Analyse en Composantes Principales

Charlotte Pelletier
(Basé sur le cours de N. Courty)

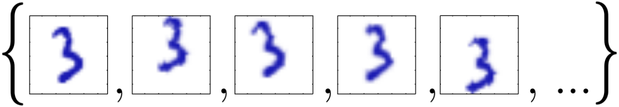
29 janvier 2020

Dimensionnalité des données

Soit un jeu de données généré en prenant une seule image de “3” pour laquelle trois transformations différentes sont appliquées :

1. translations verticales
2. translations horizontales
3. rotations

Chaque image est considérée comme une *observation* :

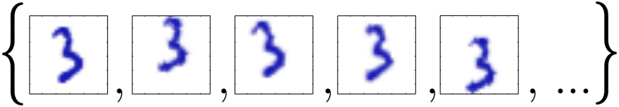


Dimensionnalité des données

Soit un jeu de données généré en prenant une seule image de “3” pour laquelle trois transformations différentes sont appliquées :

1. translations verticales
2. translations horizontales
3. rotations

Chaque image est considérée comme une *observation* :



Même si chaque image a une taille 100×100 ($d = 10\ 000$), la dimensionnalité intrinsèque des données est $d' = 3$.

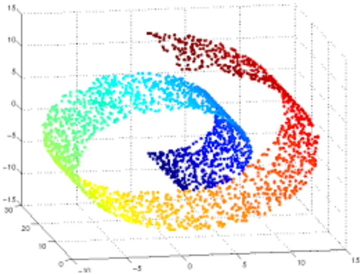
Source : Christopher M. Bishop (2006). Pattern Recognition and Machine Learning. Springer.

Variétés

Lorsque d est grand, on s'attend à ce que les données se trouvent autour d'une **variété** (*manifold*) de dimension $d' < d$.

Formalisme :

- 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*.



Soit des données $\mathbf{x} \in \mathbb{R}^d$, la **réduction de dimensionnalité** consiste à

- projeter les données dans un espace de dimension d' inférieure ($d' \ll d$).
- Bénéfices multiples (valide aussi pour la projection dans un nouvel espace) :
  - encodage plus compact, et donc diminution de l'empreinte mémoire
    $\sim$ compression de données
  - facilite la visualisation des données
  - étape de pré-traitement des systèmes d'apprentissage automatique pour réduire la malédiction de la dimensionnalité

Idée

- les d dimensions où vivent les observations $x_i \in \mathbb{R}^d$ ne sont pas toutes "intéressantes" de la même faccon
- l'ACP cherche à projeter les données dans des dimensions plus "intéressantes" ; *i.e.*, là où on observe une plus grande variation des observations dans chacune des directions
- l'ACP est une technique d'analyse **linéaire** $\rightarrow$ chaque nouvelle dimension trouvée par la PCA est une combinaison linéaire des d dimensions
- l'ACP est une méthode d'aprentissage non supervisé

Des définitions...

1. aussi appelé transformation de Karhunen-Loève (KLT)
2. projection des données dans un espace orthogonal (de plus petite dimension) – Hotelling, 1933
3. projection linéaire qui minimise le coût moyen de projection, définit comme la distance au carré entre les observations et leurs projections – Pearson, 1901

Formalisme

- Soit f un opérateur de projection
  - $f : \mathbb{R}^d \rightarrow \mathbb{R}^{d'}$ projette les données dans un espace de dimension réduite ($d' \ll d$).
  - $\mathbf{y} = f(\mathbf{x})$
- L'ACP est une technique d'analyse **linéaire** :
  - f est linéaire
  - Soit $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_m] \in \mathbb{R}^{d \times m}$ 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 $\mathbf{Q}^T \in \mathbb{R}^{d \times d}$
    - **reconstruction** : $\mathbf{X} = \mathbf{Q}\mathbf{Y}$

- Soit f un opérateur de projection
  - $f : \mathbb{R}^d \to \mathbb{R}^{d'}$ projette les données dans un espace de dimension réduite ($d' \ll d$).
  - $\mathbf{y} = f(\mathbf{x})$
- L'ACP est une technique d'analyse **linéaire** :
  - f est linéaire
  - Soit $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_m] \in \mathbb{R}^{d \times m}$ un jeu de données, soit $\mathbf{Y} = [\mathbf{y}_1, \dots, \mathbf{y}_m] \in \mathbb{R}^{d \times m}$ sa projection (espace latent) :
    - **projection** : $\mathbf{Y} = \mathbf{Q}^T\mathbf{X}$, avec $\mathbf{Q}^T \in \mathbb{R}^{d \times d}$
    - **reconstruction** : $\mathbf{X} = \mathbf{Q}\mathbf{Y}$

Comment déterminer la matrice Q pour projeter les données $\mathbf{X}$?

