Comparaison de méthodes de réduction de dimension pour des analyses de données biologiques

Reunan Bellec & Malo Gillard
Enseignant référent : David Rousseau

28 avril 2019

Table des matières

Introduction

Le dérèglement climatique, le développement de nouvelles maladies des plantes, la maîtrise des rendements, amènent les états et l'ensemble des acteurs de l'agriculture en charge de la sélection variétale, à identifier des semences performantes, résistantes aux maladies, à des périodes de sécheresse ou de brusques variations environnementales durant leur développement. Ces travaux peuvent bénéficier d'avancées technologiques récentes en matière de traitements de l'information, applicables sur de larges populations de plantes. Une échelle particulièrement importante est celle de la graine, dont la qualité germinative conditionne la suite du développement de la plante. Dans ce projet annuel, nous nous intéressons à des graines de betterave sucrière pour laquelle la France est l'un des plus gros producteurs au monde.

L'objectif de l'expérience est d'élargir la variabilité génétique de la betterave, dans le but de la rendre plus compétitive, en doublant le rythme de croissance annuelle de son rendement en sucre. Nous allons donc tester différents génotypes de betteraves (200 individus) et étudier leur germination. À partir des semences des populations de betteraves sélectionnées et des résultats de leur germination, nous observons plusieurs différences, que l'on va chercher à caractériser. Parmi les 200 génotypes, les variables à expliquer sont la surface, la longueur, la largeur, l'imbibition, la vitesse de germination, etc.

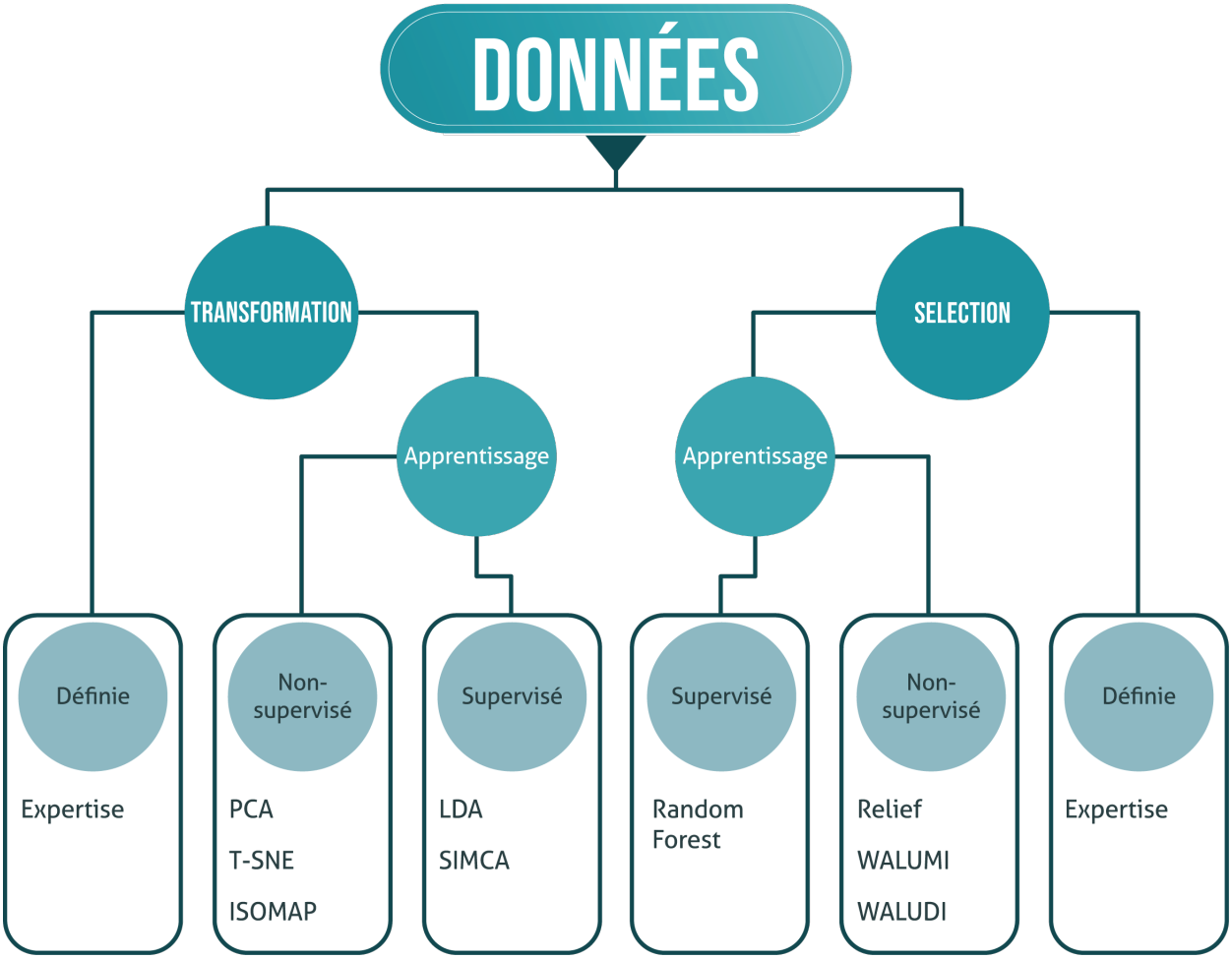
Notre rôle ici sera de comparer plusieurs méthodes de réduction de dimension de l'espace de toutes ces variables, de manière à classifier les différents génotypes par la suite. Nous allons donc considérer plusieurs méthodes afin de répondre à cette problématique : PCA, t-SNE, Random Forest, LDA, etc.

