Le Machine Learning pour coder dans une nomenclature: les libellés de caisse

Jérémy L'Hour
Data scientist au SSP Lab
DMCSI, Insee

Insee
Mesurer pour comprendre

Codification automatique de libellés

- **Contexte**: utilisation de données de caisse de la grande distribution pour le calcul de l'Indice des Prix à la Consommation (IPC) et d'indicateurs d'activité mensuelle.
- **Problème**: Comment classer des libellés de caisse dans une nomenclature précise de façon automatique?
- **Objectif** : Construire, de façon supervisée, un algorithme **L_i → f(L_i)** qui, à chaque libellé, associe une catégorie parmi les **K** catégories de la nomenclature (~50 pour les indicateurs d'activité vs. >100 pour l'IPC).
- **Exemples de libellés et de classification en postes COICOP**:

  PIZZA CRUST 580G 4FROM → « pizza, quiche et plats cuisinés à base de céréales »,

  MONSAVON AU LAIT GD PASSION X2 → « savons, dentifrices et produits de toilettes »,

  FROMAGE DE TETE BOL 350G → « autres préparations charcutières », etc.
- **Actuellement** : dans l'IPC, les produits sont classés grâce à un référentiel acheté à une société privée, qui ne couvre pas la totalité des ventes en grandes surfaces – il y a donc une partie labellisée et une partie non labellisée.

Projet en cours.

Phase A: Exploration et Décomposition du Problème

Etape 1: Constitution du dictionnaire

- **Objectif**: passer d'une chaîne de caractères L_i à un vecteur numérique X_i.
- L'information contenue sur un libellé est limitée, mais il faut savoir distinguer ce qui est discriminant.
- Pour cela, on constitue une liste de tokens (« dictionnaire »).
- **Qu'est-ce qu'un token ?** Chaîne de caractères bordée par des espaces

  - assez fréquente (deux apparitions ou plus),

  - assez longue (trois caractères ou plus),

  - dont les caractères numériques sont supprimés,

  - dont la ponctuation est remplacée par un espace.
- Parfois : étape supplémentaire de réduction de dimension via une Décomposition en Valeurs Singulières ou une sélection des mots les plus discriminants via le calcul de la Mutual Information.
- Mais ici, cette étape est longue et les performances sont dégradées.
- **Au final**, X_i compte le nombre d'apparitions des tokens du dictionnaire dans L_i.

Etape 2 : entraînement de l'algorithme

- **Objectif** : Estimer **X_i → f(X_i)** par des méthodes plus ou moins conventionnelles.
- **Mauvaises solutions** :

  - algorithmes à base d'arbres (forêts aléatoires etc.) car implémentés avec sélection aléatoire des variables et mal adaptés à la grande dimension,

  - modèles de type Logit car certaines catégories avec peu d'exemples.
- **Solution basique** : Naive Bayes.
- **Solution la plus performante** : Machine à Vecteur Support (SVM), qui s'adapte très bien à la grande dimension (grâce à l' « astuce du noyau »).
- Sur un échantillon de 30,000 libellés (70 % entraînement / 30 % test) et 43 catégories,

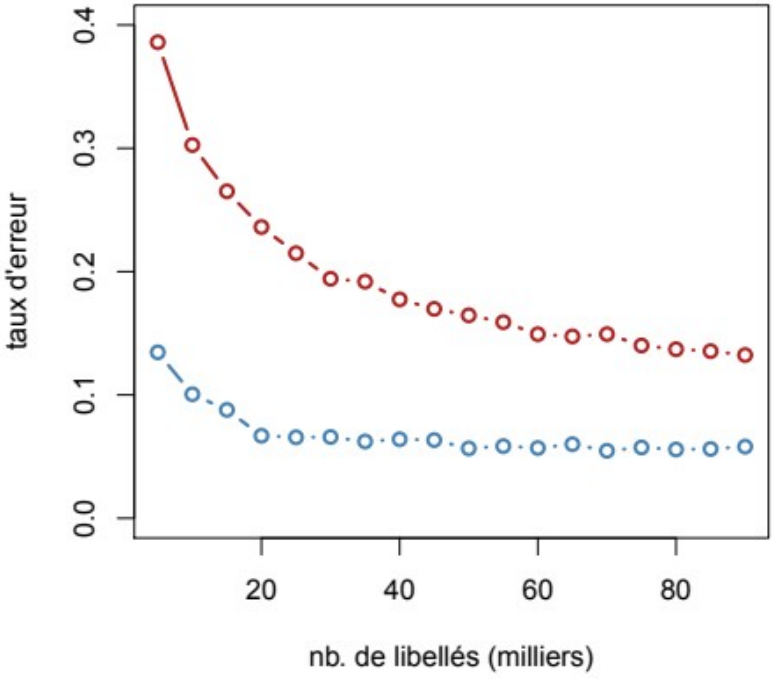
taux d'erreur:

