Catégorisation de textes

Text Mining

Ricco Rakotomalala

1. Catégorisation de textes
2. Evaluation des performances
3. Réduction de la dimensionnalité
4. Techniques d'apprentissage supervisé
5. Bibliographique

Document classification, indexation automatique

CATÉGORISATION DE TEXTES

L'objectif de la catégorisation de textes est de pouvoir associer automatiquement des documents à des classes (catégories, étiquettes, index) prédéfinies. Nous nous plaçons dans le cadre de l'apprentissage supervisé.

<document>
<sujet>acq</sujet>
<texte>
Resdel Industries Inc said it has agreed to acquire San/Bar Corp in a share-for-share exchange, after San/Bar distributes all shgares of its Break-Free Corp subsidiary to San/Bar shareholders on a share-for-share basis.
</texte>
</document>
<document>
<sujet>acq</sujet>
<texte>
Warburg, Pincus Capital Co L.P., an investment partnership, said it told representatives of Symbion Inc it would not increase the 3.50-dlr-per-share cash price it has offered for the company. In a filing with the Securities and Exchange Commission, Warburg Pincus said one of its top executives, Rodman Moorhead, who is also a Symbion director, met April 1 with Symbion's financial advisor, L.F. Rothschild, Unterberg, Towbin Inc.
</texte>
</document>

Exemple « Reuters »

- « sujet » est la variable cible
- « texte » est la « variable » prédictive

Objectif : identifier une liaison fonctionnelle…

$$sujet = \Phi(texte, \alpha)$$

…qui soit la plus « efficace » possible.

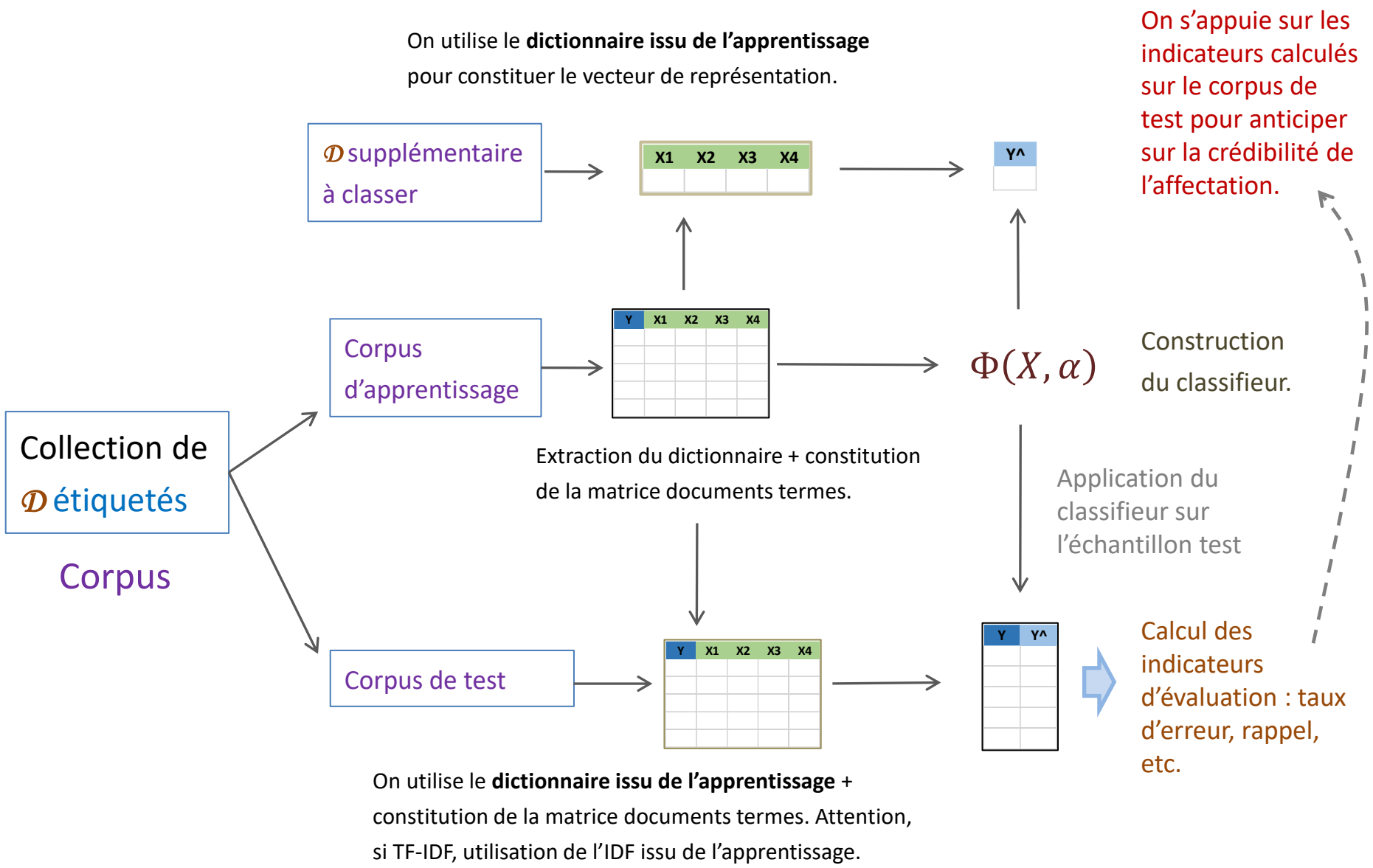
Nous nous situons dans un schéma de classement par le contenu.

La solution la plus simple consiste à s'appuyer sur la représentation en sac de mots pour obtenir une présentation tabulaire des données, propice à l'utilisation des algorithmes usuels de data mining.

N° doc	classes	acquir	acquisit	…	yesterday	yet	yield	york	zinc	zirconium
1	acq	1	0		0	0	0	0	0	0
2	acq	0	0		0	0	0	0	0	0
3	acq	1	0		0	0	0	0	0	0
4	acq	0	0		0	0	0	0	0	0
5	acq	2	0		0	0	0	0	0	0
6	acq	0	1		1	0	0	0	0	0
71	acq	0	1		1	0	0	0	0	0
…										
72	crude	0	0		0	0	0	1	0	0
73	crude	0	0		0	0	0	0	0	0
74	crude	0	0		0	0	0	0	0	0
75	crude	0	0		0	0	0	3	0	0
76	crude	0	0		0	0	0	0	0	0
77	crude	0	0		0	0	0	0	0	0
78	crude	0	0		3	0	0	1	0	0
…										
116	crude	1	0		0	0	0	0	0	0
117	crude	0	0		0	0	0	0	0	0

Variable cible (Y), souvent binaire {+, - }, mais pas forcément.