Dans un premier temps, nous étudierons le principe de chaque méthode, ce qui nous permettra d'aborder les notions d'apprentissages supervisé et non supervisé, mais également de classer toutes ces méthodes selon leur fonctionnement. Dans un second temps, après avoir compris chacun des précédés, nous les utiliserons sur les données issues de l'expérience de germination et nous comparerons les résultats.

Chapitre 1

Principe des méthodes

Le schéma suivant nous résume rapidement où se situe chaque méthode.



1.1 Groupe 1

1.1.1 PCA

PARTIE REUNAN

Principe

Explication mathématique

Utilisation du module python sklearn

1.1.2 T-SNE

PARTIE REUNAN

Principe

Explication mathématique

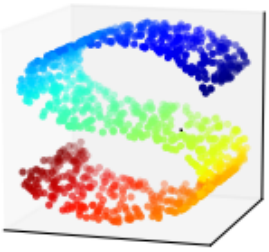
Utilisation du module python sklearn

1.1.3 ISOMAP

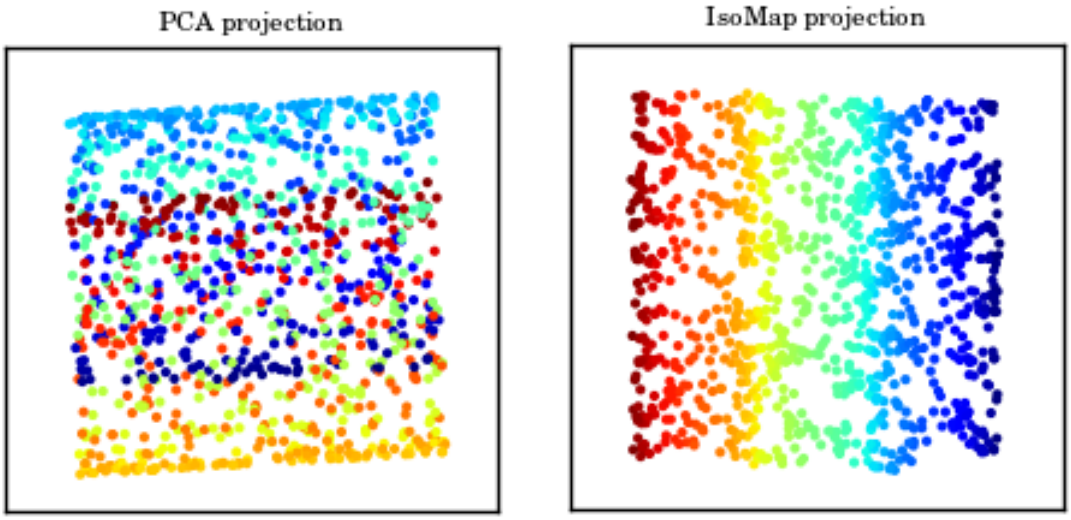
Principe

Tout comme la méthode T-SNE, la méthode ISOMAP est une méthode de réduction de dimension non linéaire. Cependant, contrairement à la méthode PCA, l'approche utilisée ici propose de mieux approximer la structure géométrique réelle de l'ensemble de données à travers la réduction de dimension. Elle est dite non-linéaire du fait qu'elle s'adapte très bien aux structures géométriques non linéaires.

Par exemple, supposons que notre jeu de données soit représenté en trois dimensions par une courbe en S (donc non linéaire) :



La méthode PCA appliquée à ce jeu de données nous donnera comme résultats (en réduisant le nombre de dimension de 1) des valeurs désorganisées, tandis que la méthode ISOMAP préservera la structure locale après avoir réalisé une projection.



La différence vient du fait que deux points peuvent être proches selon une distance euclidienne (utilisée dans le premier cas), mais très éloignés si on mesure la distance sur la **surface** définie par les points, appelée **distance géodésique** (ce que fait la méthode ISOMAP dans le deuxième cas).

Néanmoins il faut prendre en compte le fait que nous ne connaissons pas cette surface (dans le cas où nous avons un ensemble discret de points par exemple), ce qui rend compliquée la tâche d'évaluation des distances géodésiques.
Pour résoudre ce problème, la méthode ISOMAP va construire un graphe d'adjacence des points et approcher la distance géodésique en cherchant le chemin le plus court à travers ce graphe.

Explication mathématique

L'algorithme va se dérouler en 3 étapes.

Étape 1 : *construction du graphe d'adjacence*

Considérons que les données sont représentées par un ensemble X de dimension d. Pour construire le graphe d'adjacence, nous pouvons utiliser deux méthodes : pour chaque point x_i de X, soit chercher les k plus proches voisins $x_1, ..., x_k$ de x_i, soit, utiliser la distance euclidienne pour trouver l'ensemble des points x_j situés dans un certain rayon r (en effet, pour des poins voisins, la distance euclidienne fournit

une approximation juste de la distance géodésique).
Nous représentons ensuite les relations de voisinage par un graphe G : les noeuds sont les points x_i, et le poids de l'arête qui relie x_i à un point de son voisinage correspond à la distance euclidienne entre ces deux points (par défaut le poids de l'arête qui relie deux points ne faisant pas partie du même voisinage est fixé à ∞).
Nous supposerons au préalable que notre structure géométrique est entièrement connectée, c'est-à-dire qu'il n'y a pas de groupes de points isolés.

