Aspects locaux et globaux de l'apprentissage machine KNN

Noé Aubin-Cadot

5 février 2020

But

But : Déterminer le continent d'un point terrestre :

$$(\text{longitude } \phi, \text{latitude } \theta)$$

i.e. établir une fonction :

$$f : [-\pi, \pi) \times [-\pi/2, \pi/2] \rightarrow \{0, 1, 2, 3, 4\}$$

où :

$$\begin{aligned} 0 &= \textit{Amérique} \\ 1 &= \textit{Europe} \\ 2 &= \textit{Afrique} \\ 3 &= \textit{Asie} \\ 4 &= \textit{Océanie} \end{aligned}$$

Idée : Pour déterminer le continent où je me trouve, demander à mes k plus proches voisins le continent où ils se trouvent.

Plan

Plan :

1. Trouver des données.
2. Préparer les données.
3. Visualiser les données.
4. Apprentissage machine sur les données.
5. Frontières de décision.
6. Analyse des résultats.
7. Ouverture.

Trouver des données

On veut des données qui contiennent :

- *source* $\mathbf{X}$ = position (ϕ, θ).
- *but* $\mathbf{y}$ = continent.

Considérons les deux tables suivantes :

- table [1] : pays et continents (249 lignes).
- table [2] : villes, positions (ϕ, θ), pays (>3M lignes).

La jointure de [1] et [2] selon le pays donne, en se limitant aux villes de population de plus de 200K habitants, la table qui nous intéresse :

- table 3 : positions (ϕ, θ), continents (1661 lignes).

Préparer les données

Problème 1 : Saint-Pierre-et-Miquelon, par exemple, appartient à la France mais est en Amérique et non en Europe. On se ferme les yeux là-dessus.

Problème 2 : La surface de la Terre est une sphère $S^2 \subset \mathbb{R}^3$.

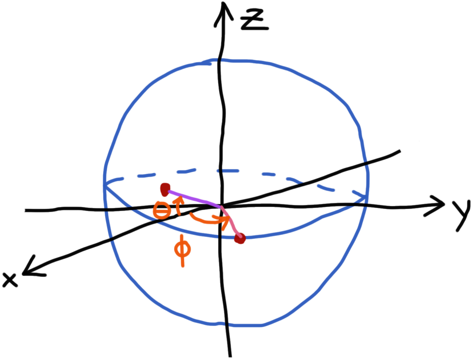
$\Longrightarrow$ La distance euclidienne $d_{\mathbb{R}^2}$ sur le plan (ϕ, θ) n'est pas réaliste pour deux raisons :

1. **Aspect local :** la sphère S^2 est à courbure scalaire positive, donc non localement isométrique à un plan $\mathbb{R}^2$.
2. **Aspect global :** la sphère S^2 n'est pas homéomorphe à $\mathbb{R}^2$.

Préparer les données

<u>Idée :</u> un peu de *feature engineering*, on passe en $3D$:

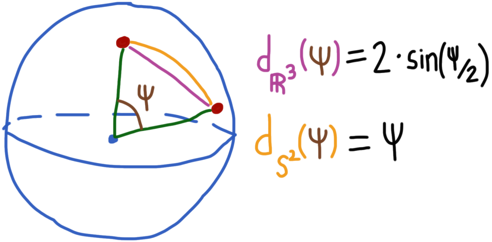
$$\begin{aligned} x &= \cos(\theta)\cos(\phi) \\ y &= \cos(\theta)\sin(\phi) \\ z &= \sin(\theta) \end{aligned}$$



Préparer les données

Sur $S^2 \subset \mathbb{R}^3$ il y a deux distances :

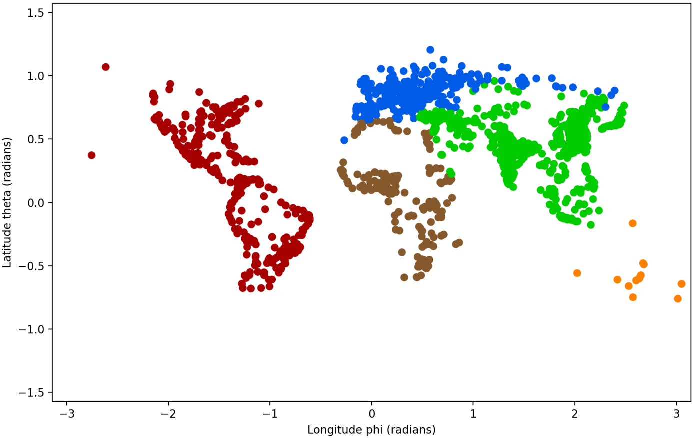
- la distance euclidienne $d_{\mathbb{R}^3}$
- la distance longueur d'arc de cercle d_{S^2}



Pour KNN, les deux métriques $d_{\mathbb{R}^3}$ et d_{S^2} sont équivalentes.
J'utiliserai $d_{\mathbb{R}^3}$.

