TP2 Naive Bayes

Thibault Cart & Rami Albadri

Apprentissage et Test

Pour évaluer les performances de l'algorithme de Naive Bayes, nous avons divisé le jeu de données adult.csv en deux parties : deux tiers pour l'apprentissage et un tiers pour le test. Cette division a été répétée **cinq fois** avec des répartitions aléatoires différentes pour garantir la robustesse de l'évaluation.

À chaque itération, un modèle Naive Bayes a été entraîné à l'aide de la fonction naiveBayes de la librairie e1071, puis testé sur les données restantes. Le **taux de bien classés (TBC)** a été calculé et la **moyenne des cinq TBC** a été utilisée comme mesure globale de performance.

Nous avons également utilisé « set.seed() » pour garantir la reproductibilité des répartitions aléatoires.

Enfin, nous avons comparé le TBC moyen obtenu au pourcentage de la classe majoritaire. Si le TBC moyen est significativement supérieur à cette classe, alors le modèle peut être considéré comme performant. Dans notre cas, Naive Bayes a montré une meilleure performance (0.829) que la classe majoritaire (0.759).

```
Classe majoritaire : <=50K
TBC moyen sur 5 essais : 0.8293
Taux de la classe majoritaire : 0.7592
Le modèle Naive Bayes fait mieux que la classe majoritaire.
-> NB a une performance acceptable.
```

Gestion des attributs qualitatifs

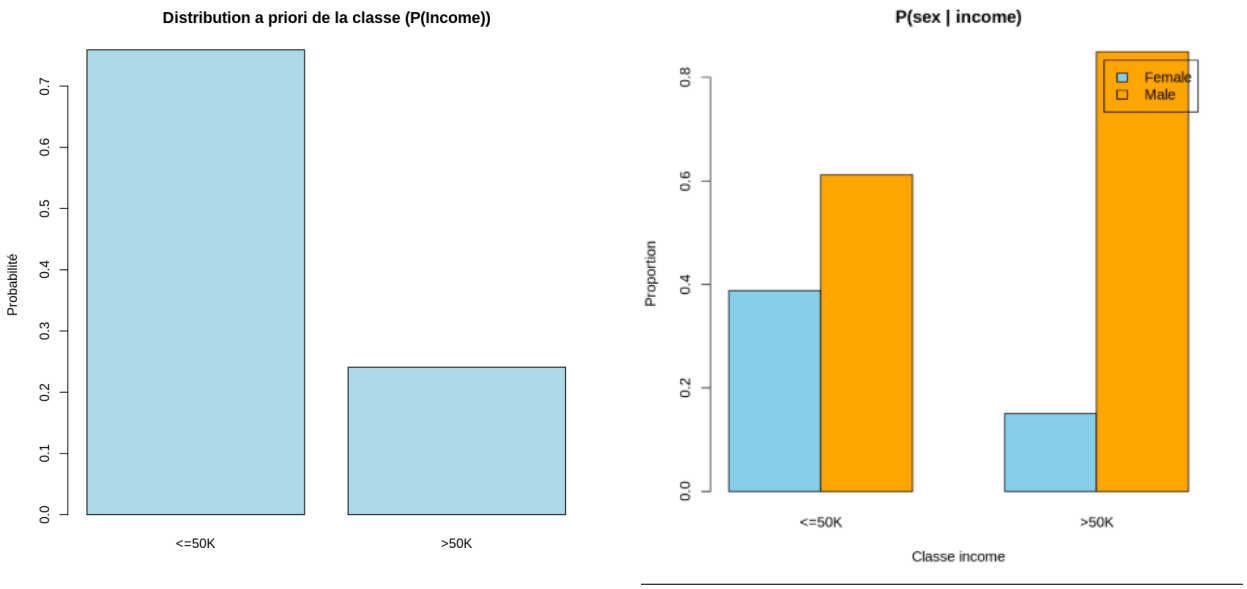
Naive Bayes prend en compte les attributs qualitatifs en apprenant, pour chaque classe, la probabilité conditionnelle associée à chaque modalité. Par exemple, il évalue des expressions du type :

- P(education = Bachelors | income = <=50K)
- P(sex = Female | income = >50K)

Ces probabilités sont extraites automatiquement à partir des fréquences observées dans les données d'apprentissage, et stockées dans le modèle.

Dans le TP1, nous avions déjà estimé ce genre de distribution manuellement (par exemple, la proportion de personnes mariées dans chaque classe de revenu). Le modèle Naive Bayes automatise ce processus pour l'ensemble des attributs qualitatifs.

Nous avons visualisé un exemple concret de distribution conditionnelle avec l'attribut sex. La table croisée table(sex, income) permet d'obtenir la fréquence de chaque modalité, que l'on peut ensuite normaliser pour approcher P(sex | income). Cette distribution est facile à représenter avec un barplot. On observe clairement que les hommes sont sur-représentés dans la classe >50K, ce que le modèle utilise pour ses prédictions.



La distribution a priori des classes (P(C)) a également été observée. On note que la classe <=50K représente environ 76 % du total.

Lien avec les distributions du premier TP

Les distributions apprises par Naive Bayes pour les attributs qualitatifs sont très proches de celles que nous avions estimées manuellement dans le TP1. Nous avions notamment mesuré les proportions de personnes ayant un certain niveau d'éducation, selon leur classe de revenu.

On retrouve ici les mêmes logiques que dans le TP1, mais automatisées par le modèle.

