1 Encodage des variables catégorielles

Les modèles de machine learning nécessitent des variables numériques. L'encodage catégoriel consiste à transformer des variables qualitatives en représentations numériques exploitables par les algorithmes.

On distingue :

— Variables **nominales** : sans ordre (ex : couleur, ville)
— Variables **ordinales** : avec ordre (ex : faible $<$ moyen $<$ fort)
— Variables **cycliques** : périodiques (heure, mois, jour)

2 One-Hot Encoding et get_dummies

2.1 Principe

Chaque catégorie est représentée par une variable binaire.

2.2 Formulation mathématique

Soit une variable catégorielle

$$X \in \{c_1, c_2, \ldots, c_K\}$$

On définit :

$$x^{(j)} = \mathbb{1}(X = c_j)$$

Cela génère K variables (ou $K - 1$ pour éviter la colinéarité).

2.3 Quand l'utiliser

— Variables nominales
— Faible cardinalité
— Modèles linéaires, SVM, kNN, réseaux de neurones

2.4 Interprétation

Dans un modèle linéaire :

$$y = \beta_0 + \sum_{j=1}^{K-1} \beta_j x^{(j)}$$

Chaque coefficient β_j mesure l'effet de la catégorie c_j par rapport à la catégorie de référence.

3 Sum Encoding (Deviation Coding)

3.1 Principe

Les catégories sont comparées à la moyenne globale plutôt qu'à une catégorie de référence.

3.2 Propriété

Les coefficients vérifient :

$$\sum_{j=1}^{K} \beta_j = 0$$

3.3 Interprétation

— L'intercept correspond à la moyenne globale

— Chaque coefficient représente un écart à cette moyenne

4 Helmert Encoding

4.1 Principe

Chaque catégorie est comparée à la moyenne des catégories suivantes (selon un ordre choisi).

4.2 Interprétation

Le coefficient mesure la différence entre un niveau et la moyenne des niveaux ultérieurs.

4.3 Usage

— Analyses statistiques

— Variables ordinales ou ordonnées artificiellement

5 Binary Encoding

5.1 Principe

Chaque catégorie reçoit un identifiant entier, converti en représentation binaire.

5.2 Formulation

Si $id(c) \in \{0, \ldots, K-1\}$:

$$id(c) = \sum_{b=0}^{B-1} bit_b \cdot 2^b \quad \text{avec} \quad B = \lceil \log_2 K \rceil$$

5.3 Avantages et limites

— Réduction de dimension

— Introduit des similarités artificielles

6 Hashing Encoding

6.1 Principe

Une fonction de hachage projette les catégories dans un espace de dimension fixée.

6.2 Formulation

$$j = h(c) \bmod n \quad , \quad x_j \leftarrow x_j + 1$$

6.3 Usage

— Très grande cardinalité

— Données textuelles ou logs

6.4 Limite

Présence de collisions rendant l'interprétation difficile.

7 Label Encoding

7.1 Principe

Chaque catégorie est remplacée par un entier :

$$\{c_1, c_2, c_3\} \rightarrow \{0, 1, 2\}$$

7.2 Attention

Introduit un ordre artificiel. À éviter pour les modèles basés sur distance ou linéaires.

8 Ordinal Encoding

8.1 Principe

Encodage numérique respectant un ordre réel :

$$faible = 0, \quad moyen = 1, \quad lev = 2$$

8.2 Interprétation

Une augmentation d'une unité correspond à un changement de niveau.

9 Target / Mean Encoding

9.1 Principe

Chaque catégorie est remplacée par la moyenne de la cible conditionnellement à la catégorie.

9.2 Formulation

$$enc(c) = \mathbb{E}[y \mid X = c] \approx \frac{1}{n_c} \sum_{i:x_i=c} y_i$$

9.3 Risque

Fuite d'information (data leakage) sans validation croisée ou régularisation.

10 Frequency Encoding

10.1 Formulation

$$enc(c) = \frac{n_c}{N}$$

10.2 Interprétation

La valeur encode la popularité de la catégorie.

11 Probability Ratio Encoding

11.1 Formulation

$$enc(c) = \log\left(\frac{P(y=1 \mid c)}{P(y=0 \mid c)}\right)$$

11.2 Interprétation

— Valeur positive : favorise la classe 1

— Valeur négative : favorise la classe 0

12 Weight of Evidence (WoE)

12.1 Formulation

$$WoE(c) = \log\left(\frac{P(X=c \mid y=1)}{P(X=c \mid y=0)}\right)$$

12.2 Usage

Très utilisé en scoring crédit et régression logistique.

13 Leave-One-Out Encoding

13.1 Formulation

$$enc_i = \frac{\sum_{j:x_j=c} y_j - y_i}{n_c - 1}$$

13.2 Avantage

Réduction du sur-apprentissage.

14 James–Stein Encoding

14.1 Formulation

$$enc(c) = \lambda_c \bar{y}_c + (1 - \lambda_c)\bar{y}$$

14.2 Interprétation

Lissage adaptatif vers la moyenne globale.

15 M-estimator Encoding

15.1 Formulation

$$enc(c) = \frac{\sum y_c + m\bar{y}}{n_c + m}$$

16 Thermometer Encoding

16.1 Principe

Encodage cumulatif ordinal :

$$X = 3 \Rightarrow [1, 1, 1, 0, 0]$$

16.2 Formulation

$$x^{(j)} = \mathbb{1}(X \geq c_j)$$

17 Encodage trigonométrique (variables cycliques)

17.1 Formulation

$$x_1 = \sin\left(2\pi \frac{t}{T}\right), \quad x_2 = \cos\left(2\pi \frac{t}{T}\right)$$

17.2 Usage

Heure, mois, jour de la semaine.

18 Bases radiales (RBF)

18.1 Formulation

$$\phi_k(t) = \exp\left(-\frac{d(t,\mu_k)^2}{2\sigma^2}\right)$$

avec distance cyclique :

$$d(t,\mu) = \min(|t-\mu|, T - |t-\mu|)$$

18.2 Interprétation

Chaque base mesure la proximité à un centre.

19 Résumé des choix

- Faible cardinalité : One-Hot
- Forte cardinalité : Target / Frequency / Hashing
- Ordinale : Ordinal ou Thermometer
- Cyclique : Sin/Cos ou RBF
- Interprétation statistique : Sum, Helmert, WoE