Apprentissage statistique : modélisation descriptive et introduction aux réseaux de neurones (RCP208)

Sélection de variables

Michel Crucianu
(prenom.nom@cnam.fr)
http://cedric.cnam.fr/vertigo/Cours/ml/

EPN05 Informatique
Conservatoire National des Arts & Métiers, Paris, France

12 décembre 2024

Plan du cours

Réduction du nombre de variables d'entrée : objectifs

n données dans $\mathbb{R}^m \rightarrow n$ données dans $\mathbb{R}^k$, $k \ll m$

1. Réduire le volume de données à traiter, tout en conservant au mieux l'information « utile » ← définir ce qu'est information utile
2. Améliorer le rapport signal / bruit en supprimant des variables non pertinentes ← définir ce qu'est une variable non pertinente
3. Réduire la complexité d'un modèle décisionnel (pour améliorer sa généralisation) : le nombre de variables en est une composante
4. Simplifier la maintenabilité de la solution en réduisant le nombre de variables à recueillir pour les nouvelles observations
5. Améliorer la « lisibilité » des données
   - Mettre en évidence des relations entre variables ou groupes de variables
   - Permettre la visualisation ← définir ce qu'il faut mettre en évidence
6. Répondre à la « malédiction de la dimension » (*curse of dimensionality*)

Objectifs différents, critères différents → méthodes différentes

Réduction du nombre de variables : approches

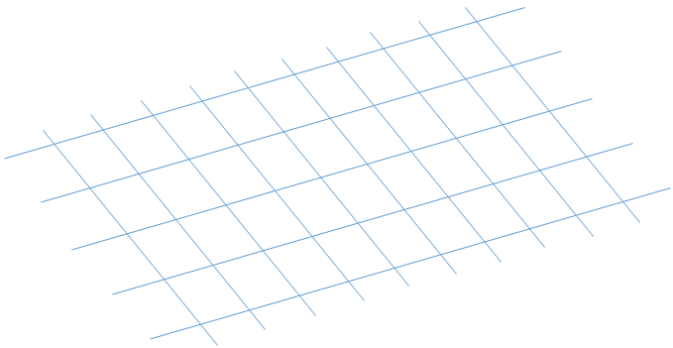
1. Réduction de dimension
   - Modélisation à partir d'un nombre plus faible de variables obtenues par construction de nouvelles variables à partir des variables initiales (par ex. combinaisons linéaires)
   - Plus de flexibilité par rapport à la sélection
   - Mais les nouvelles variables sont rarement interprétables

2. Sélection de variables (*feature selection*)
   - Sélection d'un sous-ensemble de k variables parmi les m variables initiales
   - Les variables sélectionnées gardent leur signification initiale
   - Mais solution potentiellement sous-optimale car cas particulier de la construction de nouvelles variables

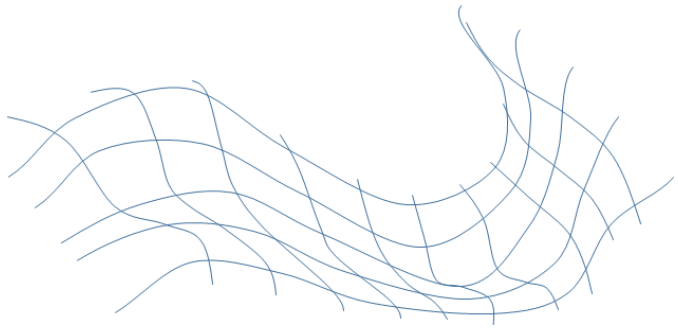
Réduction de dimension : rappel

La dépendance entre les nouvelles variables et les variables initiales peut être

- Linéaire : trouver un sous-espace linéaire de dimension k dans l'espace initial $\mathbb{R}^m$
  - Chapitre 2 : ACP, AFD, ACM
- Non linéaire : trouver un sous-espace non linéaire de dimension faible
  - Chapitre 5 : LLE, t-SNE, UMAP



