4 - Classification de texte

1. Principe
2. Arbres de décision
3. Classification bayésienne naïve

Références : *livre en ligne sur NLTK* *http://nltk.org/book/ch06.html*

4.1 - Principe

- La classification consiste à ranger un objet donné dans une classe, l'ensemble des classes étant donné à l'avance.

- Pour la classification textuelle, les objets à classer sont des objets langagiers : textes, phrases, mots.

- On se limitera aux méthodes statistiques fondées sur l'apprentissage qui peuvent s'appliquer quel que soit le domaine.

- La première étape consiste à sélectionner les **traits** des objets à classer qui sont pertinents pour le classement et à définir la façon de les **coder** (booléens, entiers, réels, chaînes de caractères)

4.1 - Principe

- Un classifieur est **supervisé** s'il est construit à partir d'un ensemble d'apprentissage où les objets concernés sont correctement classés et où les traits sont renseignés.

- Le **choix des traits** est crucial et dans une approche supervisée, on va s'aider de l'ensemble d'apprentissage pour pondérer les traits qui peuvent être nombreux, en fonction de leur pouvoir disciminant par rapport aux classes choisies.

- Si on choisit trop de traits au départ avec un trop petit corpus d'apprentissage, on aura du mal à généraliser à partir du corpus (surajustement ou **overfitting**).

- Un moyen de raffiner l'ensemble des traits est d'utiliser un **corpus de développement** à côté du corpus d'apprentissage, qui va permettre de détecter les erreurs et de corriger les traits en fonction.

4.1 - Principe

- Pour évaluer un classifieur, on a besoin d'un corpus de test qui est un corpus où les objets sont correctement rangés dans des classes et où les traits sont renseignés.

- En général, on choisit un seul corpus qu'on divise en 90% pour l'entrainement et 10% pour le test. En répétant ceci et modifiant la subdivision, on peut faire de la validation croisée.

- L'évaluation se fait par comparaison de la sortie du classifieur sur le corpus de test avec la référence. La mesure de fidélité est le rapport du nombre d'objets correctement classés sur le nombre d'objets à classer.

- Pour une classification binaire, on distingue les vrais positifs (TP) les vrais négatifs (TN), les faux positifs (FP) et les faux négatifs (FN) d'où deux mesures : le rappel, TP/(TP +FN), et la précision, TP/(TP+FP).

4.2 - Arbres de décision

- Un arbre de décision est composé de nœuds internes ou **nœuds de décision**, où des valeurs de traits sont testées et des **feuilles** où est fait l'affectation des classes.

- Pour un objet à classer dont on connaît les valeurs de trait, on parcourt l'arbre de la racine vers les feuilles et à chaque nœud de décision, le choix du fils se fait en fonction des valeurs de traits associées à l'objet et des indications du nœud.

- Pour la construction de l'arbre de décision, le choix des traits est crucial et il est guidé par l'ensemble d'apprentissage.

4.2 - Arbres de décision

- L'arbre est construit récursivement à partir de l'ensemble d'apprentissage à partir de la racine. Si on considère qu'à chaque nœud de décision est attaché un seul trait, on cherche sur la partie du corpus d'apprentissage concernée, le trait le plus discriminant.

- Une méthode consiste à chercher le trait qui augmente au maximum l'organisation de l'ensemble des objets, c'est-à-dire en diminue l'entropie.

- L'entropie du classement d'un ensemble selon les classes c_1, ..., c_n qui ont des probabilités respectives $P(c_1)$, ..., $P(c_n)$ est donné par la formule :

$$H = -\Sigma_{i=1}^{i=n} P(c_i) log_2(P(c_i))$$

4.2 - Arbres de décision

- Dans la construction de l'arbre de décision de la racine vers les feuilles, chaque nœud de décision correspond à un sous-ensemble E de l'ensemble d'apprentissage. Si on choisit le trait f pour partionner E selon ses valeurs x_1, ..., x_p , on obtient les sous-ensembles E_1, ..., E_p et on calcule le gain d'information G(f) obtenu en faisant cette partition à l'aide de la formule :

$$G(f) = H - \Sigma_{i=1}^{i=n} \frac{|E_i|}{|E|} H_i(f)$$

$H_i(f)$ représente l'entropie du classement sur le sous-ensemble E_i où le trait f prend la valeur x_i et H représente l'entropie du classement sur E.

- On choisit comme trait discriminant celui qui entraîne le gain d'information le plus grand.

4.2 - Arbres de décision

Exemple : On cherche à savoir de quoi dépend la réussite aux examens d'un étudiant à partir de 3 paramètres : son assiduité et sa participation en classe exprimées sous forme booléenne, et le plat qu'il a mangé à Noël. On utlise pour cela un arbre de décision créé par apprentissage.

