Statistiques pour Big Data

Thomas Chuffart

thomas.chuffart@parisnanterre.fr

Informations générales

- 24h CM
- Évaluation : Devoir sur table (60%) + projet (40%)

Le projet :

- Doit faire intervenir du machine learning ou des méthodes big data
- Doit être fait dans un notebook Jupyter
- Date de rendu : 01/12/2021

Overview

Introduction
- Machine learning, data science, AI et d'autres concepts
- Historique
- Apprentissage supervisé
- Apprentissage non supervisé

Régression linéaire pénalisée

Classification

Données textuelles

Objectif

Objectif du cours : construire des modèles qui peuvent prédire un outcome après un choix d'inputs :

- De nombreuses méthodes existent : SVM, random forest, ...
- aucune n'est magique. Elles nécessitent de bonnes données et une validation du modèle.
- L'apprentissage machine sans le jugement humain : ça ne fonctionne pas.

Disclaimer

- Ne JAMAIS faire confiance à quelqu'un qui dit avoir un algorithme génial à moins d'avoir vu les résultats de validation eg. diviser le sample en deux, lancer l'algo sur la 1ère partie, tester sur la deuxième partie.

Overview

Introduction
Machine learning, data science, AI et d'autres concepts
Historique
Apprentissage supervisé
Apprentissage non supervisé

Des définitions

- Machine Learning : Algorithmes qui apprennent via les données à disposition.
- Statistical learning : une branche des statistiques appliquées, basée sur des règles de programmation et des hypothèses
- Data Science : extraction d'information des données

Des définitions - ML

- Murphy (2012) : Machine learning is defined as a set of methods that can automatically detect patterns in data, and then use the uncovered patterns to predict future data, or to perform other kinds of decision making under uncertainty.
- Athey (2018) : Machine learning is a field that develops algorithms designed to be applied to datasets, with the main areas of focus being prediction (regression), classification, and clustering or grouping tasks.

Exemples

- Prévoir l'apparition d'un cancer en fonction des caractéristiques des patients
- Prévoir le prix d'une action ou d'un indice boursier pour les prochains mois (inputs : performance de la firme et données économiques)
- Identifier des nombres sur un code postal écrit manuellement sur une image numérisée
- Prévoir le prix de vente d'un logement selon ses caractéristiques et celles de la ville où il se trouve.

Exemple

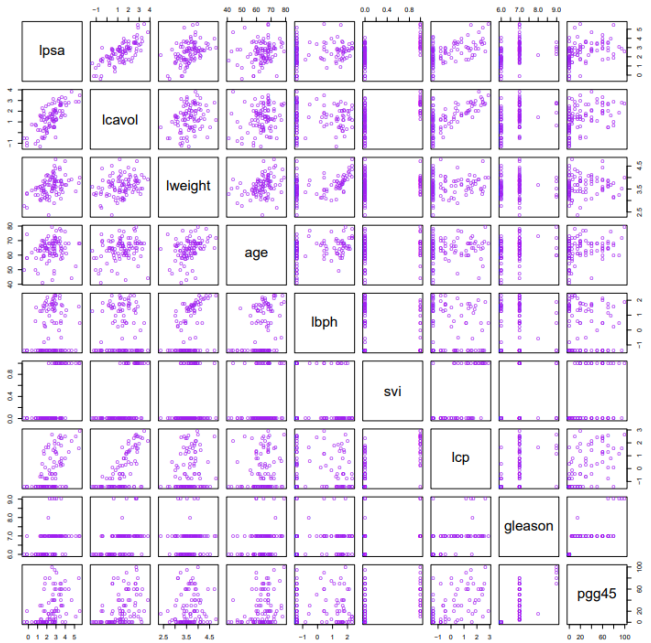


Figure – Data visualisation

Overview

Introduction
- Machine learning, data science, AI et d'autres concepts
- Historique
- Apprentissage supervisé
- Apprentissage non supervisé

Historique

In God we trust, all others bring data. – William Edwards Deming (1900-1993)

Des premières traces d'imaginaire autour des robots et de l'IA dans la Grèce antique :

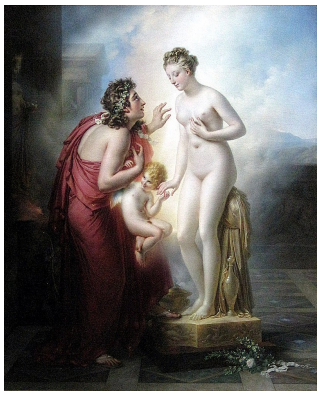


Figure – Pygmalion amoureux de sa statue, d'Anne-Louis Girodet par Jean-Pol Grandmont

- Le sculpteur Pygmalion est tombé amoureux de sa statue, Galatée, rendue vivante grâce à Aphrodite, la déesse de l'amour.
- Aristote - Autour 350 avant JC, il s'essaye à la logique formelle.

