M1 PluriTAL

Introduction à la fouille de textes

Examen final

23 avril 2018

Durée de l'examen : deux heures

Tous les documents non-électroniques sont autorisés.

1 Classification par apprentissage automatique

On dispose d'un ensemble de textes représentés par deux attributs que l'on souhaite ranger dans deux classes distinctes notées + et −. Des exemples déjà classés sont fournis dans le tableau suivant :

Identifiant	Attribut 1	Attribut 2	Classe
1	1	4	+
2	2	1	−
3	3	5	+
4	4	4	+
5	4	1	−
6	6	2	−
7	7	5	−

1. Dessiner dans un repère cartésien à 2 dimensions les points représentant chacun de ces textes (en mettant des symboles + et − sur votre figure).
2. Expliquer ce que donnera sur ces exemples l'application de chacun des algorithmes d'apprentissage automatique de classifieurs suivants :
   (a) L'algorithme de la classe majoritaire
   (b) Un algorithme de construction d'un arbre de décision (construire un arbre qui classe correctement tous les exemples, plusieurs solutions sont possibles, deux niveaux dans l'arbre suffisent)
   (c) Un algorithme de SVM (dessiner un séparateur crédible sur la figure)
3. Classer à l'aide des trois classifieurs construits précédemment les nouvelles donné ! :es suivantes

Identifiant	Attribut 1	Attribut 2
8	2	3
9	2	2
10	5	1
11	6	5
12	7	1

4. Sachant que parmi ces nouvelles données, 8, 9 et 10 appartiennent en fait à la classe + et 11 et 12 à la classe −
   (a) Donner les matrices de confusion obtenues par les classifieurs construits précédemment pour ces nouvelles données
   (b) En déduire la précision, le rappel et la F-mesure de ces algorithmes pour la classe + sur ces données

2 Identification de tâches

La tâche d'*annotation sémantique* est une extension de la tâche de détection des entité nommées qui consiste à lier des entités d'une base de connaissances avec leurs mentions dans un texte. Par exemple, un système d'annotation sémantique doit être capable de lier la phrase

> Madame Curie est la seule personne à avoir été récompensée d'un prix Nobel dans deux domaines scientifiques distincts.

avec les entrées « Marie Skłodowska-Curie » et « prix Nobel » de sa base de connaissances mais pas avec les entrées « Pierre Curie » ou « Irène Joliot-Curie ».

1. Reformuler la tâche d'annotation sémantique comme un enchaînement de tâches élémentaires de la fouille de textes vues en cours (plusieurs solutions sont possibles, deux tâches distinctes suffisent)
2. Suggérer des ressources pertinentes pour un système d'annotation sémantique

3 Extraction d'information

On désire tester différentes méthodes pour apprendre à extraire automatiquement des noms complets de réalisateurs de cinéma, en distinguant leur nom et leur prénom, afin de les ranger dans deux champs distincts d'une base de données. Pour cela, on dispose des exemples suivants, dans lesquels les mots soulignés sont les *prénoms* et les mots **en gras** sont les *noms de famille* à extraire

1. Le film de Lars **von Trier**, « Melancolia », a été bien accueilli au Festival de Cannes
2. Tout le monde attend le prochain film de **Wong** Kar Wai.
3. Francis Ford **Coppola** est un des rares réalisateurs à avoir obtenu deux palmes d'or
4. L'intégralité des films de Jacques **Demy** sont sortis en DVD.

Pour extraire ces noms et prénoms, on propose d'appliquer une méthode de classification aux *séparateurs entre mots*, c'est-à-dire les espaces et les ponctuations (qui sont supprimées du texte initial) séparant deux mots ainsi que le début et la fin d'un texte. Par exemple, dans la phrase 1, en représentant les séparateurs entre mots par des barres verticales, on obtient :

| Le | film | de | Lars | **von** | **Trier** | Melancolia | a | été | bien | accueilli | au | Festival | de | Cannes |

1. Proposer un ensemble de classes dans lesquelles ranger les séparateurs, qui permettront d'extraire correctement les informations souhaitées.
2. Si on veut utiliser un programme d'apprentissage automatique pour distinguer ces séparateurs, il faut les transformer en vecteurs. Pour chaque séparateur, on suppose que les attributs choisis sont : la catégorie du mot deux places avant, la catégorie du mot précédent, la catégorie du mot suivant et la catégorie du mot deux places après. Les catégories possibles des mots sont : `DET`, `NC`, `NP` (pour « nom propre », identifiés uniquement parce ce qu'ils sont inconnus d'un dictionnaire et débutent par une majuscule), `V`, `PREP`, `ADJ`… Donner, pour chacune de vos classes de séparateurs, un exemple d'une donnée dans cette représentation.
3. D'après la régularité des données, quelle classe sera la plus facile à apprendre ?
4. Une autre solution pour extraire ces noms et prénoms est d'*annoter* les textes, en associant à chacun des mots une étiquette. Définir un jeu d'étiquettes permettant de résoudre la tâche qui nous préoccupe et annoter chacune des phrases fournies en exemples avec ce jeu d'étiquettes.

4 Questions diverses

Les questions suivantes sont indépendantes

1. Si on ne se préoccupe pas de la précision d'un système de recherche d'information, comment peut-on s'assurer simplement que son rappel soit maximal ?

2. Quand un classifieur utilise l'algorithme de la classe majoritaire, quel est son rappel pour les classes non-majoritaires ? Peut-on donner sa précision pour ces classes ? Justifier.

3. On considère un corpus de deux textes, t_1 et t_2

   (a) Dans quelle situation la forme *mange* a-t-elle un indice TF·IDF non-nul pour t_1 et nul pour t_2 ?

   (b) On suppose que l'on se trouve dans la situation de la question précédente. Est-il possible qu'en lemmatisant le corpus, le lemme *manger* ait un indice TF·IDF nul pour t_1 **et** pour t_2 ? Justifier.