0.1 Definition Communes

Dans cette section, nous définissons des concepts et des termes communs utilisés dans plusieurs algorithmes d'apprentissage automatique présentés dans ce document.

0.1.1 Distance Euclidienne

La distance euclidienne est une mesure de la distance entre deux points dans un espace euclidien. Elle est définie par la formule suivante pour deux points avec n dimensions $p = (p_1, p_2, \ldots, p_n)$ et $q = (q_1, q_2, \ldots, q_n)$:

$$d(p, q) = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}$$

0.1.2 Distance de Manhattan

La distance de Manhattan, également connue sous le nom de distance de taxicab ou distance L1, est similairement définie par la formule suivante pour deux points avec n dimensions $p = (p_1, p_2, \ldots, p_n)$ et $q = (q_1, q_2, \ldots, q_n)$:

$$d(p, q) = \sum_{i=1}^{n} |p_i - q_i|$$

0.1.3 Distance de Minkowski

La distance de Minkowski est une généralisation des distances euclidienne et de Manhattan. Elle est définie par la formule suivante pour deux points avec n dimensions $p = (p_1, p_2, \ldots, p_n)$ et $q = (q_1, q_2, \ldots, q_n)$, et un paramètre $r \geq 1$:

$$d(p, q) = \left(\sum_{i=1}^{n} |p_i - q_i|^r \right)^{\frac{1}{r}}$$

0.1.4 Standarisation

La standarisation, également appelée normalisation Z-score, est une technique de mise à l'échelle des données qui transforme les valeurs d'une variable pour qu'elles aient une moyenne nulle et un écart-type unitaire. La formule de la standarisation pour une valeur x est la suivante :

$$z = \frac{x - \mu}{\sigma}$$

où μ est la moyenne de la variable et σ est son écart-type.

0.1.5 Dummy Encoding

Le Dummy Encoding est une technique d'encodage des variables catégorielles qui transforme chaque catégorie en une variable binaire distincte (0 ou 1). Pour une variable catégorielle avec k catégories, le Dummy Encoding crée k nouvelles variables binaires, où chaque variable indique la présence ou l'absence d'une catégorie spécifique.

0.1.6 Ordinal Encoding

L'Ordinal Encoding est une technique d'encodage des variables catégorielles qui attribue un entier unique à chaque catégorie en fonction de son ordre ou de sa hiérarchie. Contrairement au Dummy Encoding, l'Ordinal Encoding préserve l'ordre des catégories, ce qui peut être utile lorsque les catégories ont une relation intrinsèque (par exemple, "petit", "moyen", "grand" peuvent être encodés comme 0, 1, 2 respectivement), ou quand on souhaite minimiser le nombre de nouvelles variables créées.

0.1.7 F1 Macro

La F1 Macro est une métrique d'évaluation utilisée pour mesurer la performance d'un modèle de classification, en particulier dans les cas de classes déséquilibrées. Elle est calculée en prenant la moyenne non pondérée des scores F1 pour chaque classe. Le score F1 pour une classe est la moyenne harmonique de la précision et du rappel, définie comme suit :

$$F1 = 2 \times \frac{\text{Précision} \times \text{Rappel}}{\text{Précision} + \text{Rappel}}$$

et la F1 Macro est donnée par :

$$\text{F1 Macro} = \frac{1}{K} \sum_{k=1}^{K} F1_k$$

0.1.8 ROC AUC

La ROC AUC (Receiver Operating Characteristic - Area Under the Curve) est une métrique d'évaluation utilisée pour mesurer la performance d'un modèle de classification binaire. La courbe ROC trace le taux de vrais positifs (TPR) contre le taux de faux positifs (FPR) à différents seuils de classification. L'aire sous la courbe (AUC) quantifie la capacité du modèle à distinguer entre les classes positives et négatives. Une AUC de 1 indique une classification parfaite, tandis qu'une AUC de 0,5 indique une performance équivalente à un classement aléatoire.

0.1.9 L'entropie

L'entropie est une mesure de l'incertitude ou de la pureté d'un ensemble de données dans le contexte des arbres de décision. Elle est définie par la formule suivante pour un ensemble de données avec c classes :

$$H(D) = -\sum_{i=1}^{c} p_i \log_2(p_i)$$

ou p_i est la proportion d'instances appartenant à la classe i dans l'ensemble de données D. Une entropie élevée indique une plus grande incertitude, tandis qu'une entropie faible indique une plus grande pureté.

0.1.10 Indice de Gini

L'indice de Gini est une mesure de l'impureté ou de la pureté d'un ensemble de données, souvent utilisée dans les arbres de décision. Il est défini par la formule suivante pour un ensemble de données avec c classes :

$$G(D) = 1 - \sum_{i=1}^{c} p_i^2$$

ou p_i est la proportion d'instances appartenant à la classe i dans l'ensemble de données D. Un indice de Gini de 0 indique une pureté totale (toutes les instances appartiennent à une seule classe), tandis qu'un indice de Gini proche de 0,5 indique une impureté maximale (les instances sont réparties uniformément entre les classes).