Visualiser les données

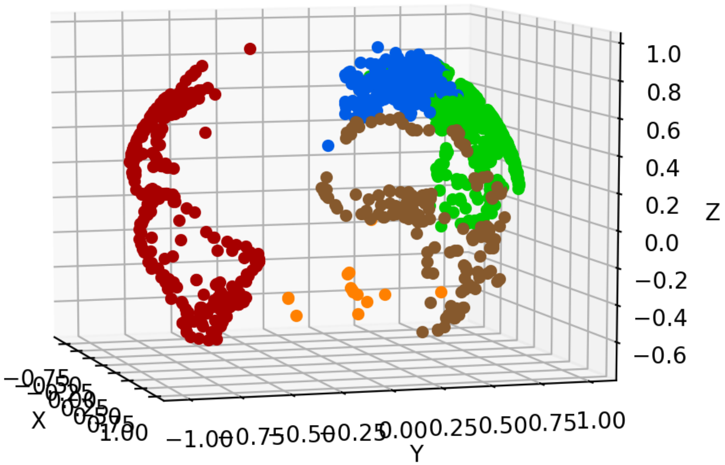
Continents des villes (>200K habitants) dans le plan (ϕ, θ) :



Amérique, Europe, Afrique, Asie, Océanie.

Visualiser les données

Continents des villes (>200K habitants) sur $S^2 \subset R^3$:



Amérique, Europe, Afrique, Asie, Océanie.

Apprentissage machine sur les données

On scinde les données $(\mathbf{X}, \mathbf{y})$ en deux sous-ensembles :

- 75% : *entraînement* $(\mathbf{X}_{\text{train}}, \mathbf{y}_{\text{train}})$, 1245 lignes.
- 25% : *test* $(\mathbf{X}_{\text{test}}, \mathbf{y}_{\text{test}})$, 416 lignes.

On entraîne un classificateur sur les données d'entraînement et on évalue ses résultats sur les données de test.

On peut essayer divers classificateurs `scikit-learn` e.g. KNN, BNG, BNB, SVM, lbfgs, liblinear, RFC, Perceptron, SGDC, DTC, etc.

Apprentissage machine sur les données

Scores d'apprentissage pour la métrique $d_{\mathbb{R}^2}$ sur l'espace (θ, ϕ) et pour la métrique $d_{\mathbb{R}^3}$ sur $\mathbb{R}^3$:

Nom	Train	Test
KNN	100.0%	98.8%
BNG	97.2%	96.4%
BNB	71.7%	67.8%
SVM	95.3%	94.7%
lbf	95.7%	95.2%
lib	92.5%	93.0%
RFC	100.0%	98.3%
Per	80.1%	79.8%
SGD	94.5%	92.1%
DTC	100.0%	97.1%

Scores pour $d_{\mathbb{R}^2}$

Nom	Train	Test
KNN	100.0%	99.0%
BNG	96.2%	96.4%
BNB	74.4%	71.2%
SVM	95.7%	95.0%
lbf	96.5%	95.7%
lib	94.1%	95.0%
RFC	100.0%	98.8%
Per	90.9%	88.5%
SGD	96.7%	94.7%
DTC	100.0%	97.4%

Scores pour $d_{\mathbb{R}^3}$

Apprentissage machine sur les données

Matrice de confusion $(i, j) = $ (réel, prédit) pour classificateur KNN, $k = 1$, et métrique $d_{\mathbb{R}^3}$:

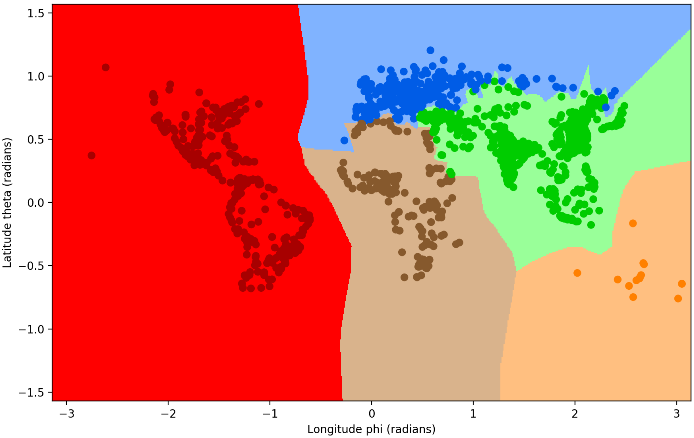
	0	1	2	3	4
0	101	0	0	0	0
1	0	80	0	0	0
2	0	1	47	0	0
3	0	3	0	181	0
4	0	0	0	0	3

Il y a quatre mauvaises classifications sur 416 prédictions :

- 1 ville africaine (Oran en Algérie) prédite en Europe.
- 3 villes asiatiques (Aqtöbe, Kostanaï et Pavlodar au Kazakhstan) prédites en Europe.

