

TP4 : Classification binaire avec SVM, Mélange de modèles, et Modèles Probabilistes Mixtes

Exercice 1 :

Exercice 2 :

Exercice 3 :

Exercice 4 :

Exercice 5 :

Question 1 :

- **Voting Classifieur (Soft)** : Combine les forces des autres modèles le scores est équilibrés entre précision, rappel et F1-score.
- **Logistic Regression** : Est très proche en termes d'AUC-ROC mais a un F1-score légèrement inférieur.
- **Naive Bayes** : A un excellent F1-score mais un AUC-ROC légèrement inférieur.
- **SVM** : Est performant mais légèrement moins robuste pour la classe spam 19 (FP).

Question 2 :

| | |
|--|----------------|
| Analyse temps d'entraînement : | |
| Naïve Bayes | 0.13 secondes |
| SVM | 22.84 secondes |
| Regression logistique | 0.15 secondes |
| Classifieur Hard | 33.99 secondes |
| Classifieur Soft | 16.75 secondes |
| Observations : | |
| <ul style="list-style-type: none">• Naive Bayes et Logistic Regression sont très rapides et légers, adaptés pour des tâches où les ressources sont limitées.• SVM est beaucoup plus exigeant en temps et en ressources.• Voting Classifiers combinent plusieurs modèles, ce qui explique le cout plus élevé. | |

Question 3 :

Préférences pour Voting Classifieur :

- **Données complexes** : Lorsque les données sont très variées, le Voting Classifieur peut équilibrer les forces de plusieurs modèles.
- **Robustesse** : Si les performances d'un seul modèle sont insuffisantes comme dans des tâches très légers et rapide, le Voting Classifieur peut réduire les erreurs.
- **Equilibre** : Pour un bon compromis entre précision et rappel, le Voting Classifiers (Soft) sont souvent plus robustes.

Question 4 :

| Modeles | Avantages | Limites |
|--------------------|--|--|
| SVM | Performant pour des données linéaires | Lent pour les gros datasets |
| GMM | Flexible avec des données non gaussienne | Plus adapté pour des taches de clustering que pour la classification |
| Voting classifieur | Combine plusieurs modeles benefice donc la force de chacun | Peut être très couteux à entrainer ! |

Question 5 :

Dans ce TP nous avons pu voir que des modèles légers sont plus adapter pour des prédiction rapide (naive Bayes) alors que des modèles comme le voting classifieur sont plus couteux mais on de meilleure performance sur de grande base de données. Avec une bonne optimisation de chaque modèle et un bon mélange de modèle dans le Classifieur voting on peut avoir de bon compromis entre rapidité et précision.

Annexes :

Matrix Logistic Regression :

Matrice de confusion :

| | | Réalité | |
|------------|--------------|--------------|--------------|
| | | Négative : 0 | Positive : 1 |
| Prédiction | Négative : 0 | 955 | 0 |
| | Positive : 1 | 22 | 128 |

Rapport de classification:

| | Metric precision | recall | f1-score | support |
|--------------|------------------|--------|----------|---------|
| 0 | 0.98 | 1 | 0.90 | 965 |
| 1 | 1 | 0.85 | 0.90 | 150 |
| | | | | |
| Accuracy | | | 0.98 | 1115 |
| Macro avg | 0.90 | 0.93 | 0.95 | 1115 |
| weighted avg | 0.98 | 0.98 | 0.98 | 1115 |

SVM

Matrice de confusion :

| | | Réalité | |
|------------|--------------|--------------|--------------|
| | | Négative : 0 | Positive : 1 |
| Prédiction | Négative : 0 | 960(VN) | 5(FN) |
| | Positive : 1 | 19(FP) | 131(VP) |

Rapport de classification

| | Metric precision | recall | f1-score | support |
|--------------|------------------|--------|----------|---------|
| 0 (ham) | 0.98 | 0.99 | 0.99 | 965 |
| 1 (spam) | 0.96 | 0.87 | 0.92 | 150 |
| | | | | |
| Accuracy | | | 0.98 | 1115 |
| Macro avg | 0.97 | 0.93 | 0.95 | 1115 |
| weighted avg | 0.98 | 0.98 | 0.98 | 1115 |

| | | | | | | | | | |
|------------------------------------|------------------|---------|--------------|--------------|------------------------------------|------------------|---------|-------------|--------------|
| Naives Bayes : | | | | | Classifieur voting Soft : | | | | |
| Matrice de confusion : | | | | | Matrice de confusion : | | | | |
| | | Réalité | | | | | Réalité | | |
| | | | Negative: 0 | Positive: 1 | | | | Negative: 0 | Positive : 1 |
| Prédiction | Negative: 0 | 957 | 8 | | Prédiction | Negative: 0 | 994 | 1 | |
| | Positive : 1 | 12 | 138 | | | Positive : 1 | 18 | 132 | |
| Rapport de classification : | | | | | Rapport de classification : | | | | |
| | Metric precision | recall | f1-score | support | | Metric precision | recall | f1-score | support |
| 0 | 0.99 | 0.99 | 0.99 | 965 | 0 | 0.98 | 1 | 0.99 | 965 |
| 1 | 0.95 | 0.92 | 0.93 | 150 | 1 | 0.99 | 0.88 | 0.93 | 150 |
| | | | | | | | | | |
| Accuracy | | | 0.98 | 1115 | Accuracy | | | 0.98 | 1115 |
| Macro avg | 0.97 | 0.96 | 0.96 | 1115 | Macro avg | 0.99 | 0.94 | 0.96 | 1115 |
| weighted avg | 0.98 | 0.98 | 0.98 | 1115 | weighted avg | 0.98 | 0.98 | 0.98 | 1115 |
| | | | | | | | | | |
| GMM : | | | | | Classifieur voting Hard : | | | | |
| Matrice de confusion : | | | | | Matrice de confusion : | | | | |
| | | Réalité | | | | | Réalité | | |
| | | | Negative : 0 | Positive : 1 | | | | Negative: 0 | Positive : 1 |
| Prédiction | Negative : 0 | 962 | 3 | | Prédiction | Negative : 0 | 964 | 1 | |
| | Positive : 1 | 36 | 114 | | | Positive : 1 | 20 | 130 | |
| Rapport de classification : | | | | | Rapport de classification : | | | | |
| | Metric precision | recall | f1-score | support | | Metric precision | recall | f1-score | support |
| 0 | 0.96 | 1 | 0.98 | 965 | 0 | 0.98 | 1 | 0.99 | 965 |
| 1 | 0.97 | 0.76 | 0.85 | 150 | 1 | 0.99 | 0.87 | 0.93 | 150 |
| | | | | | | | | | |
| Accuracy | | | 0.97 | 1115 | Accuracy | | | 0.98 | 1115 |
| Macro avg | 0.97 | 0.88 | 0.92 | 1115 | Macro avg | 0.99 | 0.93 | 0.96 | 1115 |
| weighted avg | 0.97 | 0.97 | 0.96 | 1115 | weighted avg | 0.98 | 0.98 | 0.98 | 1115 |

Figure 3 : Tableau des performances des différents modèles