RÉSUMÉ MACHINE LEARNING

SELMI OUSSAMA

CHAPITRE 1

Introduction au machine learning

NOTIONS DE BASE

Le machine learning: un programme informatique qui apprend a partir d'une expérience avec des résultats ameliorés.

Notions génerals:

- Regression, estimation: prédire des valeurs continues (numériques)
- Classification: Prédire un classe ou une catégorie (valeur discrète)
- Clustering: Regroupement de individus avec cas similaire (homogènes)
- Séquence mining: prédire les événements suivants
- Réduction des dimensions (feature sélection): réduire la taille de dataset

Les données d'apprentissage sont divisés en 3 types:

1. Données d'entrainement: construire un ensemble des exemples
2. Données de test: candidats sur qui on applique le modèle d'apprentissage
3. Données de validation: réajuster ou valider es paramètres de l'algorithme

NOTIONS DE BASE

Autres notions:

- Sur apprentissage (overfitting): absence de géneralisation, le modèle s'adapte qu'au training set.
- Sous apprentissage: un modèle qui s'adapte mal au training set, avec un cout d'erreur large.

Types d'apprentissage:

- Supervisé: construire un modèle a partir des exemples avec labels connues.

+ Classification, Régression

- Non supervisé: algorithme doit découvrir lui meme la structure de données (découvrir des clusters)

❑TDSP (Team Data Science Process): méthode itérative pour les solutions d'analyse prédictive. (étapes: Compréhension métier, acquisition des abc, modélisation, déploiement)

❑Méthodologie CRISP-DM (Cross Industry Standard Process - Data Mining): agile et itérative pour les projets Data Science

CHAPITRE 2

Préparation de données

Il faut ici avoir: Un nettoyage de données, ou transformation.

savoir s'il ya des exigences sur les données, corriger les erreurs (bruit statistique), Sélectionner des variables, ingénierie des variables (dérivation des nouvelles variables), réduire les dimensionnalités

1) Nettoyage des données (imputation) y inclut l'élimination des individus avec n.a, remplacer les données avec une valeur fixe, arbre de décision ou les valeurs les plus proches (similarité)

2) Feature selection est le processus de choisir les caractéristiques de données pour une meilleure prédiction

+ réduire le surajustement (valeurs rédondantes ou non nécessaires), améliorer la précision, réduire le temps du training.

On a 3 méthodes: filter, wrapper, embedded

3) transformation de données (normalisation, agrégation, généralisation) elle nous aide a rapprocher les échelles des attributs.

3.1 Agrégation: collecter de données a partir diverses sources avec un format unique.

3.2 discrétisations: conversion de données continues en ensemble d'intervalles.

3.3 géneralisation (OLAP, AOI): conversion de données afin de les géneraliser (exemple: age 20 -> jeunes)

4) ingénierie des variables: y inclut la création des features (somme, min, produit..), topic extraction, extraction des features a partir des images (variance, écart, skewness..), bag of words (occurence du mot dans un texte)

CHAPITRE 3

Apprentissage supervisé

KNN: basé sur la classification qui prédire l'integrité d'un élément a un groupe ou catégorie.

Fonctionnement du KNN:

1- Prendre des points labels et les utiliser pour prédire les labels d'autres points

2- classer les cas similaires

3- détecter les voisins

Fonctionnement algorithmique (pseudo code)

1- Choisir valeur de K (K< sqrt(n nombre échantillons)

2- Calculer les distances entres les cas (distance euclidienne pour valeurs qauntitatives et distance de manhattan pour les variables diverses)

3- Chercher les K observations proches

4- Voter la classe du cas

+ Simple, facile, efficace pour données non linéaire, polyvalent

- cout de calcul, choix K crucial, sensible au bruit

SVM: cet algorithme mappe les données afin de le catégoriser meme s'il sont pas séparables linéairement. avec l'objectif de trouver une hyperplane qui sépare les classes de donées de manière optimale.

Il a des fonctions noyaux (Linéaire, polynomial, RBF, Sigmoide)

Si les données sont sur un seul axe x, on peut avoir un espace basé sur x et x carré.

Le **kernelling**: c'est le mapping de données dans un éspace dimensionnel.

SVM multi classes: ici on prend la méthode one vs all, on prend chacun des classes et on le compare individuellement avec une autre classe.

+ efficace a grandes dimensions, robuste aux données bruitées

- choix du noyau difficile, sensible a l'échelle, pas idéal pour les multi classes.

Arbre de décision: un modèle prédictif utilisant une structure arborescente pour prendre des décisions basés sur des règles (hiérarchie descendante)

On a un nœud racine qui est le point de départ de l'arbre (on le choisit pur et celui qui sépare mieux les exemples positives et négatives, on se base a des mésures d'impureté (entropie de shannon, entropie de boltzmann, index de gini)

- un noeud interne qui répresente un test sur un attribut.
- une branche répresente un choix qui résulte du test.
- une feuille c'est un noeud terminal qui donne la décision finale.

Fonctionnement:

1) Initialiser a partir du racine
2) pour chaque noeud on choisit la variable qui sépare mieux les individus
3) séparation des individus
4) fin

Conditions arret: fin des attributs, tous les enregistrements d'un noeud sont dans la meme classe.

L'algorithme CART (Classification and Regression Trees) est une technique d'apprentissage supervisé utilisée pour la construction d'arbres de décision.

il utilisie l'indice de gini qu'on cherche a minimiser, IG= 1-ensemble fi*2

\+ simple a interpréter, fonctionne avec tous les types de données

\- possibilité du surapprentissage, peut etre complexe

Température	Oui	Non	Nbr D'instance
Chaud	2	2	4
Moyen	3	1	4
Frais	4	2	6

Gini(Temp=Chaud) = 1 - $(2/4)^2$- $(2/4)^2$ = 0.5

Gini(Temp=Moyen) = 1 - $(3/4)^2$- $(1/4)^2$ = 0.375

Gini(Temp=Frais) = 1 - $(4/6)^2$ - $(2/6)^2$ = 0.445

Ensuite, nous allons calculer la somme pondérée des indices de Gini pour la caractéristique « **Température** ».

Gini(Temp) = (4/14) × 0.5 + (4/14) x 0.375 + (6/14) × 0.445 = 0.439

CHAPITRE 3

Métriques évaluation

c'est primordial pour savoir si le modèle est significatif et pour donner une idée des performances en déploiement ou pour comparer des modèles candidats.

*** Cross validation**: c'est une technique qui consiste a diviser les données en sous ensembles pour entrainer et tester le modèle plusieurs fois.

- exemple: train test split: partitionner les données en deux ensembles, apprentissage 80% et test 20%

autres exemples: shuffle split cross validation, time series cross validation..

*** Matrice de confusion**: contenant plusieurs mesures de performances

elle contient 4 métriques: fausses prédiction FN, FP, et bonnes prédictions VP, VN.

1) accuarcy (pourcentage de bonnes prédictions VP+VN/Total
2) recall (proportion des résultats positifs rééles identifés correctement VP/VP+FN)
3) precision (proportion d'identifications positives correcte VP/VP+FP)
4) F1 score (combine a précision et le recall 2*(recall*precision)/recall+precision)

*** Graphiquement**

- * AUC (Area Under the Curve) ROC (Receiver Operating Characteristic)
- *ROC, représente les taux de VP en fonction de taux FP
- *AUC, l'aire sous la courbe ROC
- * Courbe PR (precision recall) met en oeuvre la relation netre précision et recall + aproprié pour les classes déséquilibrées au contraire du ROC + détecte les modèles optimaux.