Variables prédictives numériques (**X**), dimensionnalité élevée, beaucoup de 0.


On utilise le **dictionnaire issu de l'apprentissage** pour constituer le vecteur de représentation.
On s'appuie sur les indicateurs calculés sur le corpus de test pour anticiper sur la crédibilité de l'affectation.
$\mathcal{D}$ supplémentaire à classer
X1 X2 X3 X4
Y^
Corpus d'apprentissage
Y X1 X2 X3 X4
$\Phi(X, \alpha)$
Construction du classifieur.
Collection de $\mathcal{D}$ étiquetés
Corpus
Extraction du dictionnaire + constitution de la matrice documents termes.
Application du classifieur sur l'échantillon test
Corpus de test
Y X1 X2 X3 X4
Y Y^
Calcul des indicateurs d'évaluation : taux d'erreur, rappel, etc.
On utilise le **dictionnaire issu de l'apprentissage** + constitution de la matrice documents termes. Attention, si TF-IDF, utilisation de l'IDF issu de l'apprentissage.

1 Nous sommes souvent dans un schéma une classe à identifier (+) contre les autres (-). En cas de plusieurs classes non exclusives (ex. indexation automatique), nous créons un classifieur binaire pour chaque classe. L'évaluation est spécifique.

2 Celles liées à la constitution de la matrice de documents termes (dimensionnalité, matrice creuse). La variable cible permet de guider la réduction de dimensionnalité.

3 L'étiquetage se fait souvent manuellement, par expertise. Le nombre de documents utilisables pour la modélisation est rare. Le ratio nombre de lignes et de colonnes dans le tableau de données est inversé. Certaines méthodes de data mining seront plus à même d'appréhender cet écueil.

Au-delà du taux d'erreur. Rappel, Précision, F-Mesure.

EVALUATION DES PERFORMANCES

La matrice de confusion résulte de la confrontation entre la classe observée et la classe prédite (sur l'échantillon test c'est mieux).

Y	Y^

Prédite

Observée: Y vs Y^	+^	-^	Σ
+	a	b	a+b
-	c	d	c+d
Σ	a+c	b+d	n

Indicateurs usuels

Taux d'erreur *(error rate)* $\varepsilon = \frac{c+b}{n}$

Traite de manière symétrique les « + » et les « - ».

Rappel *(recall)* $R = recall = \frac{a}{a+b}$

Précision *(precision)* $P = precision = \frac{a}{a+c}$

On sait que rappel et précision sont souvent antinomiques !

		Predicted condition			
	Total population	Predicted Condition positive	Predicted Condition negative	Prevalence $= \frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$	
True condition	condition positive	tp **True positive**	fn **False Negative** (Type II error)	True positive rate (TPR), Sensitivity, Recall, probability of detection $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False negative rate (FNR), Miss rate $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$
	condition negative	fp **False Positive** (Type I error)	tn **True negative**	False positive rate (FPR), Fall-out, probability of false alarm $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$	True negative rate (TNR), Specificity (SPC) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$
	Accuracy (ACC) $= \frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$	Positive predictive value (PPV), Precision $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Test outcome positive}}$	False omission rate (FOR) $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Test outcome negative}}$	Positive likelihood ratio (LR+) $= \frac{\text{TPR}}{\text{FPR}}$	Diagnostic odds ratio (DOR) $= \frac{\text{LR+}}{\text{LR−}}$
		False discovery rate (FDR) $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Test outcome positive}}$	Negative predictive value (NPV) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Test outcome negative}}$	Negative likelihood ratio (LR−) $= \frac{\text{FNR}}{\text{TNR}}$	

(Wikipédia)

Lorsque le classifieur sait fournir un score (proportionnel à la probabilité d'être positif), il est possible de définir des scénarios de seuils d'affectation et de calculer les rappels et précisions associés.

Classer les données selon un score décroissant

Individu	Score (+)	Classe
1	1	+
2	0.95	+
3	0.9	+
4	0.85	-
5	0.8	+
6	0.75	-
7	0.7	-
8	0.65	+
9	0.6	-
10	0.55	-
11	0.5	-
12	0.45	+
13	0.4	-
14	0.35	-
15	0.3	-
16	0.25	-
17	0.2	-
18	0.15	-
19	0.1	-
20	0.05	-

Positifs = 6
Négatifs = 14

Seuil = 1

	^positif	^négatif	Total
positf	1	5	6
négatif	0	14	14
Total	1	19	20

recall = 1/6 = 0.2 ; precision = 1/1 = 1

Seuil = 0.95

	^positif	^négatif	Total
positf	2	4	6
négatif	0	14	14
Total	2	18	20

recall = 2/6 = 0.33 ; precision = 2/2 = 1

Seuil = 0.9

	^positif	^négatif	Total
positf	3	3	6
négatif	0	14	14
Total	3	17	20

recall = 3/6 = 0.5 ; precision = 3/3 = 1

Seuil = 0.85

	^positif	^négatif	Total
positf	3	3	6
négatif	1	13	14
Total	4	16	20

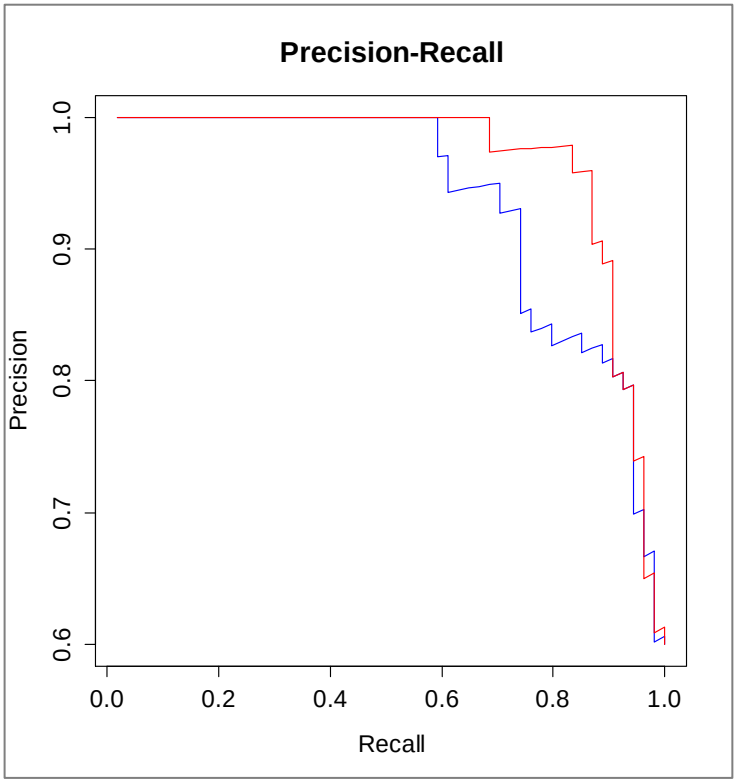
recall = 3/6 = 0.5 ; precision = 3/4 = 0.75

Seuil = 0

	^positif	^négatif	Total
positf	6	0	6
négatif	14	0	14
Total	20	0	20

recall = 6/6 = 1 ; precision = 6/20 = 0.3

La courbe précision-rappel permet de situer l'arbitrage entre le rappel et la précision à mesure que l'on augmente la taille de la cible (le nombre de documents classés positifs).



