Substitution du One-Hot Encoding

Groupe de données

February 28, 2025

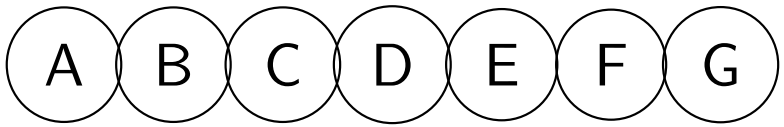
Pourquoi trouver une substitution au One-Hot Encoding ?

Catégorie	One-Hot Encoding Vector
A	[1, 0, 0, 0]
B	[0, 1, 0, 0]
C	[0, 0, 1, 0]
D	[0, 0, 0, 1]

One-Hot Encoding pour ces catégories crée des vecteurs binaires où chaque catégorie est représentée par un vecteur de taille 4 (car il y a 4 catégories). Chaque vecteur a un seul "1" à la position correspondant à la catégorie, et tous les autres éléments sont "0".

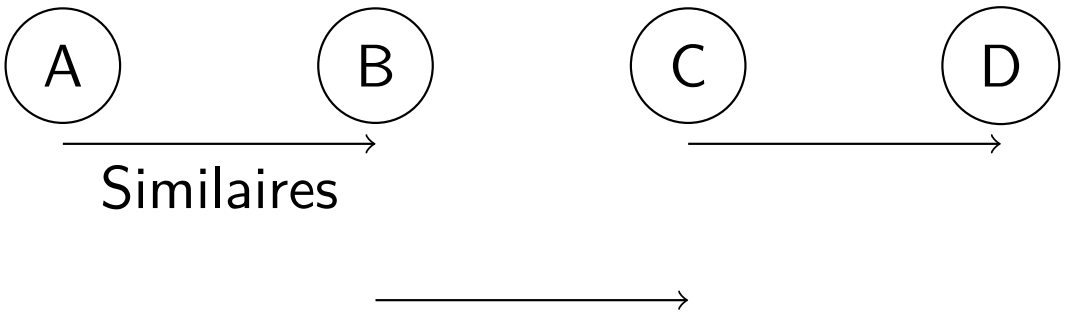
Limites du One-Hot Encoding dans ce contexte:

- **Dimensionnalité élevée** : Le One-Hot Encoding crée une explosion du nombre de colonnes, ce qui augmente la consommation mémoire et le temps de calcul.(A/Upgrade, A/Downgrade..)

