

Mini-projet - Classifieur Bayésien Naïf

Thème : Recommandation publicitaire pour une entreprise de pantalons

1. Contexte

Nom : Pantalon+ (P+) - boutique en ligne spécialisée dans les pantalons.

P+ exploite un système de recommandation et d'activation marketing. Lorsqu'un internaute recherche des mots clés liés aux pantalons, le système décide d'**envoyer ou non une publicité ciblée**.

Le CEO souhaite une extension stratégique : **introduire des « pantalons populaires » et organiser des périodes de promotion (Noël, Pâques, Fête des Pères, etc.)**.

2. Objectifs pédagogiques

- Mettre en œuvre un classifieur bayésien naïf (Bernoulli) avec lissage de Laplace.
- Passer du dénombrement à la probabilité (prioris, vraisemblances, postérieurs).
- Concevoir un pipeline robuste (validation, tolérance aux entrées ad hoc, abstention).
- Analyser l'impact de nouvelles features (« populaire », « promo ») et des effets saisonniers.
- Produire des visuels et une argumentation interprétable orientée décision marketing.

3. Données fournies

Fichier : train_pantalon.csv

Colonnes initiales :

- id : identifiant de l'internaute
- pas_cher $\in \{0,1\}$: présence de la notion de prix bas / promotion
- anglais $\in \{0,1\}$: recherche comportant « anglais », « english », etc.
- achat $\in \{\text{OUI}, \text{NON}\}$: achat effectif (ou signal fort de conversion)

Extension (phase 2) :

- populaire $\in \{0,1\}$: requête liée à des pantalons populaires (tendances, best-sellers).
- promo $\in \{0,1\}$: période promotionnelle (Noël, Pâques, Fête des Pères, etc.).

4. Étapes du mini-projet

Phase A - Baseline :

- 1) Features : pas_cher, anglais.
- 2) Estimer : prioris $P(\text{Achat})$, vraisemblances $P(\text{feature}|\text{classe})$ avec Laplace ($\alpha=1$).
- 3) Implémenter : predict_proba, predict, tableau de comptages, matrice de confusion, accuracy.

Phase B - Extension :

- 1) Ajouter les features : populaire, promo.
- 2) Réentraîner le modèle. Comparer : prioris, tableaux $p(1|\text{classe})$, accuracy, confusion.
- 3) Interpréter l'effet des périodes promotionnelles et de la popularité.

5. Livrables attendus

- 1) Notebook exécuté (.ipynb)
- 2) (Optionnel) Résumé PDF (1 page) pour le CEO : résultats clés & recommandations.

Annexe A - Dictionnaire de variables

- pas_cher : 1 si la requête évoque prix bas, réduction, promo ; 0 sinon.
- anglais : 1 si la requête comporte "anglais/english/traduction" ; 0 sinon.
- populaire : 1 si la requête renvoie à des best-sellers / tendances ; 0 sinon.
- promo : 1 si la période est promotionnelle (Noël, Pâques, Fête des Pères, ...) ; 0 sinon.
- achat : {OUI, NON}.

Annexe B - Pseudo-code (Bayes naïf Bernoulli, Laplace $\alpha=1$)

Pour chaque classe c :

- Compter N_c (observations de classe c).
- Pour chaque feature f : compter $N_{\{f=1|c\}}$, $N_{\{f=0|c\}}$.
- $p(f=1|c) = (N_{\{f=1|c\}} + \alpha) / (N_{\{f=1|c\}} + N_{\{f=0|c\}} + 2\alpha)$
- $p(c) = N_c / N$

Pour une observation x :

- $\text{score}(c) = \log p(c) + \sum_f \log p(f=x_f | c)$

- $p(c|x) = \text{softmax}(\text{score})$; prédiction = $\text{argmax}_c p(c|x)$

Annexe C - Exemple (interprétation)

Observation : pas_cher=1, anglais=0, populaire=1, promo=1

- Si $p(1|OUI)$ pour 'promo' est forte, la période promotionnelle peut accroître la probabilité d'achat.
- Discuter l'impact de la corrélation entre 'pas_cher' et 'promo'.