Courbe précision-rappel sur échantillon test. Données Reuters.

- Nombre de termes : 1085 (après filtrage sur la fréquence ≥ 2)
- Ech. Apprentissage : 27 doc.
- Ech. Test : 90 doc.
- Pondération : TF
- « + » = ACQ

SVM (Radial)

SVM (Linéaire)

L'interprétation directe n'est pas très intéressante. L'outil sert surtout à comparer les performances des classifieurs.

F-Mesure (F_1) : combinaison du rappel et de la précision

On veut se focaliser sur l'indentification des « + ». La F-Mesure (*F-score, F-Measure*) est une mesure synthétique qui combine la précision et le rappel. Elle donne la même importance aux deux critères.

$$F_1 = 2\frac{precision \times recall}{precision + recall}$$

$$F_1 = 2\frac{1}{\frac{1}{precision} + \frac{1}{recall}}$$

Qui correspond en fait à une moyenne harmonique.

Y vs Y^	+^	-^	Σ
+	46	6	52
-	11	27	38
Σ	57	33	90

$$\varepsilon = \frac{11+6}{90} = 0.19$$

$$recall = \frac{46}{46+6} = 0.88$$

$$precision = \frac{46}{46+11} = 0.81$$

Un indicateur unique facilite la comparaison entre classifieurs.

$$F_1 = 2\frac{0.81 \times 0.88}{0.81 + 0.88} = 0.84$$

La F_β permet de moduler l'importance que l'on donne à la précision et au rappel avec une moyenne harmonique pondérée.

$$F_\beta = (1+\beta^2)\frac{precision \times recall}{(\beta^2 \times precision) + recall}$$

- β=1, même importance
- β=2, deux fois plus d'importance au rappel que le précision
- β=0.5, deux fois plus d'importance à la précision
- Etc.

Y vs Y^	+^	-^	Σ
+	46	6	52
-	11	27	38
Σ	57	33	90

$\varepsilon = 0.19 \quad recall = 0.88 \quad precision = 0.81$

$$F_1 = 2\frac{0.81 \times 0.88}{0.81 + 0.88} = 0.84$$

$$F_2 = (1+2^2)\frac{0.81 \times 0.88}{2^2 \times 0.81 + 0.88} = 0.86$$

$$F_{0.5} = (1+0.5^2)\frac{0.81 \times 0.88}{0.5^2 \times 0.81 + 0.88} = 0.82$$

! On peut aussi y voir une manière d'introduire les coûts de mauvaise affectation.

Dans certains cas, nous traitons un problème où nous devons traiter **C** classes mutuellement exclusives (multi-classes) ou non mutuellement exclusives (multi-labels) (ex. indexation automatique, un document peut être associé à plusieurs thématiques). Nous pouvons calculer des rappels et précisions par classe.

Y vs Y^	C_1^	C_2^	C_3^	Σ
C_1	30	50	20	100
C_2	20	60	20	100
C_3	10	10	80	100
Σ	60	120	120	300

$$R_1 = \frac{30}{30+50+20} = \frac{30}{100} = 0.30$$

$$P_1 = \frac{30}{30+20+10} = \frac{30}{60} = 0.50$$

$$R_2 = \frac{60}{100} = 0.60 \qquad P_2 = \frac{60}{120} = 0.50$$

$$R_3 = \frac{80}{100} = 0.80 \qquad P_3 = \frac{80}{120} = 0.67$$

Comment combiner ces résultats pour disposer d'un indicateur de performance unique ?

On effectue une moyenne non pondérée des mesures, en considérant que les catégories ont la même prévalence.

$$R_{macro} = \frac{1}{C}\sum_{c=1}^{C} R_c \qquad P_{macro} = \frac{1}{C}\sum_{c=1}^{C} P_c$$

On peut déduire la F-Mesure à partir de ces deux valeurs

Exemple

Y vs Y^	C_1^	C_2^	C_3^	Σ
C_1	30	50	20	100
C_2	20	60	20	100
C_3	10	10	80	100
Σ	60	120	120	300

$$R_1 = \frac{30}{100} = 0.30 \qquad R_2 = \frac{60}{100} = 0.60 \qquad R_3 = \frac{80}{100} = 0.80$$

$$R_{macro} = \frac{1}{3}(0.3 + 0.6 + 0.8) = 0.567$$

$$P_{macro} = \frac{1}{3}(0.5 + 0.5 + 0.67) = 0.556$$

Remarques

On donne le même poids à toutes les classes :

- Bien pour ne pas masquer les résultats sur les classes rares
- Pas bien parce qu'une partie de l'information est perdue (une classe rare pèse autant qu'une classe très présente)

On effectue une moyenne pondérée des mesures, en considérant que la prévalence des catégories.

$$R_{weigthed} = \frac{1}{n}\sum_{c=1}^{C} n_c \times R_c \qquad P_{weighted} = \frac{1}{n}\sum_{c=1}^{C} n_c \times P_c$$

On peut déduire la F-Mesure à partir de ces deux valeurs

Y vs Y^	C_1^	C_2^	C_3^	Σ
C_1	30	50	20	100
C_2	20	60	20	100
C_3	10	10	80	100
Σ	60	120	120	300

$$R_1 = \frac{30}{100} = 0.30 \qquad R_2 = \frac{60}{100} = 0.60 \qquad R_3 = \frac{80}{100} = 0.80$$

$$R_{weighted} = \frac{1}{300}(100 \times 0.3 + 100 \times 0.6 + 100 \times 0.8) = 0.567$$

$$P_{weighted} = \frac{1}{300}(100 \times 0.5 + 100 \times 0.5 + 100 \times 0.67) = 0.556$$

Remarques

- Bien : corrige ‘’macro-average’’ en tenant compte du poids des classes
- Pas bien : le F-Score global peut ne pas être compris entre les rappels et précisions des classes (R_c, P_c)

Construire une matrice de confusion globale à partir des (TP, FP, FN, TN) par classe, puis en déduire le rappel et la précision.

