# Une image contenant guitare Description générée automatiquement

# On supprime les colonnes dont le pourcentage des données manquantes dépasse 7%.

L’idée sous-jacente

# La sélection des variables

La sélection de variables est devenue l'objet de nombreuses recherches et particulièrement dans les domaines d’applications ou de nombreuses variables sont disponibles (des dizaines à des centaines de milliers variables). Lorsque le nombre de variables augmente, le traitement de données fait face à un phénomène connu sous le terme « le fléau de la dimension ou la malédiction de la dimension ». Ce terme a été introduit par le mathématicien Richard Bellman en 1961. Il désigne divers problèmes qui ont lieu seulement dans les espaces de grande dimension.

Par ailleurs, Verleysen et François [1] précisent que les espaces de haute dimension présentent des propriétés géométriques surprenantes et contre-intuitives qui exercent une grande influence sur les performances des outils d’analyse de données. Ils considèrent que les normes euclidiennes et les noyaux gaussiens, qui sont couramment utilisés dans les modèles d’apprentissage automatique, deviennent inappropriés dans les espaces de haute dimension à cause « de la concentration du phénomène normatif ».

Parmi les solutions à ce phénomène est la réduction de la dimension. Dans notre projet, la réduction de la dimension, par l’élimination de certaines variables, présente un intérêt supplémentaire dans l’optimisation de l’emplacement des capteurs. Réduire le nombre de capteurs revient à identifier indirectement l’emplacement des capteurs qui maximisent les performances pour les taches finales ( prédiction de la consommation électrique, évaluation de nombre d’occupants dans le bâtiment ou la détection des ouvertures des fenêtres)

Cet article présente des mesures de distance et des noyaux alternatifs, ainsi que des méthodes géométriques pour diminuer la dimension de l’espace. La méthodologie est appliquée à un exemple typique de prédiction de séries chronologiques.

La qualité du modèle prédictif dépend directement du bon choix des variables et de leurs contenues. La sélection des variables ou de fonctionnalités est le processus qui permet de réduire le nombre de variables d’entrée lors du développement d’un modèle d’apprentissage automatique. Il est essentiel à la fois pour réduire le coût de calcul et, dans certains cas, pour améliorer les performances du modèle.

Ce processus nous intéresse

L'objectif de la sélection de variables est triple : améliorer les performances de prédiction des modèles, fournir des prédicteurs plus pertinents, et fournir une meilleure compréhension du processus sous-jacent qui a généré les données.

En effet, la sélection des variables a pour objet de choisir les caractéristiques les plus significatives. En d’autres termes, choisir les variables qui contiennent le plus d’information et d’éliminer les variables superflues ou redondantes.

[Le fléau de la dimension : techniques de sélection de variables - Quantmetry](https://www.quantmetry.com/blog/le-fleau-de-la-dimension-techniques-de-selection-de-variables/)

[Microsoft Word - Thèse Lyon 2\_Legrand.doc (univ-lyon2.fr)](http://theses.univ-lyon2.fr/documents/lyon2/2004/legrand_g/pdfAmont/legrand_g_chapitre02.pdf)

Dans cette partie, nous aborderons les algorithmes de sélection des variables. Il est à noter que la sélection des variables s’impose lorsqu’il s’agit d’appliquer des algorithmes de machine Learning en présence d’une grande quantité de variables. Elle présente également un intérêt dans la détermination des meilleurs emplacements des capteurs ainsi que leur nombre pour réaliser une tâche de prédiction donnée (classification, régression ou prévision).

Pourquoi ?

Un nombre important de variable -> le fléau de la dimension ou malédiction de la dimension (curse of dimensionality) inventé en 1961 par le mathématicien Richard Bellman. Il désigne un problème qui surgit lorsqu’on tente d’analyser des données dans un espace de grande dimension.