Frontières de décision

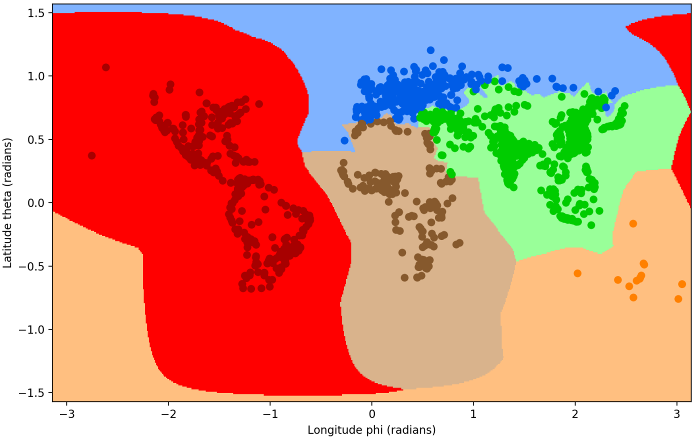
Frontières de décision selon la métrique $d_{\mathbb{R}^2}$ sur le plan (ϕ, θ) :



Amérique, Europe, Afrique, Asie, Océanie.

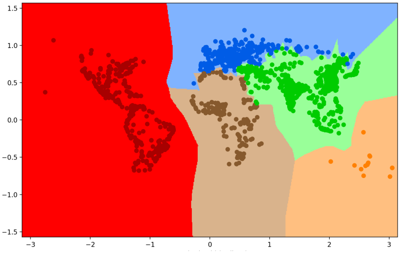
Frontières de décision

Frontières de décision selon la métrique $d_{\mathbb{R}^3}$ sur $S^2 \subset \mathbb{R}^3$:

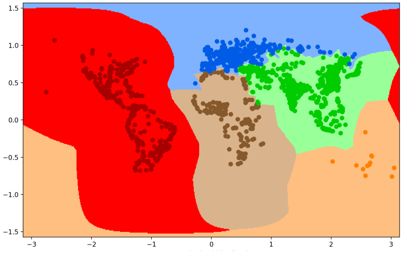


Amérique, Europe, Afrique, Asie, Océanie.

Analyse des résultats



Frontières selon $d_{\mathbb{R}^2}$.



Frontières selon $d_{\mathbb{R}^3}$.

1. **Aspect local :** la forme des frontières de décision change d'une métrique à l'autre, surtout près des pôles.
2. **Aspect global :** la frontière selon $d_{\mathbb{R}^2}$ coupe à $\phi = \pm\pi/2$, mais non celle de $d_{\mathbb{R}^3}$.

Ouverture

Possibilité d’étudier d’autres buts **y** que le *continent* :

1. Risque d’inondations
2. Risque de feux de forêts
3. Risque de tremblements de terre
4. Risque d’accidents nucléaires
5. Accès à l’eau potable
6. Économie locale
7. Employabilité
8. Pollution dans l’air
9. Direction moyenne du vent
10. Diversité des ressources énergétiques
11. Dépendance d’une région envers les énergies fossiles

Ouverture

Possibilité d'étudier d'autres sources $\mathbf{X}$ que la position géographique sur la sphère S^2 :

1. L'heure locale à valeurs en le cercle S^1
2. La position des étoiles sur la sphère céleste S^2
3. La position géographique de deux personnes en $S^2 \times S^2$
4. Le vent sur Terre en $S^2 \times \mathbb{R}^2$
5. Les événements au voisinage d'un trou noir $\mathbb{R}^4 \backslash \mathbb{R}$

Il est aussi possible d'étudier d'autres métriques que la métrique euclidienne sur $\mathbb{R}^n$.

Plus généralement, $\mathbf{X}$ vit sur un espace métrique (X, d).

Merci de votre attention ☺

Références

[1] Chaitanya Gokhale, *Kaggle, country to continent,* `https://www.kaggle.com/statchaitya/country-to-continent`.

[2] Max Mind, *Kaggle, world cities database,* `https://www.kaggle.com/max-mind/world-cities-database`.