sous-espace linéaire

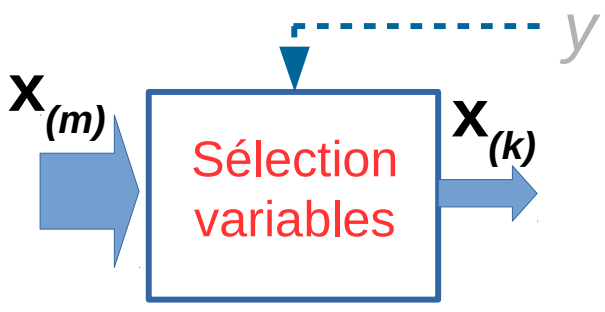


sous-espace non linéaire

Plan du cours

Sélection de variables : approches (voir [2], [4])

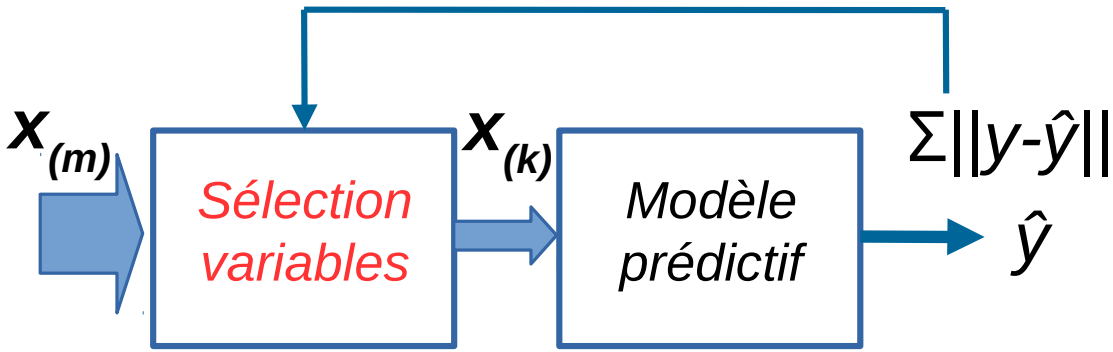
1. Filtrage : appliquées sans faire appel à un modèle prédictif (ou décisionnel) ultérieur
   - Critères de sélection : maximisation de l'information mutuelle entre une variable d'entrée et la variable à prédire, minimisation de la redondance entre variables d'entrée



   - Coopération sous-optimale avec le modèle prédictif ultérieur car celui-ci n'intervient pas dans le processus de sélection

Sélection de variables : approches (2)

2. *Wrapper* : coopération directe avec le modèle prédictif qui emploie les variables
   - Critère de sélection typique : choix du groupe de variables qui maximise les performances du modèle prédictif ultérieur (modèle prédictif et groupe de variables « emballés » ensemble)



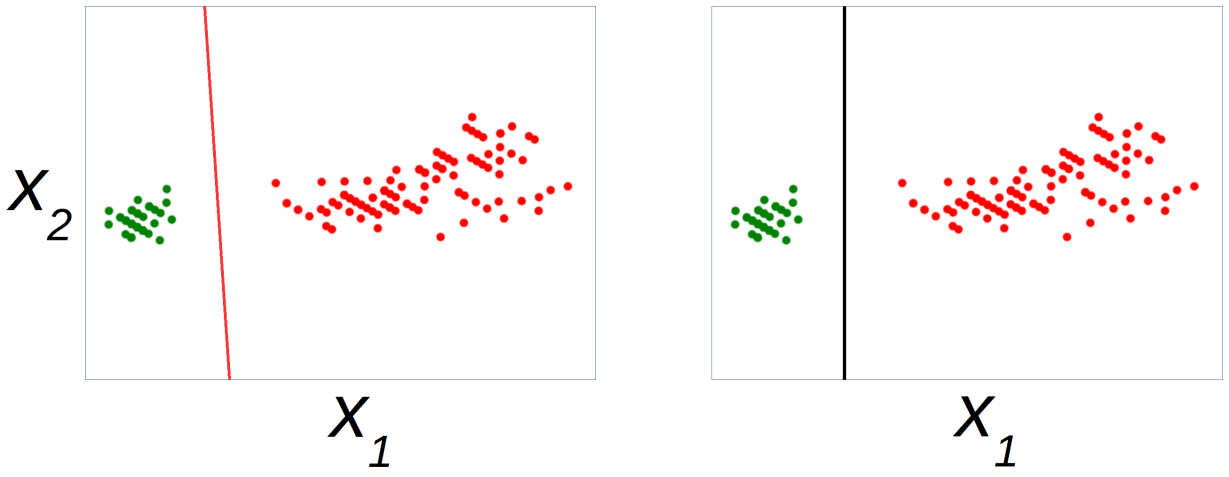
   - Coût élevé : pour évaluer chaque groupe de variables il faut construire un modèle prédictif...