Les données d'apprentissage sont fournies par le tableau suivant :

Assiduité	Participation	Repas de Noël	Réussite
oui	oui	dinde	oui
oui	non	canard	non
non	oui	dinde	non
oui	oui	canard	oui
oui	non	dinde	non
oui	non	canard	oui

4.2 - Arbres de décision

On va déterminer le gain d'information obtenu en choisissant l'assiduité comme trait initial discriminant.

On calcule l'entropie relative au succès aux examens sur tout l'ensemble d'apprentissage.

$$H = -(P(succes = 1) * log_2(P(succes = 1) + P(succes = 0) * log_2(P(succes = 0))$$

$$H = -(3/6 * log_2(P(3/6) + 3/6 * log_2(3/6) = 1$$

On calcule ensuite l'entropie sur chacun des sous-ensembles déterminés par le trait "assiduité".

$$H(ass = 1) = -(3/5 * log_2(P(3/5) + 2/5 * log_2(2/5) \approx 0,97$$

$$H(ass = 0) = -(0/1 * log_2(P(0/1) + 1/1 * log_2(1/1) = 0$$

4.2 - Arbres de décision

On calcule ensuite le gain d'information comme différence entre l'entropie initiale et la moyenne pondérée des entropies sur les deux sous-ensembles déterminés par le train "assiduité

$$G(ass) = H - (5/6 * H(P(ass = 1) + 1/6 * H(ass = 0))$$

$$G(ass) = 1 - 5/6 * 0,97 \approx 0,19$$

On fait le même calcul pour les deux autres traits et on choisira pour la racine de l'arbre de décision le trait qui fait gagner le maximum d'information.

On itère ensuite le processus à partir de chacun des nœuds fils engendrés.

4.2 - Arbres de décision

- Le modèle des arbres de décision est simple à comprendre et à adapter aux systèmes hiérachiques de traits.

- Un inconvénient vient du fait que chaque nœud de décision entre un découpage de l'ensemble d'entrainement en ensembles plus petits, et ceux-ci peuvent être trop petits entrainant du surajustement.

- Un remède à cela est d'élaguer l'arbre de décision en s'aidant de l'ensemble de développement.

- Le modèle des arbres de décision n'est pas adapté lorsque les traits sont relativement indépendants et lorsque l'on veut tenir compte de certains à faible pouvoir discriminant.

4.2 - Arbres de décision

1. L'objectif est de prédire dans un ensemble d'articles de journaux quels sont les articles qui sont des articles économiques. L'idée est d'utliser un arbre de décision se fondant sur les mots clés "économie", "finance", "croissance" et "crise". Les données d'apprentissage sont celles du tableau ci-dessous, l'ensemble d'apprentissage étant formé de 3000 articles.

économie	finance	croissance	crise	nb d'articles économiques	nb d'autres articles
oui	oui	oui	oui	185	60
oui	oui	oui	non	220	42
oui	oui	non	oui	190	58
oui	non	oui	oui	255	73
non	oui	oui	oui	80	59
oui	oui	non	non	178	52
oui	non	oui	non	40	32
non	oui	oui	non	86	101
oui	non	non	oui	230	54
non	oui	non	oui	74	87
non	non	oui	oui	12	86
oui	non	non	non	45	34
non	oui	non	non	2	99
non	non	oui	non	5	206
non	non	non	oui	8	115
non	non	non	non	0	232

4.2 - Arbres de décision

1. En calculant pour chacun des mots clés le gain d'information obtenu en le choisissant comme trait discriminant, déterminer le mot clé le meilleur pour la racine de l'arbre de décision.
2. L'objectif est de prédire la tenue ou non d'épreuves de golf en fonction de la météo à l'aide d'un arbre de décision. On considère que la prédiction dépend de 4 traits, l'ensoleillement, la température, l'humidité et le vent.
   Les données d'apprentissage sont celles du tableau ci-dessous.

Numéro	Ensoleillement	Température	Humidité	Vent	Jouer
1	soleil	75	70	oui	oui
2	soleil	80	90	oui	non
3	soleil	85	85	non	non
4	soleil	72	95	non	non
5	soleil	69	70	non	oui
6	couvert	72	90	oui	oui
7	couvert	83	78	non	oui
8	couvert	64	65	oui	oui
9	couvert	81	75	non	oui
10	pluie	71	80	oui	non
11	pluie	65	70	oui	non
12	pluie	75	80	non	oui
13	pluie	68	80	non	oui
14	pluie	70	96	non	oui

4.2 - Arbres de décision

Calculer le gain d'information obtenu en choisissant chaque trait à tour de rôle comme trait discriminant.

A partir du calcul précédent, quel leçon peut-on tirer du modèle des arbres de décision et comment améliorer l'arbre de décision utilisé pour cet exemple ?

4.3 - Classifieurs bayésiens naïfs

- On suppose que les classes dépendent de n traits $t_1,..., t_n$, qui sont **indépendants** les uns des autres et on utilise un modèle probabiliste fondé sur un calcul de probabilités conditionnelles.

