Présentation de l'étude

Manipulation des données

données

Variables qualitativ

Résultats primaires

Amenoration du modèle

du modele

Transformation des

Variables

Comparaison avec les mélanges gaussiens Transformation des

Résultats

Pertinence du mélange

Classification non supervisée Etude des étoiles

DUQUESNE Théo GUEYE Taliesin NAVARRE Victor PINEAU Benjamin Manipulation des données

Représentations de données

Variables qualitatives Résultats primaires

Amélioration

Sélection de modèl Transformation des données

Variables

Comparaison avec le mélanges gaussiens Transformation des données

Pertinence d mélange gaussien

Sommaire

- 1 Présentation de l'étude
- Manipulations des données Représentations des données Variables qualitatives Résultats primaires
- 3 Amélioration du modèle
 Sélection de modèle
 Transformation des données
- Variables qualitatives Comparaison avec les mélanges gaussiens Transformation des données
 - Transformation des données Résultats
- 6 Pertinence du mélange gaussien

Présentation de l'étude

Manipulation des données

données

Variables qualitative

Kesultats primaires

du modèle

Transformation des

Variables qualitative

Comparaison avec les mélanges gaussiens Transformation des

Résultats

Pertinence d mélange

- 1 Présentation de l'étude
- 2 Manipulations des données
- 3 Amélioration du modèle
- 4 Variables qualitatives
- 5 Pertinence du mélange gaussien

Présentation de l'étude

Manipulation des données

Représentations de

Variables qualitati

Résultats primaires

Sélection de modèle
Transformation des

Variables qualitatives

Comparaison avec les mélanges gaussiens Transformation des données

Résultats

Pertinence mélange gaussien

Jeu de données

On dispose d'un jeu de données comprenant des informations sur des étoiles tel que :

- Température (K)
- Rayon
- Couleur
- Classe spectrale
- ..

Objectif : classifier les étoiles selon leur classe : Naine rouge, Supergiant, Hypergiant ...

Jeu de données

Présentation de l'étude

Manipulation des données

Représentations de données

Variables qualitative

Résultats primaires

Sélection de modè Transformation de

Variables

Comparaison avec les mélanges gaussiens Transformation des données

Résultats

Pertinence omélange gaussien

Comment faire ?

Classification non supervisée!

- 1 Le jeu de données contient les classes des étoiles
- 2 On supprime cette colonne lors de l'apprentissage : on essaie de classifier sans connaître la classe
- 3 On vérifie le résultat en comparant avec la classe connue

Présentatio de l'étude

Manipulations des données

Représentations des données

Variables qualitatives

Amélioration

Sélection de modèle Transformation des

données

Variables qualitative

Comparaison avec le mélanges gaussiens Transformation des données

Résultats

Pertinence omélange gaussien

- 1 Présentation de l'étude
- 2 Manipulations des données Représentations des données Variables qualitatives Résultats primaires
- 3 Amélioration du modèle
- 4 Variables qualitatives
- 5 Pertinence du mélange gaussien

DUQUESNE
Théo
GUEYE
Taliesin
NAVARRE
Victor
PINEAU
Benjamin

Présentatio de l'étude

Manipulation des données

Représentations des

Variables qualitative

Amélioration du modèle

Sélection de modèle Transformation des

Variables

Comparaison avec les mélanges gaussiens Transformation des données

Résultats

Pertinence du mélange

Visualisation

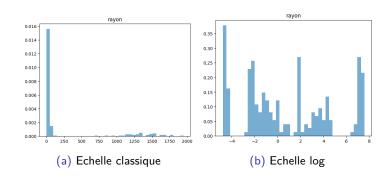


Figure: Histogramme selon le rayon

Manipulation des données

Représentations des données

Variables qualitatives

Résultats primaires

du modéle Sélection de mod

Transformation des

Variables

Comparaison avec les mélanges gaussiens Transformation des données

Résultats

Pertinence d mélange gaussien

Visualisation

Objectif:

- Meilleure séparation des clusters
- Réduction de l'effet des valeurs extrêmes
- Le mélange gaussien suppose que chaque composante suit une distribution gaussienne

Représentations de

Variables qualitatives

Résultats primaires

A ... (II I . . .

