Analyse en composantes principales

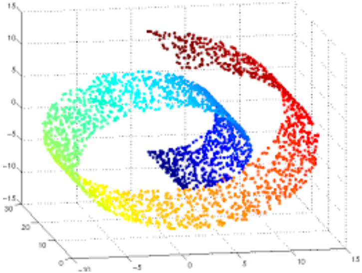
5 février 2019

Réduction de dimensions

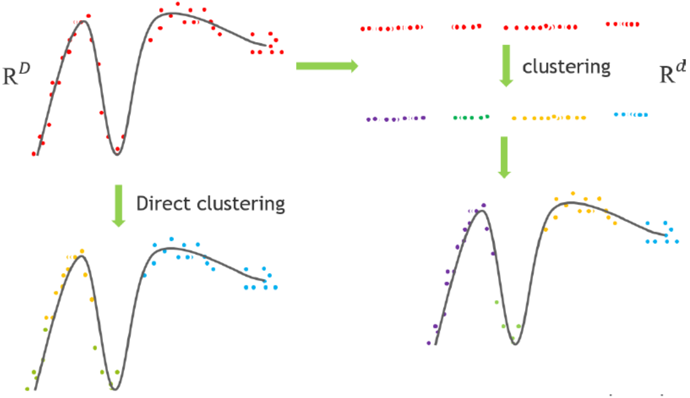
- Les données sont généralement en grande dimension ($d > 100$)
- Les données sont des mesures indirects de source qui sont par essence faiblement dimensionnel
- **réduction de dimensions :** projection des données dans un espace de taille inférieure
- Bénéfices multiples :
  - diminution de l'empreinte mémoire, encode plus compact
  - visualisation des données
  - Prétraitement des données avant classification ou autre ; réduction du fléau de la dimensionalité
- formalisation :
  - Soit $\mathbf{x} \in \mathbb{R}^n$ une donnée.
  - Un opérateur de projection $f : \mathbb{R}^n \rightarrow \mathbb{R}^d$ projette les données dans un espace de dimension réduite $d << n$
  - $\mathbf{y} = f(\mathbf{x})$

Espace Euclidien / variétés

- Espace métrique : on dispose d'une distance (Euclidienne par exemple) pour calculer la distance entre deux données
- **Variété** (*manifold*) : en chaque point, il existe un voisinage (plan tangent) homéomorphe à un espace Euclidien
- Ce voisinage est *localement Euclidien*



Exemple sur du clustering



- **L'analyse en composantes principales (ACP)** est une technique de réduction de dimensions **linéaire** (f est linéaire).
- Soit $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_m] \in \mathbb{R}^{n \times m}$ un jeu de données.
- Soit $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_m] \in \mathbb{R}^{d \times m}$ sa projection (espace latent),
- **Projection** : $\mathbf{Y} = \mathbf{Q}^T \mathbf{X}$, avec $\mathbf{Q}^T \in \mathbb{R}^{d \times n}$
- **Reconstruction** : $\mathbf{X} = \mathbf{Q}\mathbf{Y}$

- On cherche une direction de projection $\mathbf{u}$ qui **maximise la variance** des données projetées
- Variance $\sigma_{\mathbf{u}}$

$$
\begin{aligned}
\sigma_{\mathbf{u}} &= \frac{1}{m}\sum_i (\mathbf{u}^{\mathbf{T}}\mathbf{x}_i)^2 \\
&= \mathbf{u}^{\mathbf{T}}(\frac{1}{m}\sum_i \mathbf{x}_i\mathbf{x}_i^T)\mathbf{u} \\
&= \mathbf{u}^{\mathbf{T}}\Sigma_{\mathbf{X}}\mathbf{u}
\end{aligned}
$$

- $\Sigma_{\mathbf{X}}$ est la matrice de tous les produit scalaires de toutes les données (matrice de Gram)
- Généralement, on cherche $\mathbf{u}$ sous la contrainte de norme unitaire, $||u||_2 = \mathbf{u}^{\mathbf{T}}\mathbf{u} = 1$

Résolution du problème

- utilisation des multiplicateurs de Lagrange pour insérer la contrainte :

$$L(\mathbf{u}, \alpha) = \mathbf{u}^{\mathbf{T}} \Sigma_{\mathbf{X}} \mathbf{u} - \alpha(1 - \mathbf{u}^{\mathbf{T}} \mathbf{u}) \tag{1}$$

- On dérive par rapport à $\mathbf{u}$

$$\frac{\partial L(\mathbf{u}, \alpha)}{\partial \mathbf{u}} = \Sigma_{\mathbf{X}} \mathbf{u} - \alpha \mathbf{u} \tag{2}$$

- On annule ce gradient. La solution vérifie $\Sigma_{\mathbf{X}} \mathbf{u} = \alpha \mathbf{u}$
- $\mathbf{u}$ est un vecteur propre de $\Sigma_{\mathbf{X}}$!
- C'est précisément le vecteur propre associé à la plus grande des valeurs propres

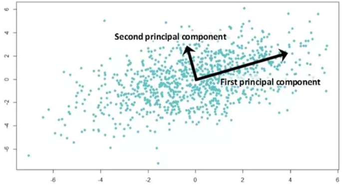
ACP

- L'ACP est une diagonalisation de la matrice de covariance
- Elle construit une base formée par les vecteurs propres
- $\Sigma_{\mathbf{X}} = \mathbf{U} \wedge \mathbf{U}^T = \sum_i \lambda_i \mathbf{u}_i \mathbf{u}_i^T$
- Soit $\mathbf{U}_d = [\mathbf{u}_1, \ldots, \mathbf{u}_d] \in \mathbb{R}^{n \times d}$ la matrice formée par les d premiers vecteurs propres, associées aux d premières valeurs propres λ_i, alors

$$\mathbf{y} = \mathbf{U}_d^T \mathbf{x}$$

- les données projetées sont 'décorrélées' (preuve en cours) :

$$\frac{1}{m} \sum_i \mathbf{y}_i \mathbf{y}_i^T = \wedge_d$$



Conclusion sur la réduction de dimensions

- Il existe un très grand nombre de stratégies pour la réduction de dimensions
  - incluant des labels par exemples (analyse discriminante de Fisher)
  - variantes non-linéaires
  - *manifold learning*
- d'autres heuristiques peuvent être utilisées
  - Préservation du voisinage entre les points
  - Par exemple méthode t-SNE (*stochastic neighbor embedding*), très utilisé en deep learning

Travaux pratiques

- Reprenez l'image hyperspectrale du précédent TP
- En utilisant le package pour faire de l'ACP de scikit-learn, effectuez une ACP sur tous les spectres de l'image
- Affichez la décroissance des valeurs propres (triées)
- Projetez les données en 2D et visualisez le nuage de points correspondant
- Affichez les images correspondant aux trois premières valeurs propres
- s'il vous reste du temps, testez aussi la méthode t-SNE