Sélection de variables : approches (3)

3. Intégration (*embedding*) : l'opération de sélection de variables est intégrée à la méthode de construction de modèle prédictif
   - Exemple simple : régularisation L_1 dans la fonction de coût du modèle prédictif

Sans régularisation L_1 : dépendance des 2 variables

Avec régularisation L_1 : dépendance d'une seule variable



   - Pas de surcoût par rapport à la construction du modèle prédictif mais ne peut pas être utilisée avec tout type de modèle

Sélection de variables : exploration de l'espace de recherche

Choisir k variables parmi $m \rightarrow$ espace de recherche $O(C_m^k)$ $\left(= \frac{m!}{k!(m-k)!}\right)$!

$\Rightarrow$ des solutions approximatives (sous-optimales) sont préférées (parfois indispensables) :

1. Critère de « pertinence » exprimable par variable initiale individuelle (indépendamment des autres), puis sélection des k variables de pertinence optimale
   $\Rightarrow$ complexité $O(m)$
2. Critère de « pertinence » exprimable par ensemble de variables initiales (par ex. la redondance pour approche filtrage ou performance globale pour approche *wrapper*)
   1. Méthodes incrémentales (par cooptation) :
      - On démarre avec une seule variable (par ex. choisie avec la méthode précédente)
      - A chaque itération on ajoute une variable, celle qui forme le meilleur ensemble avec les variables déjà sélectionnées aux itérations précédentes

      $\Rightarrow$ complexité $O(m^2)$ (dans le détail dépend du nombre de variables à ajouter)
   2. Méthodes décrémentales (par élimination) :
      - On démarre avec la totalité des variables
      - A chaque itération on teste toutes les combinaisons avec une variable en moins par rapport à l'itération précédente et on choisit celle qui est optimale (on élimine donc une variable)

      $\Rightarrow$ complexité $O(m^2)$ (dans le détail dépend du nombre de variables à éliminer)

Plan du cours

Critères de sélection de variables explicatives

1. Propriétés intrinsèques d'une variable (d'entrée)
   - Variance de la variable : si très faible, la variable explique une faible part de la variance des observations et est éliminée
2. Caractérisation d'une variable d'entrée par rapport à la variable de sortie
   - Qualité de prédiction : sont choisies individuellement les variables d'entrée qui « expliquent le mieux » la variable de sortie
3. Caractérisation d'un groupe de variables d'entrée par rapport à la variable de sortie
   - Qualité de prédiction : est choisi le groupe qui « explique le mieux » la variable de sortie
4. Caractérisation d'un groupe de variables d'entrée entre elles et par rapport à la variable de sortie
   - Qualité de prédiction et réduction de la redondance : est choisi le groupe qui présente le meilleur compromis entre ces deux aspects

Critère de variance

- Ne s'intéresse pas à la relation entre une variable d'entrée et la variable de sortie
- n observations variable $X \Rightarrow$ moyenne et variance de l'échantillon :

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad \mathrm{Var}(x) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

- Idée : les variables de variance trop faible expliquent trop peu de la variance des observations et sont éliminées $\leftarrow$ comme en ACP les axes associés aux valeurs propres les plus faibles de la matrice des covariances empiriques
- La variance dépend de l'unité de mesure : à partir d'un même échantillon de longueurs on obtient une variance bien plus grande si les valeurs sont exprimées en mm plutôt qu'en mètres $\rightarrow$ pour comparer la « dispersion » de variables différentes mieux vaut utiliser le coefficient de variation $= \frac{\text{écart-type}}{\text{moyenne}}$
- Risque : une variable (très) discriminante peut être de faible variance (ou de faible coefficient de variation), comme nous l'avons déjà vu pour l'analyse discriminante