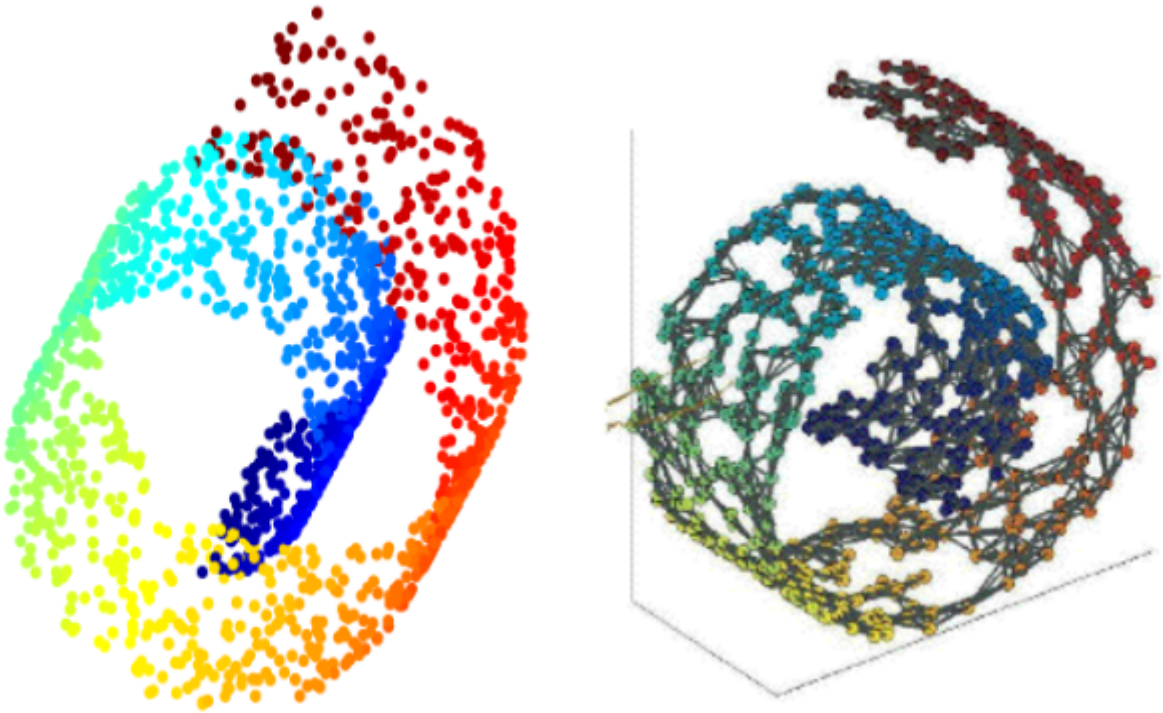


FIGURE 1.1 – Chaque point est relié à ses voisins

Étape 2 : *calcul des distances géodésiques*

L'algorithme va calculer la distance géodésique $d_G(i, j)$ pour chaque paire de points (x_i, x_j), d'après le graphe G, en appliquant un algorithme de recherche du chemin le plus court, comme par exemple l'algorithme **Dijkstra**.
Cela permet ainsi de construire la matrice des distances géodésiques D_G.

Étape 3 : *réduction de dimension*

Utilisation du module python sklearn

1.2 Groupe 2

1.2.1 LDA

Principe

Explication mathématique

Utilisation du module python sklearn

1.3 Groupe 3

1.3.1 Random Forest

PARTIE REUNAN

Principe

Explication mathématique

Utilisation du module python sklearn

1.4 Groupe 4

1.4.1 Relief

PARTIE REUNAN

Principe

Explication mathématique

Utilisation du module python sklearn

Chapitre 2

Réduction de dimension sur des données biologiques

2.1 Expertise

La première tâche à réaliser avant de travailler sur les données au travers des différentes méthodes de réduction de dimension est l'expertise, c'est à dire la préparation, le nettoyage des données. Nous avons réfléchis avec un point de vue autre que celui de data scientist : qu'est-il bon de garder ou d'écarter dans le jeu de données, faut-il rajouter des variables explicatives, et pourquoi.

Comme nous nous intéressons à l'évolution des germinations jour par jour, nous avons écarté les variables qui s'exprimaient en heures, c'est à dire les variables **5°C TMG (h)** et **5°C T50 (h)**, et gardé celles qui s'exprimaient en jours, **5°C TMG (j)** et **5°C T50 (j)**.
Cependant, cette dernière comprenait un nombre important d'entrées "Nan" (plus de la moitié), ce qui signifie qu'il n'a pas été possible de calculer le délai nécessaire pour obtenir 50% de germination pour plus de la moitié des individus. Nous l'avons donc également écartée car elle présentait peu d'intérêt dans notre étude.

Nous avons ensuite rajouté 6 variables **'v15-16j'**, **'v16-17j'**, **'v17-18j'**, **'v18-19j'**, **'v19-20j'**, et **'v20-21j'**, pour exprimer la vitesse de germination de chaque individu jour après jour à partir du jour 15.