L’apprentissage automatique

Des algorithmes de ML utilisent naturellement la sélection des variables pertinentes pour une prédiction efficace comme les arbres de décision. Cependant, la présence de plusieurs variables inutiles parmi les variables présentées au modèle augmente significativement le temps de traitement.

------------

L’objectif ici est de présenter rapidement des approches possibles pour éviter ce problème. On distingue de catégories d’approches celles qui agissent en de la modélisation et celles qui prennent en compte le modèle en question.

On appelle la première approche la sélection des variables :

Approche intégrée : la sélection des variables repose sur un critère propre du modèle d’apprentissage comme les arbres de décisions (détermination des nœuds et des règles de décision) ou pour la régression logistique (forward/backward) ..)

Approche enveloppe (wrapper)

Filtrage des prédicteurs :

------------------------------------------------------------------------

le modèle axé sur les données se prête mieux à la conception de contrôleurs en raison de sa faible complexité et pourrait servir de substitut à des modèles physiques très complexes avec une perte insignifiante de précision de prédiction pour de nombreuses applications. D’autre part, notre approche basée sur la physique est plus adaptée à la modélisation de bâtiments avec des granularités spatiales plus fines.

[Développement de méthodes et d'outils associant la modélisation numérique et des mesures in situ afin de concilier la qualité de l'air intérieur et l'efficacité énergétique (archives-ouvertes.fr)](https://tel.archives-ouvertes.fr/tel-02781482/document)

: Modèles d’émission de polluants utilisés dans la base PANDORE, d’après Abadie et Blondeau (2011). Page54

Le reste de cet article est organisé comme suit : dans la section [II](https://www.arxiv-vanity.com/papers/1603.05951/#S2), nous décrivons le banc d’essai et les données recueillies pour notre recherche. La section [III](https://www.arxiv-vanity.com/papers/1603.05951/#S3) présente le processus d’identification d’un modèle purement axé sur les données avec régression semi-paramétrique, suivie de la section [IV](https://www.arxiv-vanity.com/papers/1603.05951/#S4), qui détaille la procédure d’identification d’un modèle basé sur la physique. La section [V](https://www.arxiv-vanity.com/papers/1603.05951/#S5) comparera ensuite les performances du modèle piloté par les données et du modèle physique sous différentes mesures. Nous concluons à la section [VI](https://www.arxiv-vanity.com/papers/1603.05951/#S6) par un résumé de nos travaux actuels et futurs.

Je tiens à exprimer mes sincères remerciements et ma profonde reconnaissance à mes Directeurs et Co-Directeur de thèse, M. Karim LIMAM et M. Marc ABADIE, maîtres de conférences à l’Université de La Rochelle, pour le suivi de ma thèse, le partage de leur expertise et leur aide tout au long de mes travaux.

L’API Scikit-learn fournit la classe SelectFromModel pour extraire les meilleures fonctionnalités d’un jeu de données donné en fonction de l’importance des poids. SelectFromModel est un métaestimateur qui détermine l’importance du poids en la comparant à la valeur de seuil donnée.

Dans ce didacticiel, nous allons brièvement apprendre à sélectionner les meilleures fonctionnalités des données de régression à l’aide de SelectFromModel en Python. Le tutoriel couvre:

SelectFromModel pour les données

de régression

Liste du code source

Nous allons commencer par charger les bibliothèques et les fonctions requises.

SelectFromModel nécessite un estimateur et nous pouvons utiliser la classe AdaBoostRegressor à cette fin. Un modèle d’estimateur doit avoir des attributs pour fournir les index des données sélectionnées comme la fonction 'get\_support()'. Nous allons définir le modèle par valeur par défaut qui applique la méthode médiane pour définir la valeur seuil et ajuster le modèle sur les données x et y.

[Feature selection in machine learning using Lasso regression | Your Data Teacher](https://www.yourdatateacher.com/2021/05/05/feature-selection-in-machine-learning-using-lasso-regression/)