Sélection de mo

Transformation des

Variables qualitative

Comparaison avec les mélanges gaussiens Transformation des données

Résultats

Pertinence d mélange gaussien

Variables qualitatives

Changement des variables catégorielles :

- Couleur \longleftrightarrow 0, 1, 2 ...
- Classe spectrale M, G, O ... ←→ 0, 1, 2 ... Regroupement de certaines classes spectrales ? M et K représentent le même type d'étoile

Résultats primaires

Transformation des

Transformation des

Variables qualitatives

Premiers résultats avec 6 clusters et une matrice de covariance 'full' $(\Sigma_{\sigma} = \lambda_{\sigma} Q_{\sigma} A_{\sigma} Q_{\sigma}^{\top})$:

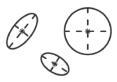


Figure: Matrice de Covariance Σ_g

⇒ Résultats non convaincants

Présentation de l'étude

Manipulation des données

Représentations des données

Variables qualitative

Résultats primaires

Amélioration

du modèle Sélection de modèle

Transformation des

Variables qualitatives

Comparaison avec le mélanges gaussiens Transformation des

Résultats

Pertinence o

- 1 Présentation de l'étude
- Manipulations des données
- 3 Amélioration du modèle Sélection de modèle Transformation des données
- 4 Variables qualitatives
- 5 Pertinence du mélange gaussien

Manipulation des données

Représentations des données

Variables qualitatives Résultats primaires

Résultats primaires

Sélection de modèle Transformation des

Variables

Comparaison avec le mélanges gaussiens Transformation des données

Pertinence omélange gaussien

Sélection du modèle

Décomposition Volume-Shape-Orientation de Σ_g :

$$\Sigma_g = \lambda_g Q_g A_g Q_g^{\top}$$

avec

- Q_g la matrice des vecteurs propores de Σ_g (orientation)
- A_g la matrice diagonale proportionnelle comosée des valeurs propres de Σ_g (forme)
- λ_g la constante de proportionalité associée (volume)

Présentation de l'étude

Manipulation des données

Représentations de

Variables qualitativ

Amélioration

Sélection de modèle

données

Variables qualitative

Comparaison avec le mélanges gaussiens Transformation des données

Kesultats

mélange gaussien

Sélection du modèle

Choix du modèle : calcul du Bayesian Information Criterion (BIC)

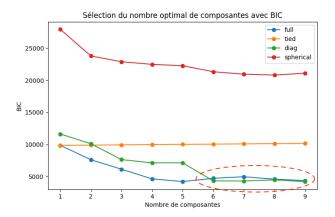


Figure: Calcul du BIC

Manipulation des données

Représentations de données

Variables qualitatives

Amélioration

Sélection de modèle

données Variables

Comparaison avec le mélanges gaussiens Transformation des

données Résultats

Pertinence omélange

Plongement dans un espace plus grand

- On est bloqués par le fait que les variables soient catégorielles
- On aimerait éviter de mettre une relation d'ordre lors de la transformation en variables scalaires

Solution : Plonger dans un espace de dimension plus grand qui respecte ces critères

des données

Représentations de données

Variables qualitatives Résultats primaires

Résultats primaires

Sélection de modèle Transformation des données

Variables

Comparaison avec le mélanges gaussiens Transformation des données

Pertinence du mélange gaussien

Plongement dans un espace plus grand

Pour chaque catégorie, on crée une dimension par classe possible (on les note "is white", "is M" etc...)

Chacun des points va être emmené sur la coordonée 0 ou 1 suivant sa classe

Avantages : On se sépare de la relation d'ordre qu'on avait

On peut toujours espérer que la distance euclidienne ait une pertinence

Inconvénients : C'est artificiel, toujours pas continu, et on passe d'une dimension 8 à une dimension 25

DUQUESNE
Théo
GUEYE
Taliesin
NAVARRE
Victor
PINEAU
Benjamin

Présentatio de l'étude

Manipulation des données

Représentations d données

Variables qualitativ

Amélioration

Sélection de modè

Transformation des données

Variables qualitatives

Comparaison avec les mélanges gaussiens Transformation des données

Pertinence d mélange

Modèle de Mixture Gaussienne

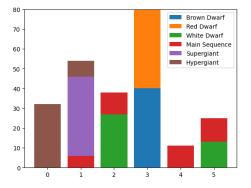


Figure: mixture gaussienne à 6 classes en dimension 25

Présentation de l'étude

Manipulation

Représentations de

Variables qualitative

Kėsultats primaires

du modèle

Sélection de modèle Transformation des données

Variables qualitatives

Comparaison avec les mélanges gaussiens Transformation des données

Résultats

Pertinence d mélange gaussien

Modèle de Mixture Gaussienne

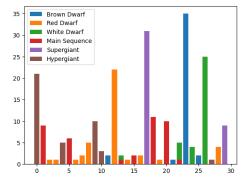


Figure: mixture gaussienne à 30 classes en dimension 25

Manipulations des données

Représentations des données

Variables qualitatives

Amélioration

Sélection de modèl

Transformation des données

Variables

Comparaison avec les mélanges gaussiens Transformation des données

Résultats

Pertinence de mélange

Solution

Idée : On garde cette mixture gaussienne à plein de classes, et on élague à l'aide d'une métrique (un peu comme un arbre)

