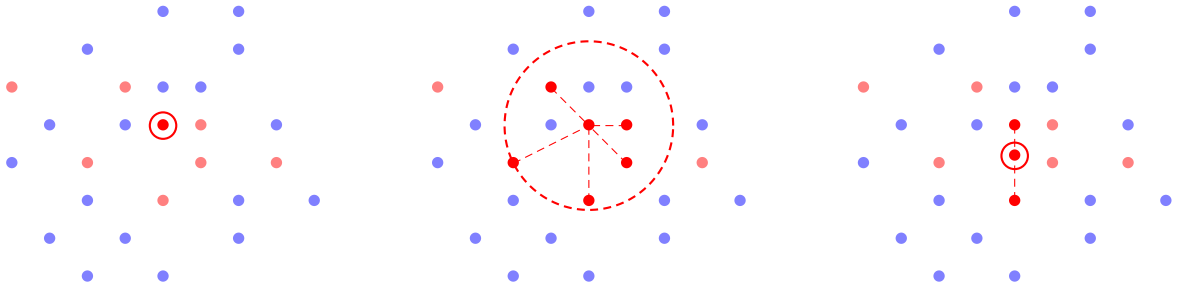
(b) Identification des k=5 voisins les plus proches

Figure – Fonctionnement de SMOTE

Hyper déséquilibre de classe

SMOTE : une approche prometteuse

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

SMOTE et One Hot Encoding

SMOTE est une méthode qui ne travaille que les datasets avec des données quantitatives. Elle fonctionne donc théoriquement sur des informations issues d'un One Hot Encoding mais on ne doit pas l'utiliser en pratique : cela n'a aucun sens. Il faut exploiter une variante de SMOTE : SMOTE-NC pour pouvoir traiter des données qualitatives.

Soit $\mathcal{D}$ un dataset comme défini dans (1) 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 $\min\limits_{x \in \mathcal{D}^-} \|\widetilde{x} - x\| \leqslant \delta$ pour $\delta \geqslant 0$ un paramètre fixé.
On peut montrer, sous l'hypothèse que $\mathcal{D} \sim \mathcal{N}(0_d, I_d)$, que :

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

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.

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

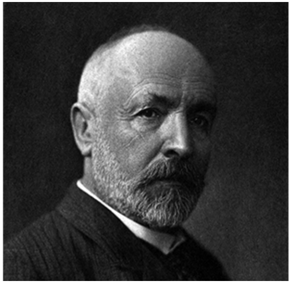
Exercice 3 (Inducteurs d'observations synthétiques ambiguës)

1. *Étudier le comportement asymptotique selon δ. Interpréter.*
2. *Calculer* $\lim\limits_{n^- \to +\infty} \mathbb{P}\left(\min_{x \in \mathcal{D}^-} \|\widetilde{x} - x\| \leqslant \delta\right)$ *et commentez.*
3. *Que peut-on dire de l'évolution de la probabilité selon la dimension du dataset ?*

Travail en grande dimension

Motivation

Tant que vous ne m'aurez pas approuvé, je ne puis que dire : je le vois mais je ne le crois pas.
— Georg Cantor (1877)



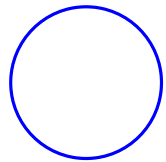
(a) En dimension 1, volume = 2

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

Travail en grande dimension

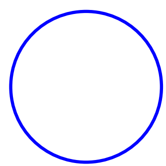
Hypersphère

(a) En dimension 1, volume = 2



(b) En dimension 2, volume = π

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



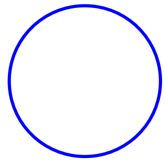
(b) Avec la norme 2

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

Travail en grande dimension

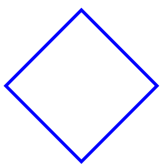
Hypersphère

(a) En dimension 1, volume = 2

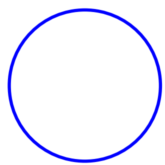


(b) En dimension 2, volume = π

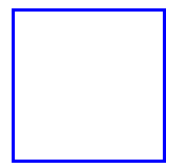
Figure – Représentation et volume d'une hypersphère de rayon 1 dans 3 espaces de dimensions différentes



(a) Avec la norme 1



(b) Avec la norme 2

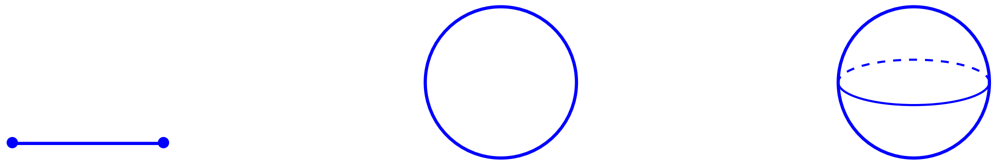


(c) Avec la norme infinie

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

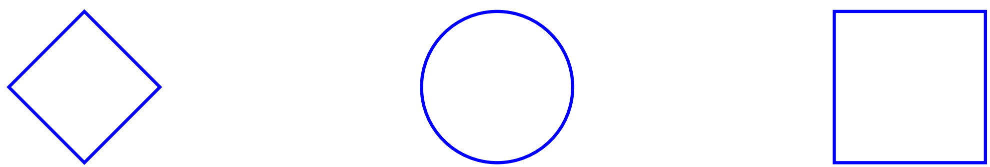
Travail en grande dimension

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

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 :*

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

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 4 (Concentration dans l'hypersphère)

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

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