Nous continuons l'exploration de l'encodage des données non numériques

Autres codages catégoriels

Imaginons que nous disposons d'une variable âge numérique, il se peut très bien pour l'application que connaître l'âge exacte de la personne ne soit pas nécessaire, on introduit alors une nouvelle représentation de l'information que nous appellerons "tranche d'âge" pour laquelle nous fixons arbitrairement les catégories *enfant* de 0 à 12 ans, *ado* de 12 à 16, *jeune adulte* de 16 à 30, *adulte* de 30 à 60 (ou 65 si la réforme passe) et *sénior* de 60 à 125. Cette nouvelle variable peut bien entendu être représenté, au choix par les valeurs 1 à 5, par un vecteur de taille 5 (et du one hot encoding) ou par un vecteur de taille 3 (codage binaire), mais on peut aussi la représenter par l'âge moyen (ou médian si on s'intéresse à notre échantillon) de la catégorie $6, 14, 23, 45, 92.5$.

"impact encoding" ou "target encoding"

L'idée est d'utiliser un codage de la variable **X** qui reflète la relation existant avec la variable à prédire **y**. Prenons le cas général où la variable **X** possède k modalité $m_1, m_2, \dots m_k$ et que la variable **y** est une variable numérique. Pour coder l'information que $\boldsymbol{X} = m_j$, on va calculer la quantité $E[y|X = m_j]$, ainsi si on dispose de n exemples $\{(x_i, y_i), 1 \leq i \leq n\}$, que n_j soit le nombre d'exemples où $\boldsymbol{X} = m_j$ on aura

$$E[y|X = m_j] = \frac{1}{n_j} \sum_{i=1}^{n_j} y_i$$

La technique du **leave one out encoding** est très similaire, on enlève juste la donnée à traiter pour faire les calculs

Exemple

On souhaite prédire l'humeur en fonction de la météo, on dispose de la base de données suivante:

météo	humeur
soleil	joyeux
soleil	joyeux
soleil	triste
nuageux	triste
nuageux	joyeux
nuageux	triste

météo	humeur
pluie	triste

Supposons que joyeux=1, triste=0. Avec l'approche "impact encoding" *soleil* sera codé $\frac{2}{3}$; *nuageux* sera codé $\frac{1}{3}$ et *pluie* 0

Tandis que dans l'approche "leave one out", la première et la seconde ligne seront codées par $\frac{1}{2}$, tandis que la troisième $\frac{2}{2}$

Le problème de cette approche est un risque de sur-apprentissage des données, mais aussi un problème de robustesse du codage lorsque de nouvelles données sont ajoutées. Par ailleurs dans le petit exemple présenté on n'a qu'un seul cas pour la valeur *pluie* on se retrouve donc à faire une forte hypothèse que lorsqu'il pleut notre humeur sera triste.

Approche IMDb

IMDb est l'abréviation de Internet Movies Database wikipedia l'un de ses objectifs est de fournir le top 250 des films les mieux notés, les concepteurs ont mis au point une formule pour éviter un trop grand bouleversement du classement après la sortie de films récents

$$\frac{R.v + C.m}{v + m}$$

où R est la note moyenne attribuée au film, v est le nombre de votes pour le film, C est la note moyenne obtenue pour tous les films de la base, m est le nombre minimum de votes requis pour apparaître dans le top 250. Avec cette approche le score d'un film lorsqu'il n'y a pas eu suffisamment de votants sera proche du score moyen, et il reflétera la note attribuée par les votants lorsque leur nombre sera suffisant

Cette idée peut s'appliquer au **target encoding** et est connue sous le nom de **smoothing** ou *lissage*

Exercice

à partir de la petite base sur météo et humeur, appliquez la technique du lissage, dans les deux approches *impact encoding* et *leave one out*