	+^	-^
+	$TP = \sum_{c=1}^{C} TP_c$	$FN = \sum_{c=1}^{C} FN_c$
-	$FP = \sum_{c=1}^{C} FP_c$	$TN = \sum_{c=1}^{C} TN_c$

$$R_{micro} = \frac{TP}{TP + FN}$$

$$P_{micro} = \frac{TP}{TP + FP}$$

On peut alors en déduire la F-Mesure

Exemple

Y vs Y^	C_1^	C_2^	C_3^	Σ
C_1	30	50	20	100
C_2	20	60	20	100
C_3	10	10	80	100
Σ	60	120	120	300

	+^	-^
+	$30 + 60 + 80 = 170$	$(50 + 20) + (20 + 20) + (10 + 10) = 130$
-	$(20+10)+(50+10)+(20+20)=130$	$(60+20+10+80)+(30+20+10+80)+(30+50+20+60)=470$

$$R_{micro} = \frac{170}{170 + 130} = 0.567$$

$$P_{micro} = \frac{170}{170 + 130} = 0.567$$

Remarque

On donne un poids identique à tous les documents cette fois-ci, qu'ils appartiennent à une classe rare ou pas. Une classe à forte prévalence peut écraser les résultats.

Stratégies adaptées au cadre de l'apprentissage supervisé

RÉDUCTION DE LA DIMENSIONNALITÉ

Dans le schéma classe d'intérêt (+) contre les autres (-), nous nous intéressons surtout aux occurrences des termes qui permettent de désigner la classe (+). De fait, on peut se contenter de constituer le dictionnaire relative à cette classe. Le nombre de termes sera forcément réduit.

Base Reuters
117 doc., 71 acq (+), 46 crude (-)
Dictionnaire : 2435 termes (sans filtrage sur fréquence)

	pred.glob	
	acq	crude
acq	46	6
crude	12	26

F_1=0.836

Constitution du dictionnaire à partir uniquement des documents relatifs à acq : 1837 termes (sans filtrage sur les fréquences)

	pred.loc	
	acq	crude
acq	45	7
crude	10	28

F_1=0.841

On augmente rarement les performances, mais la réduction de dimension y est.

Elle intervient juste après la constitution de la matrice documents termes. Elle évalue la liaison de chaque descripteur, pris individuellement, avec la variable cible. On gère la pertinence, pas la redondance. Elle n'a de sens que pour la pondération binaire (*si utilisation de l'entropie, cf. page suivante ; mais autre critère possible [ex. rapport de corrélation] si autre pondération - voir doc. de Scikit-Learn : SelectKBest*).

Avantages

- Rapidité, simplicité des calculs
- Capacité à traiter des grandes bases
- Généricité (censée fonctionner quelle que soit la méthode en aval)

Inconvénients

- Généricité très questionnable
- Calculs pour pondération binaire valables pour les autres ?
- Ne gère pas la redondance

Y \ X	x_1	…	x_l	…	x_L	Σ
y_1						
⋮			⋮			
y_k		…	n_{kl}	…		$n_{k.}$
⋮			⋮			
y_K						
Σ			$n_{.l}$			n

Fréquences conjointes et marginales

$$p_{kl} = \frac{n_{kl}}{n} \qquad p_{k.} = \frac{n_{k.}}{n} \qquad p_{.l} = \frac{n_{.l}}{n}$$

Information mutuelle (~ covariance [liaison])

$$I(Y,X) = \sum_k \sum_l p_{kl} \times \log_2 \frac{p_{kl}}{p_{k.} \times p_{.l}}$$

Entropie (~ écart-type [dispersion])

$$H(Y) = -\sum_k p_{k.} \log_2 p_{k.}$$

Incertitude symétrique
(~ corrélation)

$$\rho_{y,x} = 2 \times \left[\frac{I(Y,X)}{H(Y) + H(X)} \right]$$

Varie entre [0 ; 1]

Test de significativité

$$G = 2 \times n \times \ln(2) \times I(Y,X)$$

Sous H0 : indépendance entre X et Y, suit une loi du χ^2 à (K-1)*(L-1) degrés de liberté

Dans notre cas, L = 2 toujours, K = 2 très souvent.

Etapes :

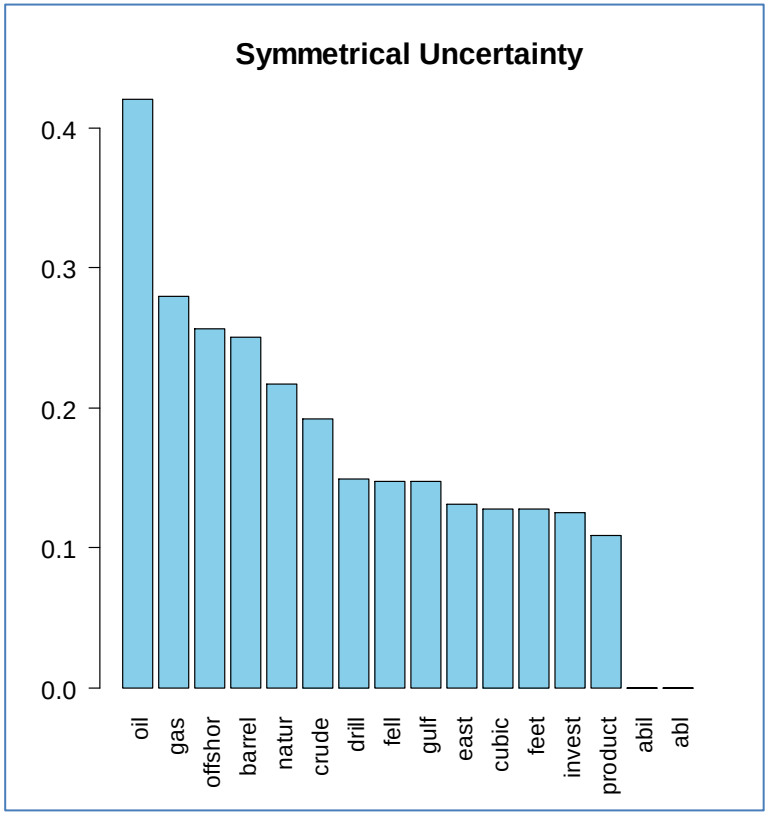
1. Calculer le critère ρ pour chaque descripteur
2. Les classer par ordre décroissant selon ρ
3. Ne retenir que les descripteurs les plus significatifs (???)

