RER- Naïves Bayes

Contexte : Comprendre l'utilisation des algorithmes (Naïves Bayes , OneR) dans le machine learning et notamment dans la classification supervisée.

Mots clés :

- Naïves Bayes : classification bayésienne probabiliste simple basée sur le théorème de Bayes avec une forte indépendance des entrées, dite naïve, des hypothèses.

Calculer la probabilité a posteriori à partir des probabilités prior et de la probabilité conditionnelle.

La probabilité posteriori de l'hypothèse

H = hypothèse = classe = décision = évènement = label = target. L'hypothèse c'est ce que l'on cherche (posteriori) à partir de l'historique (prior) et de la probabilité conditionnelle

Features = ça peut être entrée/sortie.

- Naïf : croyance en l'indépendance des features

- Théorème de Bayes : $$P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$$

- Indépendance : (définition courte) Evènements aléatoires n'ayant aucune influence l'un sur l'autre.

(Définition longue et détaillée) L'indépendance de 2 évènements est une notion statistique. Elle implique que si 2 événements A et B surviennent aléatoirement, la survenue de l'évènement A n'influera pas sur la survenue de l'évènement B et vice-versa. Cela s'exprime mathématiquement par l'équation suivante : $P(A \cap B) = P(A) * P(B)$.

- Odd of Believes : Degré de croyance qu'une hypothèse est vrai. Degré de croyance que cet évènement A est vrai sachant que B est vrai.

(Likelihood = Vraisemblance)

LN : Likelihood of Necessity : Mesure qu'une hypothèse n'est pas vrai sachant que l'évidence n'est pas présente.

$$LN = \frac{P(\neg E \mid H)}{P(\neg E \mid \neg H)}$$

(LN est la vraisemblance de nécessité et c'est une mesure du discrédit de l'hypothèse H si la preuve E est manquante)

LS : Likelihood of Sufficiency: Croyance qu'une hypothèse est vrai sachant que l'évidence/Symptôme est vrai.

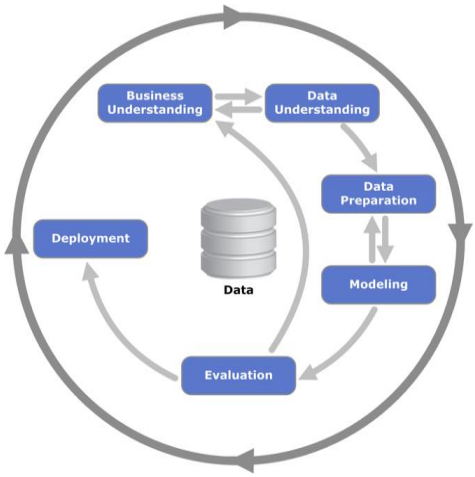
$$LS = \frac{P(E \backslash H)}{P(E \backslash \neg H)}$$

(LS est la vraisemblance de suffisance : c'est la mesure de la croyance de l'expert en l'hypothèse H si la preuve E est présente)

- Scikit-learn : Bibliothèques "Python" pour le machine learning.
- Crisp-DM : Cross industry standard process for Data-Mining.

Modele de processus itératif comportant 6 phases séquentielles :

1. Compréhension de l'entreprise
2. Compréhension des données
3. Préparation des données (Data Munging (= Synonyme =) Wrangling = Transformation, Formatage, nettoyage des données) => les rendre exploitable
4. Modélisation
5. Evaluation
6. Deployment



- Apprentissage : Entraînement d'un modèle.

- OneR : One Rule : Algorithme de classification qui prend la règle de l'attribut (champs) le plus dominant "(/discriminant)" pour réaliser les prédictions. C'est de l'apprentissage automatique supervisé simplifié.

Problématique(s) :

Comment fonctionne un classifieur Naïves Bayes ?

Comment l'utilisation d'un algorithme "naïf" à partir du Théorème de Bayes permet de classifier les données ?

Quelles sont les limites du Naïves Bayes ?

Hypothèses :

1. Le Modèle Naïve Bayes permet d'entrainer la machine plus rapidement que d'autres modèles (Adeline) VRAI
2. CRISP-DM permet de faire du Clustering (Aude) FAUX
3. L'algorithme Naîve Bayes permet la classification des données (bien trop simple vu que Bassam a répondu) VRAI !!!!
4. L'algorithme de Naîve Bayes est l'algo de classification le plus utilisé (Axel) VRAI principalement pour classification de document FAUX pour le reste
5. L'algorithme Naîve Bayes peut servir pour classifier n'importe quel type de donnée (Briand) VRAI uniquement pour tout ce qui est type de données structurées
6. L'algorithme Naîve Bayes est un algo d'IA plus intelligent que les autres (Osman) FAUX
7. On ne peut pas utiliser le Naïve Bayes classifier pour réaliser des diagnostics médicaux (Etienne) FAUX
8. Naîve Bayes est un algorithme de Data Mining qui permet de ressortir des connaissances d'une base de données (Adrien) VRAI
9. L'algorithme Bayes demande moins de ressource en entrée mais moins précis en sortie (Niko) VRAI
10. Naîve Bayes ne convient pas forcément à toutes les classifications de données (Tetyana) VRAI
11. La sortie d'un classifier Bayesien peut être une égalité (Seydou) VRAI
12. Le caractère naïf du théorème de bayes à fait sa réputation (Jean-Paul SOSSAH) BOAF ! VRAI
13. Il existe un algorithme Bayesien non naïf (Adrien Gémini) I DON'T KNOW ! Mais Faux !
14. Le théorème de Bayes étant un algorithme de probabilité, nous l'utilisons plus dans du data mining que dans le machine learning. (Panda) FAUX !

15. On utilise les même métriques que règles d’association (Solenn sans le ‘e’ by Bassam) FAUX

Plan d'action :

Etudier les ressources

Comparaison OneR / Bayes : Les deux sont simples mais OneR plus simple.
OneR attribut discriminant ≠ Bayes prend tout

Découvrir Scikit-learn

Répondre aux problématiques

Réfléchir aux hypothèses

Faires les Workshop en collaboration

Livrer les Livrables