Critères de qualité prédictive individuelle

- Idée : dans quelle mesure la variable d'entrée X « explique » la variable de sortie Y ?
- Différents critères de qualité prédictive individuelle ou « pertinence », par exemple :
  1. Corrélation linéaire : $r = \frac{\mathrm{Cov}(X,Y)}{\sigma(X)\sigma(Y)} \in [-1,1]$, Cov étant la covariance, σ l'écart-type
  2. Test du χ^2 d'indépendance ($\rightarrow$ revoir l'analyse des correspondances binaires)
     - X et Y variables discrètes, prenant chacune un nombre fini de valeurs
     - N observations au total : pour n_{ij} observations X prend sa i-ème valeur et Y sa j-ème valeur, on note $n_{i\_} = \sum_j n_{ij}$, $n_{\_j} = \sum_i n_{ij}$, $e_{ij} = \frac{n_{i\_} n_{\_j}}{N}$
     - Pour chaque variable X on calcule $t_X = \sum_{i,j} \frac{(n_{ij} - e_{ij})^2}{e_{ij}}$ (écart entre la distribution conjointe de X, Y et la distribution correspondant à l'indépendance)
     - On n'applique pas le test d'indépendance mais on trie les variables X en ordre décroissant de leurs valeurs t_X, puis on élimine les variables X pour lesquelles ces valeurs sont trop faibles

Critères de qualité prédictive individuelle (2)