L'écueil est là : test stat. souvent trop permissif, seuil sur ρ difficile à cerner ; « sauts » sur l'évolution du ρ ?

Base **Reuters**, pondération binaire, dictionnaire global, filtrage sur les fréquences (≥ 2).

14 termes se démarquent.

(Package Fselector) sous R

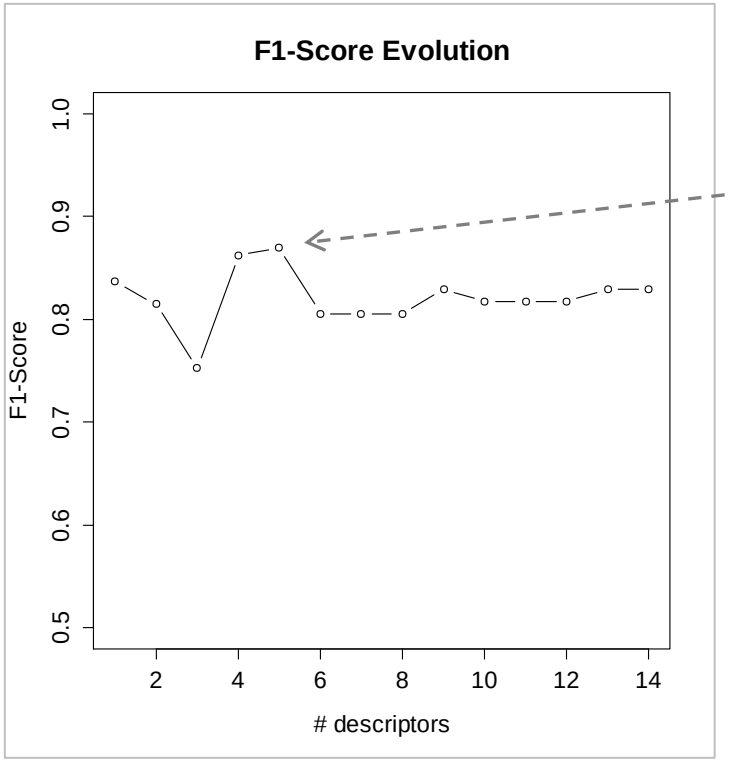


Idée :

1. S'appuyer sur le critère ρ pour ordonner les descripteurs
2. Les introduire au fur et à mesure dans le modèle (forward) et surveiller les performances en test (ou CV, bootstrap…)
3. On choisit la configuration qui optimise la performance

On tient compte des caractéristiques de la méthode d'apprentissage pour choisir la bonne solution. !

Base **Reuters**
SVM Linéaire



La solution avec les 5 premiers descripteurs semble être la meilleure. Mais attention, la courbe est très instable dans le schéma échantillon-test, surtout avec un effectif faible (60 obs. en test dans notre exemple). On a intérêt à passer par des méthodes d'évaluation à plus faible variance (ex. bootstrap).

Quelles sont les techniques adaptées à la catégorisation de textes ?

TECHNIQUES D'APPRENTISSAGE SUPERVISÉ

Peuvent convenir les méthodes fortement régularisées (pour éviter le sur apprentissage, la colinéarité,…) et capables techniquement d'appréhender les grandes dimensions.

- SVM parce que le paramètre de coût permet de contrôler l'influence de l'ensemble d'apprentissage, et elle s'appuie sur la matrice de Gram (matrice des produits scalaires entre les vecteurs individus) pour les calculs.

- Classifieurs linéaires régularisés (ex. Ridge, Lasso…) s'appuyant sur la descente du gradient (seul le vecteur gradient est nécessaire). Régression PLS.

- Les méthodes d'induction de règles (y compris les arbres de décision). Pas de matrice à construire, capacité à réaliser une sélection de variables drastique.

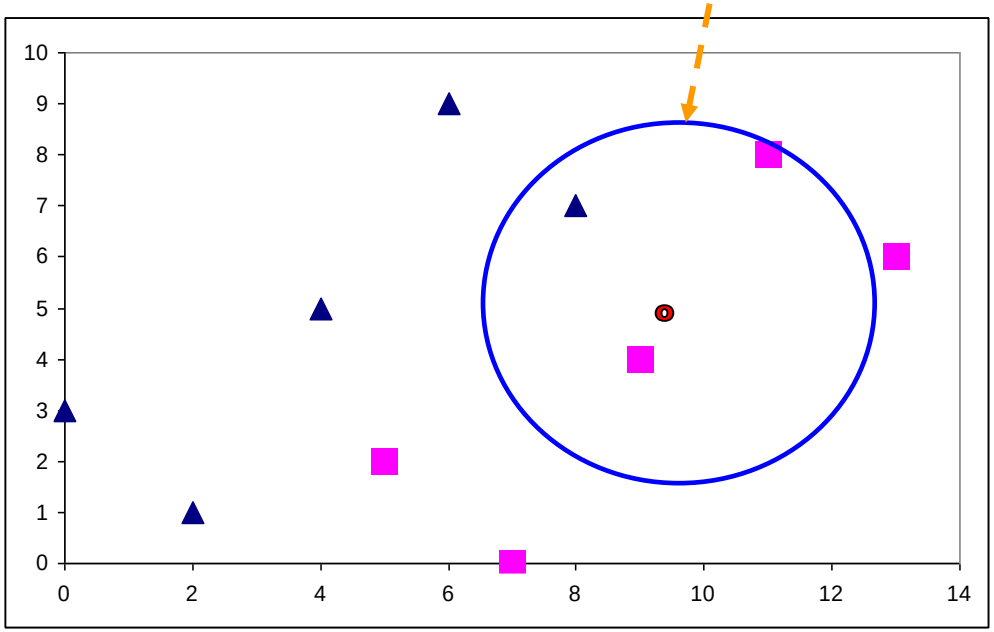
- Et d'autres méthodes que l'on retrouve dans la plupart des références et que nous allons voir maintenant.

K-NN – K nearest neighbors

MÉTHODE DES K PLUS PROCHES VOISINS (K-PPV)

La méthode des K-ppv fait partie des méthodes d'analyse discriminante non paramétrique. Elle ne fait pas d'hypothèses sur les distributions.