DUQUESNE
Théo
GUEYE
Taliesin
NAVARRE
Victor
PINEAU
Benjamin

Présentation de l'étude

Manipulation des données

Représentations de

Variables qualitativ

Résultats primaires

Amélioratior du modèle

au modele

Transformation des

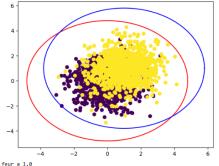
Variables qualitatives

Comparaison avec les mélanges gaussiens Transformation des données

Résultats

Pertinence d mélange

Mesure par quantiles



feur_b 1.0 0.9551847028649821

Figure: Illustration métrique

DUQUESNE Théo GUEYE Taliesin NAVARRE Victor PINEAU Benjamin

Présentation de l'étude

Manipulation des données

Représentations de

Variables qualitativ

Résultats primaires

Amenoration du modèle

Sélection de mod

Transformation des

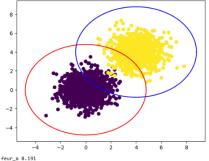
Variables qualitatives

Comparaison avec les mélanges gaussiens Transformation des données

Résultats

Pertinence d mélange

Mesure par quantiles



feur_a 0.191 feur_b 0.188 0.6193490910947782

Figure: Illustration métrique

Présentation de l'étude

Manipulation des données

Représentations de

Variables qualitativ

Résultats primaires

Amelioration du modèle

du modéle

Transformation des

Variables qualitatives

Comparaison avec les mélanges gaussiens Transformation des données

Résultats

Pertinence d mélange

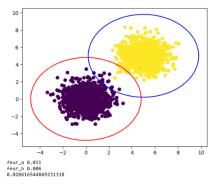


Figure: Illustration métrique

Présentation de l'étude

Manipulation des données

Représentations de

Variables qualitativ

Résultats primaires

Amélioration du modèle

du modèle

Transformation des

Variables qualitatives

Comparaison avec les mélanges gaussiens Transformation des données

Résultats

Pertinence d mélange

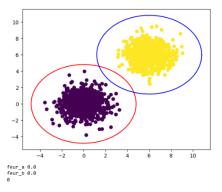


Figure: Illustration métrique

DUQUESNE Théo GUEYE Taliesin NAVARRE Victor PINEAU Benjamin

Présentation de l'étude

Manipulation des données

Représentations de données

Variables qualitative

Amélioration

du modèle

Transformation des

Variables qualitatives

Comparaison avec les mélanges gaussiens Transformation des données

Résultats

Pertinence d mélange

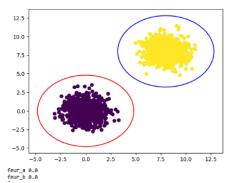


Figure: Illustration métrique

DUQUESNE Théo GUEYE Taliesin NAVARRE Victor PINEAU Benjamin

Présentation de l'étude

Manipulation des données

données

/ariables qualitative

Résultats primaires

Amélioration

du modèle

Transformation des

Variables qualitatives

Comparaison avec les mélanges gaussiens Transformation des données

Résultats

Pertinence di mélange

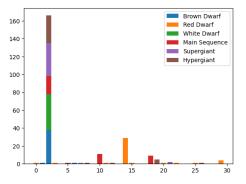


Figure: Résultat après élaguage

Victor PINEAU Benjamin

Présentatio de l'étude

Manipulation des données

données

Résultate primaires

Kėsultats primaires

du modele

Transformation des

Variables

Comparaison avec le mélanges gaussiens Transformation des

Résultats

Pertinence d mélange

Solutions?