- Etant donné un objet dont on connait les valeurs $v_1,..., v_n$ des n traits $t_1,..., t_n$, il s'agit de trouver la classe c qui maximise la probabilité conditionnelle $P(c \mid v_1,..., v_n)$.

- Selon le théorème de Bayes :

$$P(c|v_{1n}) = \frac{P(v_{1n}|c) \times P(c)}{P(v_{1n})}$$

- Compte tenu de l'indépendance des traits, il faut maximiser :

$$P(v_{1n}|c) \times P(c) = \Pi_{i=1}^{i=n} P(v_i|c) \times P(c)$$

4.3 - Classifieurs bayésiens naïfs

Exemple : On cherche à prévoir la réussite aux examens d'un étudiant à partir de 3 paramètres : son assiduité et sa participation en classe exprimées sous forme booléenne, et le plat qu'il a mangé à Noël. On utilise pour cela un classifieur bayésien naïf créé par apprentissage.

Les données d'apprentissage sont fournies par le tableau suivant :

Assiduité	Participation	Repas de Noël	Réussite
oui	oui	dinde	oui
oui	non	canard	non
non	oui	dinde	non
oui	oui	canard	oui
oui	non	dinde	non
oui	non	canard	oui

4.3 - Classifieurs bayésiens naïfs

On cherche à prévoir si un étudiant qui est assidu mais ne participe pas en cours et qui a mangé du canard à Noël réussira ses examens.

On commence par calculer la probabilité a priori pour chaque classe.

$$P(succes = 1) = P(succes = 0) = 3/6 = 0.5$$

On calcule ensuite la vraisemblance des valeurs de traits pour chaque classe. Pour la classe "succès", on obtient :

$$P(ass = 1|suc = 1) \times P(part = 0|suc = 1) \times P(rep = canard|suc = 1)$$

$$3/3 \times 1/3 \times 2/3 \approx 0,22$$

Pour la classe "échec", on obtient :

$$P(ass = 1|suc = 0) \times P(part = 0|suc = 0) \times P(rep = canard|suc = 0)$$

$$2/3 \times 2/3 \times 1/3 \approx 0,15$$

4.3 - Classifieurs bayésiens naïfs

On calcule ensuite la probabilité a posteri de chaque classe.

Pour la classe "succès" : $0,22 \times 0,5 = 0,11$

Pour la classe "échec" : $0,15 \times 0,5 \approx 0,07$

En conclusion, on prévoit que l'étudiant réussira ses examens.

4.3 - Classifieurs bayésiens naïfs

- Quand un trait donné n'apparaît jamais pour une classe donnée dans le corpus d'apprentissage, on lui attribue une probabilité non nulle calculée selon une méthode de **lissage**.

- Le défaut des classifieurs bayésiens naïfs est que l'hypothèse d'indépendance des traits est en générale fausse, ce qui donne aux traits dépendants un poids plus grand qu'ils ne devraient avoir. Les **modèles à entropie maximum** prennent en compte les interactions entre traits.

4.3 - Classifieurs bayésiens naïfs

1. L'objectif est de prédire la tenue ou non d'épreuves de golf en fonction de la météo à l'aide d'un classifieur bayesiens naïf. On considère que la prédiction dépend de 4 traits, l'ensoleillement, la température, l'humidité et le vent.
   Les données d'apprentissage sont celles du tableau ci-dessous

Numéro	Ensoleillement	Température	Humidité	Vent	Jouer
1	soleil	moyenne	faible	oui	oui
2	soleil	élevée	importante	oui	non
3	soleil	élevée	moyenne	non	non
4	soleil	moyenne	importante	non	non
5	soleil	basse	faible	non	oui
6	couvert	moyenne	importante	oui	oui
7	couvert	élevée	moyenne	non	oui
8	couvert	basse	faible	oui	oui
9	couvert	élevée	faible	non	oui
10	pluie	moyenne	moyenne	oui	non
11	pluie	basse	faible	oui	non
12	pluie	moyenne	moyenne	non	oui
13	pluie	basse	moyenne	non	oui
14	pluie	basse	importante	non	oui

4.3 - Classifieurs bayésiens naïfs

1. A l'aide d'un classifieur bayésiens naïf entrainé sur les données précédentes, prédire la tenue ou non d'épreuves sportives compte tenu des paramètres ci-dessous :

Numéro	Ensoleillement	Température	Humidité	Vent	Jouer
15	soleil	élevée	faible	non	?
16	pluie	élevée	importante	non	?

2. Reprendre l'exercice 2 du paragraphe 4.2 avec les mêmes données d'apprentissage mais avec un classifieur bayésien naïf pour déterminer si les articles contenant le mot "économie" mais pas les autres mots clés sont plutôt des articles économiques ou non.