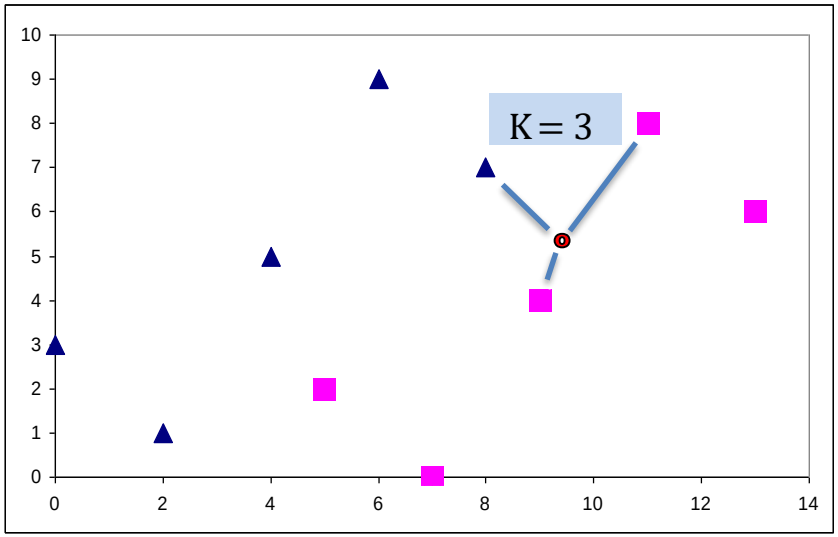
Principe : définir un voisinage autour du point « o » à classer et **estimer localement les probabilités**.



Questions clés

- Comment définir le voisinage ?
- Quelle distance utiliser ?

Paramètre **K** : nombre d'obs. à considérer autour du point à classer



$$P(Y=\Delta \mid X) = 1/3$$

$$P(Y=\blacksquare \mid X) = 2/3$$

Il n'y a pas de modèle explicite (*lazy learning*, *instance based learning*), on passe en revue la totalité de la base d'apprentissage pour chaque individu à classer (*des stratégies pour éviter d'avoir à le faire existent*).

Commentaires

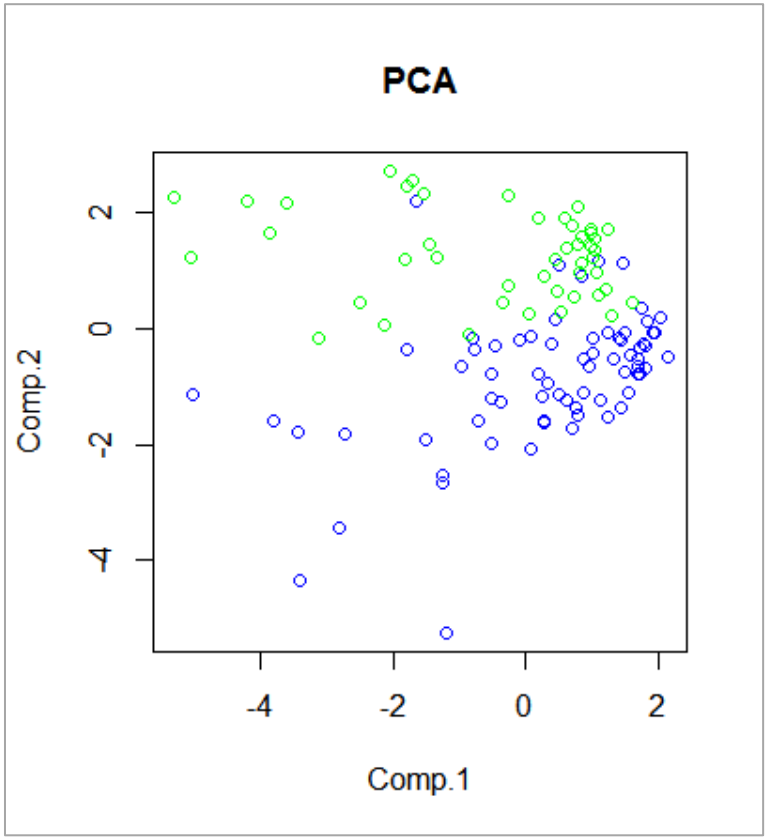
Le choix de K n'est pas toujours évident. On sait que K ↘, moins de biais, plus de variance ; K ↗ moins de variance, plus de biais.

K = 1, diagramme de Voronoï (zone d'influence autour de chaque point).

K = 1, asymptotiquement (taille échantillon d'apprentissage → ∞), $\text{Err}(1\text{-PPV}) < 2 \times \text{Err}(\text{modèle bayésien idéal})$

Base **Reuters**, pondération binaire, dictionnaire global, filtrage sur les fréquences (≥ 2). 57 obs. TRAIN, 60 obs. TEST.

Position des points dans le 1er plan factoriel ➔



3-NN ➔ F1-Score = 0.892

SVM (linéaire) ➔ F1-Score = 0.886

Dans ce contexte précis, les deux approches se tiennent. Souvent les K-NN ont une bonne tenue et peuvent servir de référence avant de se lancer dans des méthodes sophistiquées.

Avantages

- Fortes accointances avec la recherche d'information, peut bénéficier des technologies développées à cet effet (ex. index inversé)
- Simplicité, pas d'apprentissage coûteux
- Peut servir de référence pour situer les autres méthodes d'apprentissage
- Peut bénéficier des distances adaptées au text mining

Inconvénients

- Difficulté à interpréter le classement
- Déploiement nécessite la disponibilité de la base d'apprentissage (ex. PMML)
- Lenteur en classement (si programmé naïvement) (cf. Nearest Neighbor Search)
- Attention à la dimensionnalité et aux variables non pertinentes

Pistes d'amélioration

- Les techniques de réduction jouent un rôle important, ceux vus dans ce support, mais aussi d'autres (ex. LSI - latent semantic indexing, topic model)
- Il est possible de pondérer l'influence des observations selon leur éloignement dans le calcul de la probabilité d'affectation

Affectation au centroïde (barycentre conditionnel) le plus proche

CLASSIFIEUR DE ROCCHIO

Principe

- Chaque classe est représenté par un prototype (typiquement le barycentre conditionnel calculé sur l'échantillon d'apprentissage, *mais Rocchio peut introduire une variante où l'on arbitre entre l'influences des obs. de la classe et les autres*)
- Un individu supplémentaire à classer est affecté au groupe dont le prototype est le plus proche (*différentes mesures de similarité peuvent être utilisées*)