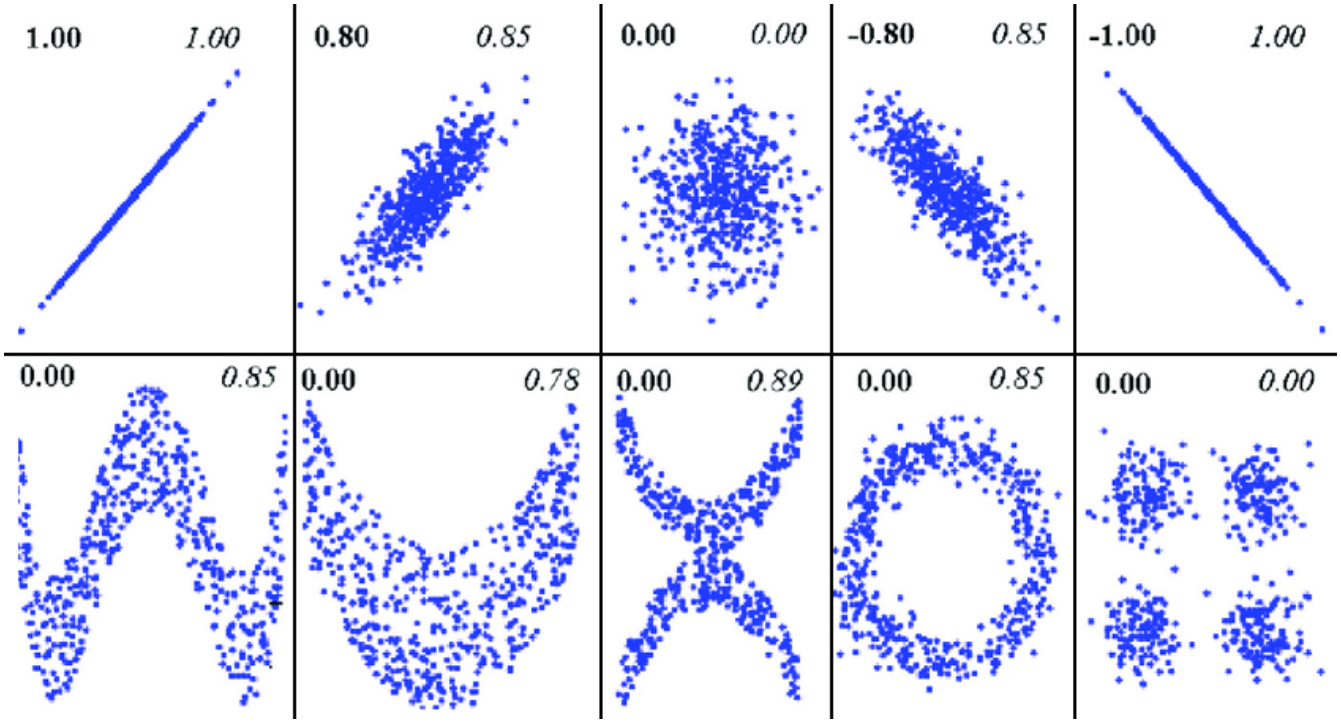


FIG. – Relations entre deux variables : bien représentées (en haut) ou mal représentées (en bas) par la correlation linéaire (illustration issue de [5]). La valeur de la corrélation linéaire est à chaque fois en haut à gauche, la valeur de l'information mutuelle en haut à droite (en italiques).

Critères de qualité prédictive individuelle (3)

3. Information mutuelle (qu'apporte la connaissance de X à la connaissance de Y ?) :
   - X et Y discrètes : $I(X;Y) = \sum_{x,y} P(x,y) \log \frac{P(x,y)}{P(x)P(y)}$
   - X et Y continues : $I(X;Y) = \int_{\mathbb{R}} \int_{\mathbb{R}} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} dxdy$
   - $I(X;Y) = 0 \Leftrightarrow X$ et Y sont indépendantes

$\rightarrow$ On garde les k variables d'entrée qui sont individuellement les plus pertinentes, ou celles dont la pertinence est supérieure à un seuil, ou à la moyenne, ou on garde les k premiers déciles, etc.

- Insuffisance de cette approche de sélection de $k \ll m$ variables parmi m :
  - Les k variables sont individuellement les plus « explicatives »
  - mais souvent redondantes : certaines apportent $\sim$ la même information que d'autres...

  $\rightarrow$ des variables individuellement moins « explicatives » mais plus complémentaires aux autres ne sont pas sélectionnées $\Rightarrow$ il y a un potentiel d'amélioration !

Qualité prédictive *de groupe*

- Caractériser un groupe de variables d'entrée par rapport à leur capacité à expliquer la variable de sortie
  - Comparer deux groupes quelconques : espace de recherche trop large...
  - → Comparer un groupe au même groupe avec une variable en plus pour caractériser l'apport de cette variable

- Suivant l'approche :
  1. *Wrapper* :
     - Avec chaque groupe de variables candidat on développe un modèle décisionnel
     - → On choisit le groupe pour lequel la performance du modèle décisionnel est meilleure
  2. Filtrage : extension au groupe de critères de qualité prédictive individuelle
     - Corrélation linéaire avec la variable de sortie : valeur moyenne entre variables du groupe
     - Information mutuelle avec la variable de sortie : valeur moyenne entre variables du groupe, ou extension à plusieurs variables de l'information mutuelle :

$$I(X^{(m)}; Y) = \int_{\mathbb{R}} \cdots \int_{\mathbb{R}} p(x_1, \ldots, x_m, y) \log \frac{p(x_1, \ldots, x_m, y)}{p(x_1, \ldots, x_m) p(y)} dx_1 \cdots dx_m dy \qquad (1)$$

($X^{(m)}$ étant un ensemble de m variables d'entrée)

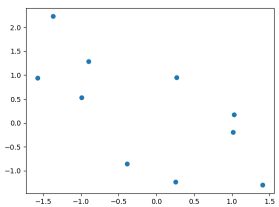
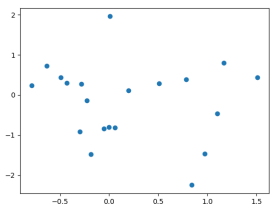
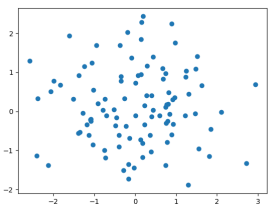
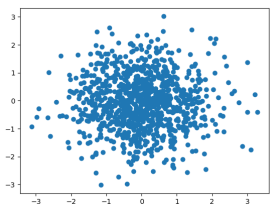
Qualité prédictive de groupe *et* réduction de la redondance