Maximisation de la variance

Cas $d' = 1$

- **Objectif** : on cherche la direction de projection $\mathbf{u}_1 \in \mathbb{R}^d$ qui **maximise la variance** des données projetées
- Par commodité, on décide que $\mathbf{u}_1$ est un vecteur unitaire, *i.e.*, $\mathbf{u}_1^T \mathbf{u}_1 = 1$
- La projection de chaque observation $\mathbf{x}_i \in \mathbb{R}^d$ donne un scalaire : $\mathbf{u}_1^T \mathbf{x} \in \mathbb{R}$
- Variance $\sigma = \sigma_{\mathbf{u}_1^T \mathbf{x}} = \mathbf{u}_1^T \boldsymbol{\Sigma}_{\mathbf{X}} \mathbf{u}_1$ avec $\Sigma_{\mathbf{X}} \in \mathbb{R}^{d \times d}$ la matrice de covariance des données [démonstration de cours à connaître]

Résolution du problème

- On cherche à maximiser σ ... sans contrainte $\mathbf{u}_1 \to \infty$
- Utilisation des multiplicateurs de Lagrange pour insérer la contrainte :

$$L(\mathbf{u}_1, \alpha) = \mathbf{u}^{\mathbf{T}} \Sigma_{\mathbf{X}} \mathbf{u} + \lambda_1 (1 - \mathbf{u}^{\mathbf{T}} \mathbf{u}) \tag{1}$$

- On dérive par rapport à $\mathbf{u}_1$ (maximisation)

$$\frac{\partial L(\mathbf{u}_1, \alpha)}{\partial \mathbf{u}_1} = \Sigma_{\mathbf{X}} \mathbf{u}_1 - \lambda_1 \mathbf{u}_1 \tag{2}$$

  ($\Sigma_{\mathbf{X}}$ est symétrique)
- On cherche à annuler ce gradient (maximum de $L(\mathbf{u}_1, \alpha)$). La solution vérifie

$$\Sigma_{\mathbf{X}} \mathbf{u}_1 = \alpha \mathbf{u}_1 \tag{3}$$

  $\mathbf{u}_1$ est un vecteur propre de $\Sigma_{\mathbf{X}}$!
- Si $\mathbf{u}_1$ est un vecteur de propre, alors la variance est égale à

$$\sigma = \mathbf{u}_1^T \Sigma_{\mathbf{X}} \mathbf{u}_1 = \mathbf{u}_1^T (\lambda_1 \mathbf{u}_1) = \lambda_1 \tag{4}$$

  Pour maximiser la variance on doit donc prendre le vecteur propre $\mathbf{u}_1$ dont la valeur propre est la plus grande.

Cas $d' > 1$

- processus itératif
- deuxième direction consiste à maximiser la variance des données projetées orthogonalement à toutes celles restantes (sur les résidus)
- on peut montrer que le résultat correspond à garder les d' vecteurs propres $\mathbf{u}_1, \cdots, \mathbf{u}_{d'}$ associés au plus grandes valeurs propres $\lambda_1, \cdots, \lambda_{d'}$

Matrice symétrique réelle

Propriétés

- Les valeurs propres réelles d'une matrice symétrique réelle M sont réelles.
- Les vecteurs propres associés à deux valeurs propres différentes sont orthogonaux [démonstration de cours à connaître].

Rappel :

- la matrice de covariance $\Sigma_{\mathbf{X}}$ est une matrice symétrique réelle,
- donc ses vecteurs propres sont orthogonaux,
- et donc les données projetées sont "décorrélées".

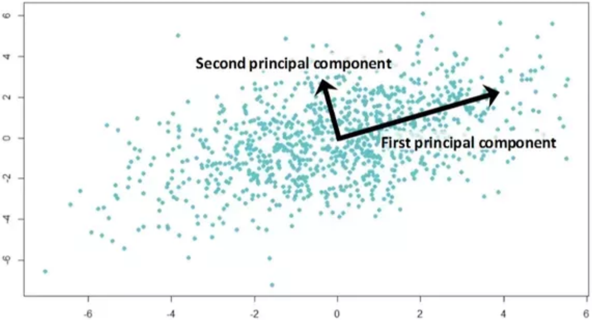
Lien entre ACP et diagonalisation

L'ACP est équivalente à **diagonaliser la matrice de covariance** :

- Soit la base orthonormée formée par les vecteurs propres de $\Sigma_{\mathbf{X}}$, on note P la matrice de changement de base (matrice de passage) :

$$P = [\mathbf{u}_1 \cdots \mathbf{u}_j \cdots \mathbf{u}_d] \tag{5}$$

alors $P^T \Sigma_{\mathbf{X}} P = diag(\lambda)$ [démonstration de cours à connaître]



Conclusion sur la réduction de dimensions

- Il existe un très grand nombre de stratégies pour la réduction de dimension
  - incluant des étiquettes par exemples (analyse discriminante de Fisher)
  - variantes non-linéaires (*e.g.*, Kernel-PCA)
  - *manifold learning*
- D'autres heuristiques peuvent être utilisées
  - préservation du voisinage entre les points
    par exemple, la méthode t-SNE (*Stochastic Neighbor Embedding*), très utilisé en apprentissage profond.