Remarques

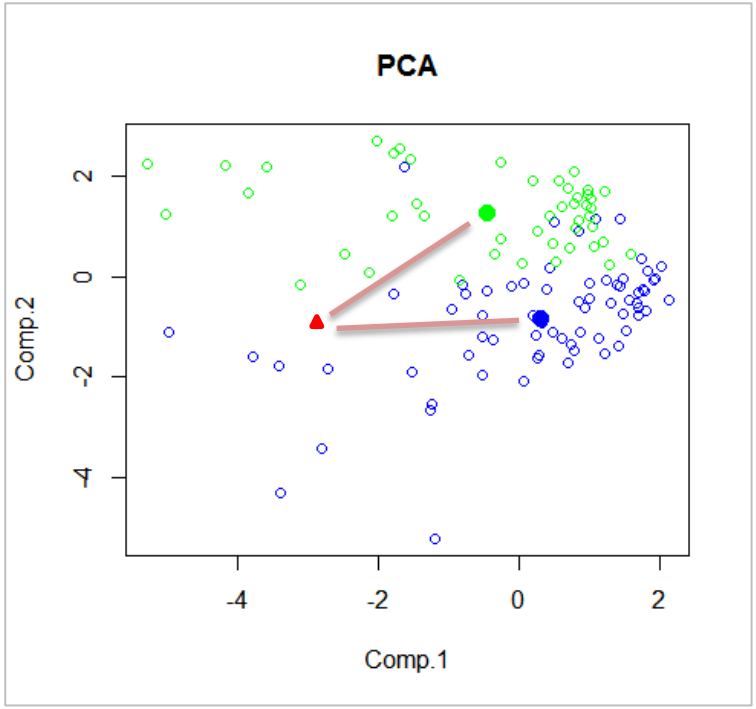
- Similitude avec les K-NN, sauf que l'on s'appuie exclusivement sur les centroïdes
- Similitude avec l'analyse discriminante linéaire, mais distance ne tenant pas compte de la forme des nuages des points. On a un séparateur linéaire. !

Avantages

- Simplicité et rapidité des calculs en apprentissage
- Facilité de déploiement (les vecteurs centroïdes suffisent pour le classement)
- Peut servir également de méthode de référence.

Inconvénients

- Pas de modèle explicite
- Ne se prête pas à l'interprétation lors du classement
- Pas très efficace lorsque les frontières entre les classes sont non linéaires



Plutôt vert ou plutôt bleu ?

Attention

Il faut calculer les barycentres conditionnels dans l'espace originel puis les projeter comme points supplémentaires dans le plan factoriel.

!

Modèle d'indépendance conditionnelle

CLASSIFIEUR BAYÉSIEN NAÏF

(NAIVE BAYES CLASSIFIER)

Règle de décision du maximum a posteriori

$$\hat{y} = \arg\max_{c} P(Y = c/X_1X_2 \cdots Xp) \qquad c \in \{1, \cdots, C\}$$

Or la probabilité conditionnelle peut s'écrire (théorème de Bayes)

$$P(Y = c/X_1X_2 \cdots Xp) = \frac{P(Y = c) \times P(X_1X_2 \cdots Xp\,/Y = c)}{P(X_1X_2 \cdots Xp)}$$

Il s'agit d'identifier le max selon c, or le dénominateur ne dépend pas de c

$$\hat{y} = \arg\max_{c}\ P(Y = c) \times P(X_1X_2 \cdots X_p\,/\,Y = c)$$

Probabilité a priori, facile à estimer, par ex. avec l'estimateur lissé des probabilités (règle de succession de Laplace)

$$\hat{P}(Y = c) = P(c) = \frac{N_c + 1}{N + C}$$

La vraie difficulté est ici. La réponse passe par : quelle hypothèse introduire pour rendre cette probabilité calculable ?

Hypothèse : On considère que les descripteurs sont deux à deux indépendants à Y fixé.

$$P(X_1 X_2 \cdots Xp \,/\, Y = c) = \mathrm{P}(\mathrm{X}_1/\mathrm{Y} = c) \times \mathrm{P}(\mathrm{X}_2/\mathrm{Y} = c) \times \cdots \times \mathrm{P}(\mathrm{X}_\mathrm{p}/\mathrm{Y} = c)$$

$$= \prod_{j=1}^{p} P(X_j/Y = c)$$

En passant par les logarithmes, l'équation dont il faut identifier le maximum devient :

$$\delta(c, \mathcal{X}) = \ln P(c) + \sum_{j=1}^{p} \ln P(X_j/Y = c)$$

! Des solutions existent pour X_j qualitatives et quantitatives. Les expressions sont grandement simplifiées pour X_j binaires c.-à-d. lorsque nous construisons la matrice documents termes avec la pondération binaire.

	classes	P(c)
acq	71	0.605
crude	46	0.395

$$P(acq) = \frac{71+1}{117+2} = 0.605$$

$$P(crude) = \frac{46+1}{117+2} = 0.395$$

Ici aussi, on utilise l'estimateur laplacien des probabilités.

	oil	
	0	1
acq	51	20
crude	1	45

	oil	
	0	1
acq	0.712	0.288
crude	0.042	0.958

$$P(oil = 1/acq) = \frac{20+1}{71+2} = 0.288$$

$$P(oil = 1/crude) = \frac{45+1}{46+2} = 0.958$$

	gas	
	0	1
acq	57	14
crude	9	37

	gas	
	0	1
acq	0.795	0.205
crude	0.208	0.792

$$P(gas = 0/acq) = \frac{57+1}{71+2} = 0.795$$

$$P(gas = 0/crude) = \frac{9+1}{46+2} = 0.208$$

Oil = 1, Gas = 0

➔ Classe = ?

$$\delta(acq, \mathcal{X}) = \ln 0.605 + \ln 0.288 + \ln 0.795 = -1.978$$

$$\delta(crude, \mathcal{X}) = \ln 0.395 + \ln 0.958 + \ln 0.208 = -2.540$$

$$\hat{Y} = acq$$

Décortiquons la fonction de classement pour une seule variable explicative X binaire :

$$d(c,X) = \ln \hat{P}(c) + \ln P(X = 1/Y = c) \times X + \ln P(X = 0/Y = c) \times (1 - X)$$

Ainsi, si X = 1 ➡ $d(c,X) = \ln \hat{P}(c) + \ln P(X = 1/Y = c)$

si X = 0 ➡ $d(c,X) = \ln \hat{P}(c) + \ln P(X = 0/Y = c)$

Il est donc possible d'écrire une expression explicite de la fonction de classement à l'aide des indicatrices X_j :

$$d(c,\mathcal{X}) = \underbrace{\left[\ln \hat{P}(c) + \sum_{j=1}^{p} \ln P(X_j = 0/Y = c)\right]}_{\mathbf{a_0}} + \sum_{j=1}^{p} \underbrace{\ln \frac{P(X = 1/Y = c)}{P(X = 0/Y = c)}}_{\mathbf{a_j}} \times Xj$$