- Constats lors de l'extension à un groupe de variables de critères de qualité prédictive individuelle :
  1. Moyenner des corrélations, ou des informations mutuelles entre variables deux à deux, fait perdre en sélectivité

  → faibles écarts de qualité prédictive entre groupes de variables

  ⇒ Sélection d'un nombre élevé de variables, entre lesquelles il y a de la redondance

  2. Calcul peu fiable de l'information mutuelle avec beaucoup de variables car nombre insuffisant d'observations (devrait croître exponentiellement avec nombre de variables)
  Exemple simple unidimensionnel : X, Y normales et indépendantes :

n :	10	20	100	1000
				
$I(X;Y)$:	0.07	0.025	0.0135	0.002

→ augmentation artificielle de l'information mutuelle

→ des variables non pertinentes semblent pertinentes et sont sélectionnées

⇒ Sélection d'un nombre excessif de variables, entre lesquelles il y a de la redondance

Qualité prédictive de groupe *et* réduction de la redondance : mRMR

- Une solution est apportée par *minimum Redundancy and Maximum Relevance* (mRMR, [3])
- Objectifs de mRMR :
  1. Augmenter la qualité prédictive mesurée par l'information mutuelle entre les variables explicatives sélectionnées et la variable de sortie : $D = \frac{1}{m} \sum_i I(X_i; Y)$
  2. Tout en réduisant la redondance des variables sélectionnées, mesurée par l'information mutuelle moyenne entre ces variables : $R = \frac{1}{m^2} \sum_{i,j} I(X_i; X_j)$

→ mRMR = sélection incrémentale avec comme critère

- *Mutual information difference* (MID) : $\max_i(D - R)$, ou
- *Mutual information quotient* (MIQ) : $\max_i \frac{D}{R}$

- [3] montre que mRMR est équivalent à l'utilisation de l'information mutuelle (1) sans l'inconvénient du calcul peu fiable sur un petit échantillon multidimensionnel
- Principe similaire : MIMR (*Maximum Information and Minimum Redundancy* [1]), MRMR (*Maximum Relevance and Minimum Redundancy* [6])...

Conclusion

- La sélection de variables peut présenter des intérêts multiples
- L'approche filtrage est moins coûteuse mais potentiellement moins performante que l'approche *wrapper*
- Une sélection de variables est intégrée à certaines méthodes de modélisation prédictive (à découvrir dans RCP209)
- Combiner optimisation de la qualité prédictive et réduction de la redondance dans une approche de filtrage peut constituer un bon compromis entre coût et performance

Références I

[1] C. Li, V. P. Singh, and A. K. Mishra.

Entropy theory-based criterion for hydrometric network evaluation and design : Maximum information minimum redundancy.

Water Resources Research, 48(5), 2012.

[2] L. Molina, L. Belanche, and A. Nebot.

Feature selection algorithms : a survey and experimental evaluation.

In *2002 IEEE International Conference on Data Mining, 2002. Proceedings.*, pages 306–313, 2002.

[3] H. Peng, F. Long, and C. Ding.

Feature selection based on mutual information : Criteria of max-dependency, max-relevance, and min-redundancy.

IEEE Trans. Pattern Anal. Mach. Intell., 27(8) :1226–1238, Aug. 2005.

[4] J. Tang, S. Alelyani, and H. Liu.

Feature selection for classification : A review.

In *Data Classification : Algorithms and Applications*, pages 37–64. 2014.

Références II

[5] T. Vu, A. Mishra, and G. Kumar.

Information entropy suggests stronger nonlinear associations between hydro-meteorological variables and enso.

Entropy, 20 :38, 01 2018.

[6] Z. Zhao, R. Anand, and M. Wang.

Maximum relevance and minimum redundancy feature selection methods for a marketing machine learning platform, 2019.