Techniques d’apprentissage automatique à développer

1. Définitions :
   1. Matrice de confusion

La matrice de confusion est la matrice qui répertorie les résultats de classification des individus dans le jeu de donnée de test afin de pouvoir quant à la pertinence de nos résultats.

C’est une matrice carrée de taille n avec n le nombre de classe dans le jeu de donnée.

Exemple :

|  |  |  |  |
| --- | --- | --- | --- |
|  | Valeurs réelle | | |
| Valeurs prédite |  | Classe 1 | Classe 2 |
| Classe 1 | Nombre d’individus classé correctement dans la classe 1  Noté TP (True Positive) pour la suite | Nombre d’individus mal classé venant de la classe 2  Noté FP (False Positive) pour la suite |
| Classe 2 | Nombre d’individus mal classé venant de la classe 1  Noté FN (False Négative) pour la suite | Nombre d’individus classé correctement dans la classe 2  Noté TN (True Négative) pour la suite |

* 1. Précision

La précision calcule le ratio de données bien prédites sur le total des prédictions d’une classe.

Cependant, si le modèle se base sur la maximisation de la précision, il ne fera pas attention aux nombres de faux positifs.

Pour contrer cela, il existe le calcul du rappel.

* 1. Recall (Rappel)

Le Rappel calcule le ratio de données bien prédites sur le total des données d’une classe.

Cependant, si le modèle se base sur la maximisation du rappel, il ne fera pas attention aux nombres de faux négatifs.

Pour contrer ces deux problématiques il existe le Fβ-Score.

* 1. F1-Score

Fβ-Score utilise les valeurs de précisions et de recall pour obtenir un nouveau score.

Donc pour le F1 Score on a :

Avec β = 1 : la précision et le rappel ont autant de poids dans le calcul du score.

* 1. K plus proches voisins (K-nearest neighbors)

1. Mentionnez la métrique de distance que vous avez choisie, et justifiez votre choix.

La métrique choisis est la distance euclidienne par élimination car c’est la plus adaptée en comparaison des autres méthodes de calcul de distance.

* Distance cosine : est plus adapté pour calculer la similarité entre deux textes.
* Distance de Jaccard : comme la distance cosine, elle est plus adaptée pour calculer la similarité entre deux textes.
* Distance de Hamming : est plus adaptée pour comparer des attributs binaires.
* Distance de Manhattan : est plus adaptée lorsque la dimensionnalité est grande.

1. Pseudo-code

Hyperparamètre :

* k : nombre de plus proche voisin
* metrics : méthode de calcul de distance

entrainer\_tester.py