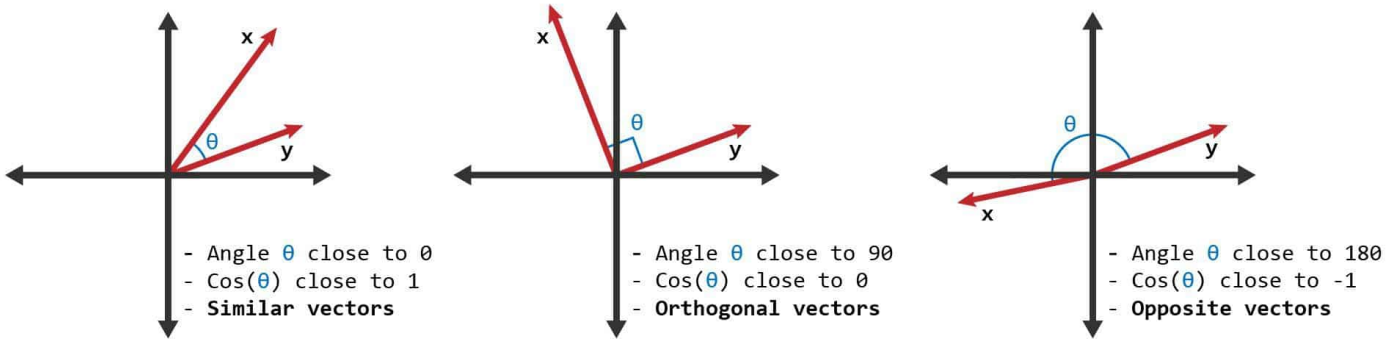


$\longrightarrow$
Explosion de la taille

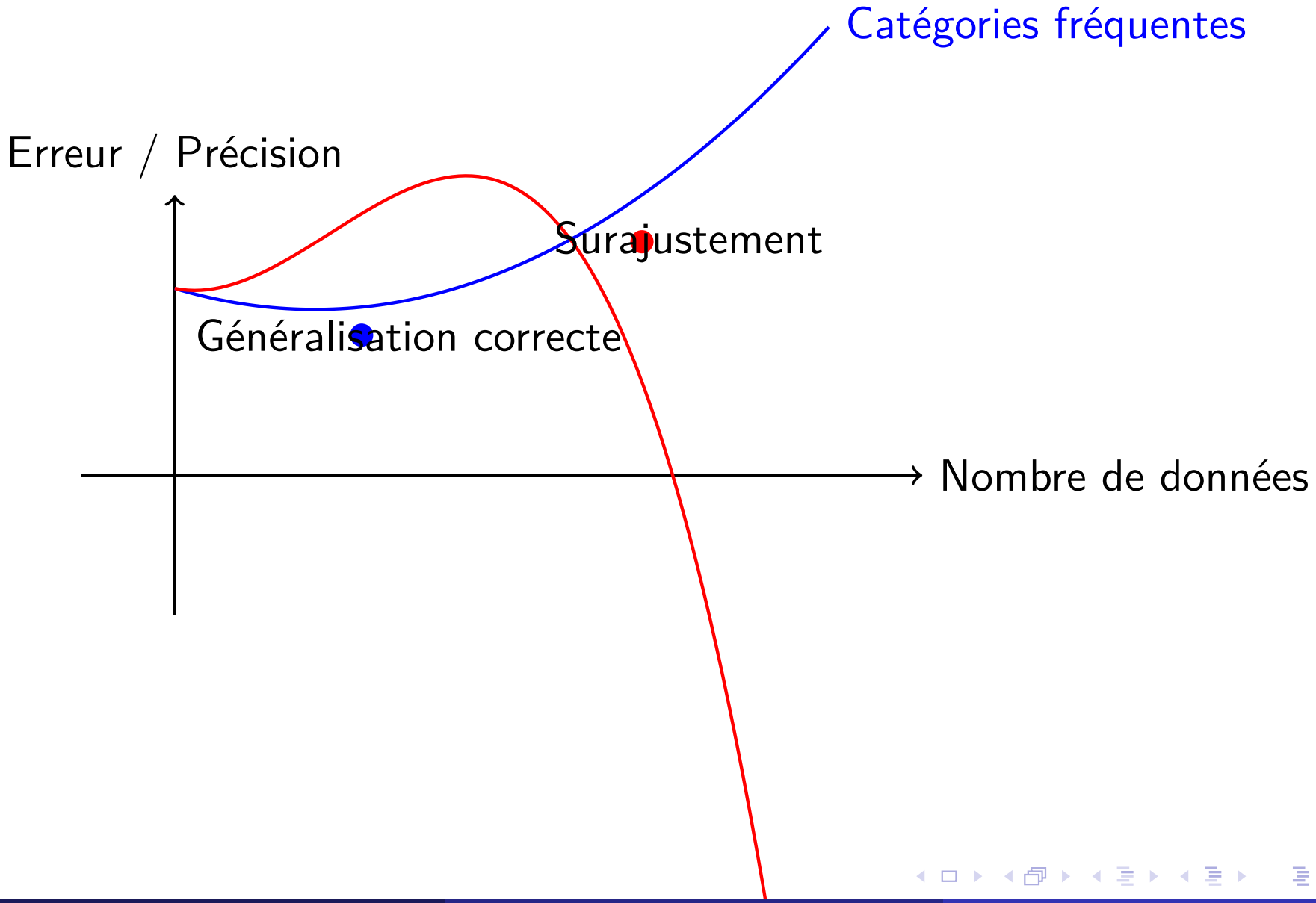
- **Perte de relations de similarité**
  Le One-Hot Encoding traite chaque catégorie comme une entité indépendante sans aucune notion de similitude ou de relation entre elles et ne permettrait pas de saisir Si 2 catégories partagent des caractéristiques proches.



- **Inadéquation avec les métriques de distance** : Le One-Hot Encoding n'est pas compatible avec les métriques comme la similarité cosinus , qui nécessitent des relations continues entre les catégories.



- **Risque de surajustement** : Les catégories rares peuvent entraîner un modèle trop spécifique, réduisant sa capacité à généraliser sur de nouvelles données.



Solutions proposées

- **Embeddings catégoriels** : Utiliser des représentations vectorielles continues pour chaque catégorie, capturant mieux les relations.
- **Encodage par fréquence ou statistique** : Utiliser des moyennes ou des fréquences pour chaque catégorie, réduisant la redondance.

Problème de la classe A

- A minoritaire, pourquoi s'agit il d'un problème?
- Comment adresser ce problème?

	MPN_x	MANUFACTURER_x	Maximum Input Offset Voltage	Maximum Single Supply Voltage	Minimum Single Supply Voltage	Number of Channels per Chip	Supplier_Package_x	Typical Gain Bandwidth Product	MPN_y	MANUFACTURER_y	Supplier_Package_
1	PN-1012382	MN-1030	0.000005	5.5	1.8	1.0	SC-70	400000.0	PN-1018128	MN-1030	SC-7
2	PN-1012382	MN-1030	0.000005	5.5	1.8	1.0	SC-70	400000.0	PN-1018129	MN-1030	SOT-2
3	PN-1012382	MN-1030	0.000005	5.5	1.8	1.0	SC-70	400000.0	PN-1018130	MN-1030	SC-7
4	PN-1012382	MN-1030	0.000005	5.5	1.8	1.0	SC-70	400000.0	PN-1018131	MN-1030	SOT-2
5	PN-1018128	MN-1030	0.000005	5.5	1.8	1.0	SC-70	400000.0	PN-1012382	MN-1030	SC-7
...	...	...	...	...	...	...	...	...	...	...	
3554	PN-1017569	MN-1030	0.010000	16.0	3.0	4.0	SOIC	3500000.0	PN-1017570	MN-1030	TSSC
3555	PN-1017570	MN-1030	0.010000	16.0	3.0	4.0	TSSOP	3500000.0	PN-1017568	MN-1030	SOI
3556	PN-1017570	MN-1030	0.010000	16.0	3.0	4.0	TSSOP	3500000.0	PN-1017569	MN-1030	SOI
3561	PN-102963	MN-1030	0.060000	28.0	4.0	2.0	SO N	350000.0	PN-102964	MN-1030	SO
3562	PN-102964	MN-1030	0.060000	28.0	4.0	2.0	SO N	350000.0	PN-102963	MN-1030	SO

$\Longrightarrow$ 2772Nouvelles observations générées pour la class A