Utilisez un attribut quantitatif de votre choix pour expliquer ce que Naive Bayes fait avec les attributs quantitatifs et comment il calcule les probabilites conditionnelles par classe en utilisant le loi Normal/Gaussien.

Dans le TP1_B, nous avons constaté que l'attribut age suit approximativement une loi normale, avec une moyenne d'environ 36 ans et un écart-type proche de 13. Naive Bayes reprend exactement ces statistiques conditionnelles : pour chaque classe de revenu (<=50K et >50K), il calcule automatiquement la moyenne et l'écart-type de l'âge.

Ainsi, si un individu a 40 ans, NB calcule deux densités gaussiennes (une pour chaque classe) Ces densités jouent le rôle de P(a|c) dans la formule de Bayes, exactement comme nous avions tracé dans TP1_B la distribution réelle d'âge (histogramme conditionnel) et la comparions à une gaussienne théorique. En résumé, Naive Bayes reprend les mêmes moyennes et écarts-types que ceux du TP1 B, mais les utilise "en arrière-plan" pour construire et évaluer une courbe de densité continue par classe, plutôt que de simples fréquences.

Nous avons analysé une ligne réelle du jeu de données (ligne 123), correspondant à un homme de 30 ans.
Voici les probabilités estimées automatiquement par le modèle :

Élément	Classe <=50K	Classe >50K
P(age = 30 | classe)	0.0253	0.0152
P(sex = Male | classe)	0.612	0.8496
P(classe)	0.7592	0.2408
Score total (produit)	0.01176	0.00310

Le modèle prédit donc la classe <=50K. Nous avons confirmé que cette prédiction correspond bien à celle retournée automatiquement par la fonction predict()

Gestion des attributs quantitatifs

Contrairement aux attributs qualitatifs, Naive Bayes ne compte pas les fréquences pour les variables numériques. Il suppose plutôt que, dans chaque classe, les valeurs suivent une loi normale (ou gaussienne).

Il apprend deux paramètres pour chaque classe :

- La moyenne (μ)
- L'écart-type (σ)

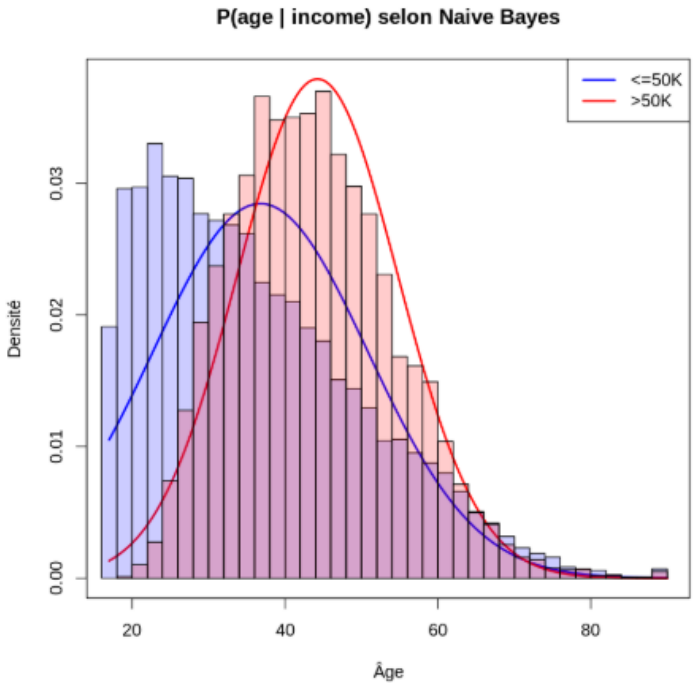
Ces deux quantités sont ensuite utilisées pour estimer P(A | C) avec la formule de la densité de la loi normale. Cela permet au modèle de construire une courbe de probabilité continue pour chaque attribut et chaque classe, plutôt que de simplement compter des occurrences.

Nous avons choisi l'attribut age, qui suit grossièrement une distribution normale d'après le TP1. Le modèle a appris les paramètres suivants :

Classe	Moyenne	Écart-type
<=50K	36.70	13.95
>50K	44.24	10.48

Grâce à ces paramètres, Naive Bayes peut estimer la densité P(age | income) avec la fonction dnorm(). Ces densités sont ensuite multipliées aux autres probabilités pour effectuer une classification.

Nous avons représenté ces courbes gaussiennes sur un graphique en les superposant aux histogrammes réels de l'âge, séparés par classe.



La courbe orange suit bien la forme globale de l'histogrammes, malgré quelques écarts mineurs observables aux extrémités des distributions. Cela confirme que l'approximation gaussienne, bien que simplifiée, reste pertinente pour ce cas précis mais on voit une nette différence avec la courbe bleue qui est plus approximatif.

Conclusion

Ce TP nous a permis d'approfondir notre compréhension du classifieur Naive Bayes en l'appliquant à un cas réel. Nous avons vu :

- Comment il gère automatiquement les attributs qualitatifs par estimation de proportions
- Comment il suppose une loi normale pour les attributs quantitatifs
- Comment il combine ces estimations avec les probabilités a priori pour prédire la classe
- Et enfin, comment interpréter ses décisions à travers un exemple concret

Le modèle s'avère simple à mettre en œuvre, rapide à entraîner, et suffisamment performant sur ce type de données. Il offre également une bonne lisibilité sur les mécanismes internes de classification, ce qui en fait un bon point d'entrée dans le domaine des modèles probabilistes supervisés.