- Naive Bayes : entraînement 18 %, test 27 %,

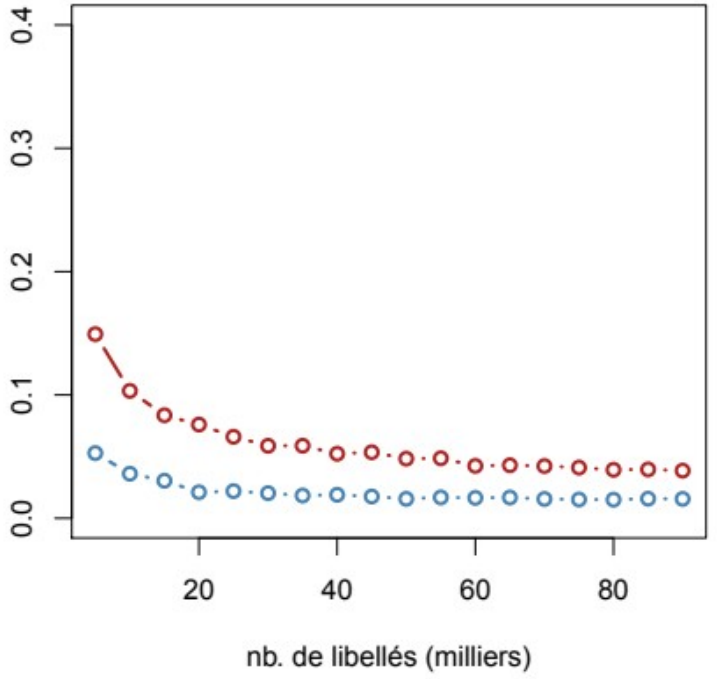
- SVM : entraînement 6 %, test 20 %.

Résultats du SVM

Prédiction parmi 43 catégories



Alimentaire vs. Non-alimentaire



Phase B: Généralisation et Utilisation de fastText

Enjeux pour généraliser / passer à l'échelle

- **Pour généraliser cet algorithme, plusieurs problèmes se posent** :

  - gestion de la grande dimension,

  - gestion des abréviations différentes entre enseignes, pluriels, faute d'orthographe,

  - gestion des mots inconnus,

  - savoir quand on se trompe (avoir une mesure du degré de « confiance » de la prédiction)

- **Solution en cours de test** : module supervisé de fastText (Joulin, Grave, Bojanowski et Mikolov, 2016).

- **Avantages** :