Reuters

$$\mathrm{d}(acq,\mathcal{X}) = -76.927 - 0.907 \times oil - 1.352 \times gas - 4.277 \times offshor + \cdots$$

$$\mathrm{d}(crude,\mathcal{X}) = -90.991 + 3.135 \times oil + 1.335 \times gas - 0.601 \times offshor + \cdots$$

Dans le cas où la cible est binaire Y ∈ {+, -}, on peut déduire une fonction score $d(\mathcal{X})$ unique:

$$d(\mathcal{X}) = d(+,\mathcal{X}) - d(-,\mathcal{X})$$

$$\begin{cases} d(+,\mathcal{X}) = a_0 + a_1.X_1 + a_2.X_2 + \ldots + a_p.X_p \\ d(-,\mathcal{X}) = b_0 + b_1.X_1 + b_2.X_2 + \ldots + b_p.X_p \end{cases}$$

$$d(\mathcal{X}) = c_0 + c_1.X_1 + c_2.X_2 + \ldots + c_p.X_p$$

La règle d'affectation est simplifiée

Si $d(X) \geq 0$ Alors $\hat{y} = +$ Sinon $\hat{y} = -$

Intérêt

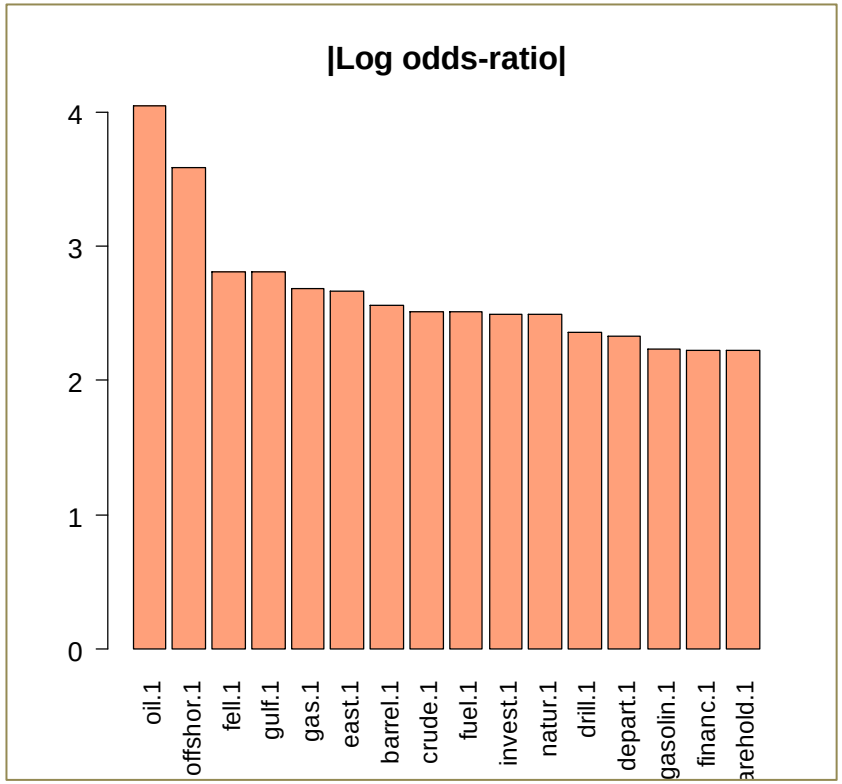
- Une seule équation à manipuler, simplicité de déploiement
- Le signe des coefficients c_j donne des indications sur l'influence des termes
- L'exponentielle des valeurs des coefficients c_j se lit comme un odds-ratio
- On peut s'en servir pour le « ranking » des variables (sélection)
- On a des résultats analogues à la régression logistique ou SVM linéaire mais les temps de calcul n'ont absolument rien à voir !

Reuters

$$d(\mathcal{X}) = d(acq,\mathcal{X}) - d(crude,\mathcal{X})$$

$$d(\mathcal{X}) = 14.064 - 4.042 \times oil - 2.687 \times gas - 3.584 \times offshor + \cdots$$

En triant les variables selon la valeur absolue des coefficients (décroissante), nous distinguons celles qui pèsent le plus dans la discrimination. On pourrait aussi imaginer une valeur de coupure pour éliminer les variables dont les coefficients sont négligeables (non significativement différents de zéro).



Remarque

Les résultats sont proches (ordre des variables les plus influentes) de ceux obtenus avec l'algorithme de filtrage basé sur l'incertitude symétrique. C'est normal, nous utilisons les mêmes tableaux de contingences. Mais, dans le filtrage nous exploitons ρ, ici nous exploitons le log-odds ratio c_j et l'ordonnancement est en relation directe avec les propriétés du classifieur utilisé.

Dans Weiss et al. (2005, page 69) est décrite une variante du Naive Bayes plutôt conseillée lorsqu'on utilise une pondération TF (ou même TF-IDF), avec

$$d(c, \mathcal{X}) = a_0 + a_j X_j$$

… mais où : $a_0 = \ln \frac{N_c}{N}$

$$a_j = \ln \frac{\lambda + N_c^j}{\lambda \times p + \sum_{j'=1}^{p} N_c^{j'}}$$

λ est un paramètre de lissage que l'on fixe souvent à λ = 1

N_c^j est le nombre d'apparition du terme j dans les documents de la classe c

Nombre total de termes apparaissant dans les documents de la classe c. L'intérêt ici est la normalisation différente, où la longueur des documents est pris en compte. !

On ne s'intéresse qu'aux occurrences des termes N_c^j

Reuters

$$\mathrm{d}(acq, \mathcal{X}) = -0.499 - 5.501 \times oil - 5.837 \times gas - 8.545 \times offshor + \cdots$$

$$\mathrm{d}(crude, \mathcal{X}) = -0.934 - 4.419 \times oil - 4.610 \times gas - 5.475 \times offshor + \cdots$$

Ouvrages

Weiss S., Indurkhya N., Zhang T., Damerau F., « Text Mining – Predictive methods for analyzing unstructured information », Springer, 2005.

Aggarwal C., Zhai C., « Mining Text Data », Springer, 2012.

Coelho L.P., Richert W., « Building Machine Learning Systems With Python », 2nd Edition, Packt Publising, 2015.

Supports et tutoriels

R.R., « Classifieur bayésien naïf », mars 2011.

R.R. « Filtrage des prédicteurs », juin 2010.

R.R., « Apprentissage-test avec Sipina », mars 2008 ; F-Mesure, introduction de la matrice de coûts en apprentissage.