- Autre métrique
- Pénalisation
- Réduction de dimension
- Featurisation

Présentatio de l'étude

Manipulation des données

Représentations de données

Variables qualitative

Résultats primaires

Amélioration

Sélection de mod

Transformation des données

Variables qualitatives

Comparaison avec le mélanges gaussiens Transformation des

Résultats

Pertinence mélange Présentation de l'étude

2 Manipulations des données

3 Amélioration du modèle

4 Variables qualitatives

Comparaison avec les mélanges gaussiens Transformation des données Résultats

5 Pertinence du mélange gaussien

Manipulation des données

Représentations des données

Résultats primaires

Amélioration

Sélection de modèle Transformation des

Variables

Comparaison avec les mélanges gaussiens

Transformation des données Résultats

Pertinence d mélange gaussien

Du continue au discret

Utilisation de variables qualitatives : Star type, Star color ou Spectral Class

Nous disposons d'un vecteur aléatoire multivarié $Y = (Y_1, ..., Y_d)$ dont les composantes sont des variables aléatoires discrètes avec $m_1, ...m_d$ niveaux

On passe du cas continu au cas discret \Rightarrow on ne calcule plus la moyenne μ_g et la matrice de covariance Σ_g mais un vecteur de probabilité d'appartenance pour les variables catégorielles Y_j

Benjamin

$$p(y) = \sum_{g=1}^{G} \tau_g f_g(y, \theta_g)$$

à la densité de mélange discrète

$$p(y) = \sum_{g=1}^{G} \tau_g \prod_{j=1}^{d} p_{g,j,y_j}$$

Présentation de l'étude

Manipulation des données

Représentations des données

Variables qualitative

Résultats primaires

Sélection de modè

Transformation des données

Variables qualitative

Comparaison avec les mélanges gaussiens Transformation des données

Résultats

mélange gaussien

Présentation de l'étude

Manipulation des données

Représentations de données

Variables qualitative

Amélioration du modèle

Sélection de modèle Transformation des

Variables

Comparaison avec les mélanges gaussiens Transformation des données

Résultats

Pertinence mélange gaussien

Transformation des données

Certaines variables ne sont pas qualitatives : le rayon, la luminosité ...

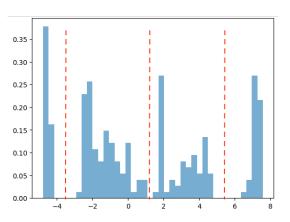


Figure: Histogramme selon le rayon (échelle log)

Présentatio de l'étude

Manipulation des données

Représentations de données

Variables qualitativ

Résultats primaires

Sélection de modèle Transformation des

Variables

Comparaison avec les mélanges gaussiens Transformation des

Résultats

Pertinence di mélange

Résultats avec StepMix

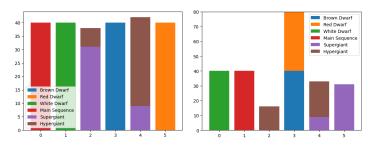


Figure: Instabilité de la méthode

Présentatio de l'étude

Manipulation des données

Représentations de données

Variables qualitativ

Amélioration

Sélection de modé Transformation de

Variables qualitative

Comparaison avec le mélanges gaussiens Transformation des données

Résultats

Pertinence du mélange gaussien

Distance de Wasserstein

Est il pertinent de modéliser les distributions par des distributions gaussiennes ? On calcule la distance de Wasserstein.

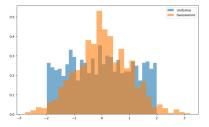


Figure: Comparaison des distributions : Uniforme vs Gaussienne

la distance de Wasserstein est W = 0.24

Présentatio de l'étude

Manipulation des données

Représentations de données

Variables qualitative

Amélioratio

Sélection de modèle
Transformation des

Variables

Comparaison avec les mélanges gaussiens Transformation des données

Pertinence du mélange gaussien

Distance de Wasserstein pour le dataset

Distance de Wasserstein entre le dataset et la loi gaussienne $\mathcal{N}(\mu_{\mathcal{G}}, \Sigma_{\mathcal{G}})$

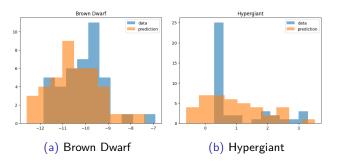


Figure: Pertinence de l'approche gaussienne

DUQUESNE
Théo
GUEYE
Taliesin
NAVARRE
Victor
PINEAU
Benjamin

Présentation de l'étude

Manipulation des données

données

variables qualitativ

Amélioration

du modèle

Transformation des

Variables

Comparaison avec les mélanges gaussiens Transformation des

Dácultate

Pertinence du mélange gaussien

Type d'étoile	Distance
Brown Dwarf	0.1972
Red Dwarf	0.1343
White Dwarf	0.3039
Main Sequence	0.3690
Supergiant	0.3780
Hypergiant	0.3689

Table: Tableau des distances de Wasserstein