  - très rapide,

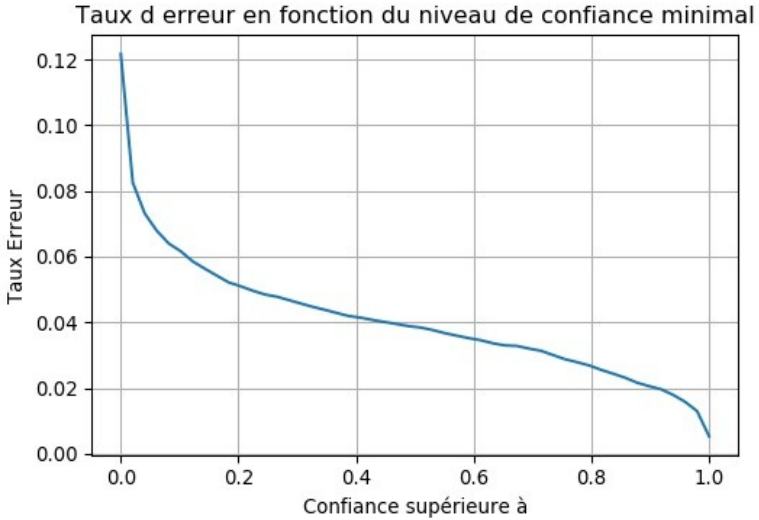
  - très performant,

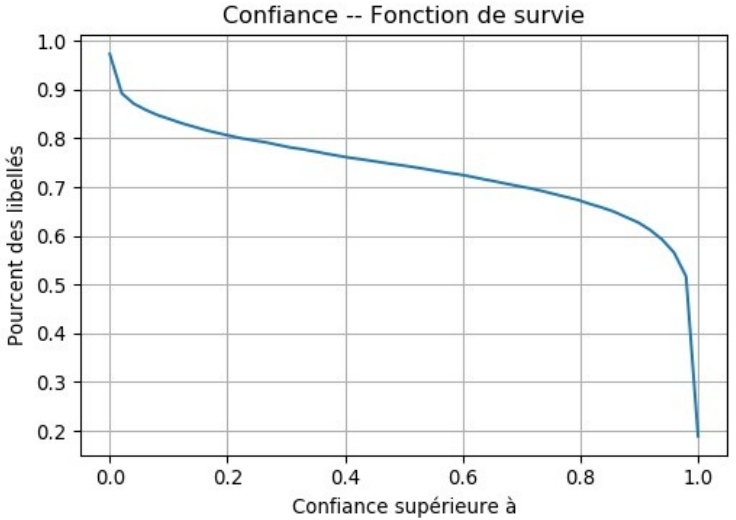
  - s'adapte aux mots inconnus, abréviations, aux fautes d'orthographes,

  - fournit une probabilité d'appartenance à une classe (possibilité de construire un indice de confiance).

Mesure du degré de « confiance » de la prédiction

- Soient p_i(1), …, p_i(K) les probabilités ordonnées d'appartenance à une catégorie L_i.
- Une mesure du degré de confiance de la prédiction est **C_i := p_i(K) – p_i(K-1)**,
- Plus **C_i** est grande, mieux on discrimine entre catégories.
- Inversement, une valeur de **C_i** faible indique soit que l'on ne sait pas discriminer, soit que l'on ne sait rien (cas où p_i(K) est très faible, ex : libellé totalement inconnu).





Note: 70,000 observations, nomenclature COICOP avec 120 catégories

En pratique

Libellé	Vrai label	Label prédit	Probabilité	Confiance
JAMBON FORET NOIRE 350G	Jambon	Jambon	99,1 %	98,8 %
BABA RHUM RHUM AGRUME M&A 320G	Pâtisseries de conservation	Eaux de vie autres que le whisky	97,7 %	94,1 %
BABA AU RHUM RHUM AGRUME M&A 320G	Pâtisseries de conservation	Pâtisseries de conservation	87,4 %	46,7%
BABA RHUM AGRUME M&A 320G	Pâtisseries de conservation	Eaux de vie autres que le whisky	1,2 %	0,4%
BAD ELECTRIQUE FILLE	Appareils électriques pour soins corporels	Appareils électriques pour soins corporels	51,6 %	50,5%
TROUSSE CRAYON GOLDF	Stylos, crayons et encre	Articles de voyage	9,5 %	0,5%
GANT EXTRA CONFORT T9	Vêtements de travail	Vêtements de travail	14,8 %	13,9 %
GANTS EXTRA CONFORT T9	Vêtements de travail	Vêtements de travail	65,1 %	64,7 %
GANT EXTRA	Vêtements de travail	Vêtements de travail	0,4 %	0,2 %

Conclusions

- La classification textuelle repose conceptuellement sur deux étapes clés (constitution du dictionnaire, choix de l'algorithme).
- Problème relativement standard où il a régulièrement été observé la bonne performance des méthodes de type Machine à Vecteur Support à noyau linéaire.
- Enjeux pour construire un outil utile aux services métiers :

  - solution rapide,

  - robuste aux différentes abréviations, fautes etc,

  - avec la capacité de fournir un score de « confiance » de la prédiction (pour décider d'une éventuelle reprise manuelle).
- **Principal problème restant :** pas d'échantillon d'apprentissage pour une vaste partie des produits (vrac, produits frais, habillement, etc.). Possibilité d'enrichir les données via d'autres sources ? Utiliser les informations contenues dans l'EAN ? Mobiliser des moyens humains pour labelliser les observations non labellisés ?

Des questions?

insee.fr