Rapport TP3

1. Analyser les résultats obtenus et discuter les avantages et inconvénients de chaque approche.

a) Naïve Bayes :

Avantages :

- Rapide à l'entraînement et à la prédiction, ce qui est un énorme avantage dans les situations où les données sont volumineuses et donc la vitesse est importante.

- Fonctionne bien pour des problèmes où les caractéristiques sont indépendantes les unes des autres.

- Modèle simple, facile à comprendre et à implémenter.

- Efficace avec de petites quantités de données.

Inconvénients :

- Moins performant sur des données complexes ou des problèmes où les caractéristiques sont fortement corrélées.

- Hypothèse d'indépendance : dans de nombreux cas réels, les caractéristiques ne sont pas indépendantes, ce qui peut nuire à la performance.

- Sensible aux données déséquilibrées.

b) Complement Naïve Bayes :

Avantages :

- Spécialement conçu pour des données déséquilibrées, ce qui en fait un bon choix pour des jeux de données avec une forte asymétrie dans les classes.

- Plus robuste que Naïve Bayes classique lorsqu'il s'agit de classes déséquilibrées.

Inconvénients :

- Comme Naïve Bayes, il repose sur l'hypothèse d'indépendance des caractéristiques, ce qui peut également limiter ses performances dans des situations plus complexes.

c) Régression Logistique :

Avantages :

- Flexible et robuste : plus adapté aux données complexes avec des caractéristiques corrélées, car il ne fait pas l'hypothèse d'indépendance.

- Permet d'interpréter facilement les coefficients et donc les relations entre les variables.

- Fonctionne bien avec des données linéaires et a une capacité à gérer des jeux de données plus complexes par rapport à Naïve Bayes.

Inconvénients :

- Plus lent à l'entraînement par rapport à Naïve Bayes, surtout lorsque le nombre de variables ou de données est élevé.

- Sensible aux valeurs aberrantes et nécessite souvent un prétraitement des données.

2. Comparer la complexité des modèles et le temps d'entraînement.

a) Naïve Bayes :

- Temps d'entraînement : Très rapide, car il utilise une approche simple basée sur des probabilités conditionnelles.

- Complexité : Faible, car il est basé sur l'hypothèse d'indépendance et ne nécessite pas de calculs complexes.

b) Complement Naïve Bayes :

- Temps d'entraînement : Un peu plus lent que Naïve Bayes classique, mais toujours rapide.

- Complexité : Modéré, car il ajuste les probabilités pour mieux gérer les déséquilibres de classes.

c) Régression Logistique (Logistic Regression) :

- Temps d'entraînement : Plus lent, en particulier si le nombre de caractéristiques est élevé, car il utilise des algorithmes d'optimisation (comme la descente de gradient).

- Complexité : Élevée par rapport aux autres modèles, car elle doit ajuster les coefficients des caractéristiques pour minimiser une fonction de coût.

3. Réfléchir sur la pertinence de chaque approche pour des problèmes de classification spécifiques.

Naïve Bayes est particulièrement adapté pour des tâches de classification textuelle (comme le filtrage de spam ou la classification de documents), où les mots peuvent être traités comme indépendants les uns des autres. Ce modèle est également utile dans des environnements où la rapidité est importante.

Complement Naïve Bayes est recommandé pour des jeux de données déséquilibrés, comme dans le cas de spam (beaucoup de messages non spam et peu de spam). Il peut donner de meilleures performances par rapport à Naïve Bayes classique dans ces cas.

Régression Logistique est plus adaptée aux cas où les relations entre les variables sont plus complexes et où une flexibilité est nécessaire. Elle est idéale pour des tâches où les données sont linéaires ou où il est nécessaire de comprendre l'influence de chaque variable sur la sortie.

4. Expliquer pourquoi Naïve Bayes est rapide mais peut être limité, tandis que la régression logistique est plus flexible.

Naïve Bayes est rapide parce qu'il repose sur une simple estimation des probabilités conditionnelles. Cependant, il limite sa capacité à modéliser des relations complexes entre les caractéristiques, car il suppose que ces caractéristiques sont

indépendantes. Cela le rend peu adapté aux jeux de données où les relations entre les caractéristiques sont importantes.

Régression Logistique est plus flexible car elle permet de modéliser des relations non linéaires et prend en compte l'interaction entre les caractéristiques. Toutefois, cette flexibilité vient au prix d'une plus grande complexité et d'un temps d'entraînement plus long.

5. Quel modèle est le plus adapté si le dataset est déséquilibré ou bruyant ?

Complement Naïve Bayes est le plus adapté pour les données déséquilibrées. Contrairement à Naïve Bayes classique, il ajuste mieux les probabilités dans les jeux de données où une classe est beaucoup plus fréquente que l'autre. Cela en fait un bon choix pour des problèmes où la distribution des classes est déséquilibrée, comme la détection de spam.