Avant le XXème siècle

- En 1642, Pascal invente la calculatrice
- En 1842, Ada Lovelace a construit le premier algorithme destiné à la machine analytique de Babbage.

Figure – Ada King, comtesse Lovelace, vers 1840 (aquarelle) - Crédits : Alfred Edward Chalon

Darmouth Workshop

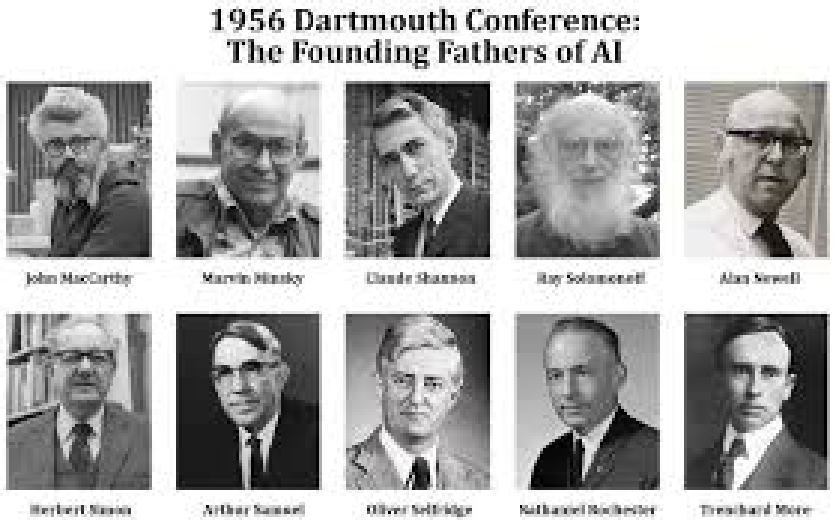


Figure – Participants du workshop

Turing (1950) : Can machines think ? à l'initiative du Dartmouth Summer Research Project on Artificial Intelligence en 1956

John Tukey (1915-2000)

- Chimiste puis mathématicien et statisticien à Princeton
- ”bit”, Exploratory Data Analysis (box plot), FFT algorithme ...
- The Future of Data Analysis (1962).

Figure – John Tukey, par Paul R. Halmos

John M. Chambers - Bell labs

Figure – John M. Chambers, Bell Labs Workshop, 2005

- connu pour avoir créé le langage de programmation S.
- Lauréat du prix ACM du logiciel de 1998

Bibliographie

- An Introduction to Statistical Learning, Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani - Lien
- The Elements of Statistical Learning, Trevor Hastie, Robert Tibshirani, Jerome Friedman - Lien
- Statistical Learning with Sparsity : The Lasso and Generalizations, Trevor Hastie, Robert Tibshirani, Martin Wainwright - Lien

Overview

Introduction
- Machine learning, data science, AI et d'autres concepts
- Historique
- Apprentissage supervisé
  - Présentation
  - Exemples
  - Estimation de f
  - Qualité de l'estimation
  - Variance et Biais, un trade-off
  - Le fléau de la dimension
- Apprentissage non supervisé

Apprentissage supervisé

L'objectif de l'apprentissage supervisé est d'apprendre une fonction de prévision à partir d'une paire d'inputs x et d'outputs y, $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}$ où :

- $\mathcal{D}$ est le training set
- n la taille du training set
- x_i est un vecteur : features, covariates, variables explicatives généralement stockés dans une matrice $n \times p$
- Y_i est la variable de réponse

Apprentissage supervisé

Deux types d'apprentissage supervisé : la classification et la régression.

- il est possible de comparer la prédiction $\hat{y}_i$ et y_i afin de calculer l'erreur de prédiction.
- quand y_i est une variable catégorique, on est dans un problème de classification.
- quand y_i est un scalaire réel, on est dans un problème de régression.

Philosophie

- Il est important de comprendre les idées des diverses techniques, afin de savoir comment et quand les utiliser.
- Il faut d'abord comprendre les méthodes les plus simples, afin de comprendre les plus sophistiqués.
- Il est important d'évaluer avec précision la performance d'une méthode, pour savoir si elle fonctionne bien ou mal (les méthodes plus simples sont souvent aussi performantes que les plus sophistiquées !)
- Il s'agit d'un domaine de recherche passionnant, ayant d'importantes applications en science, industrie et finance.
- L'apprentissage statistique est un ingrédient fondamental de la formation d'un data scientist moderne.

AS vs ML

- Le ML est apparu comme un sous-domaine de l'artificiel Intelligence.
- L'AS est apparu comme un sous-domaine de la statistique.
- Il y a beaucoup de chevauchements — les deux domaines se concentrent sur des problèmes supervisés et non-supervisés :
  - ML met davantage l'accent sur des applications grande échelle et la précision des prévisions.
  - L'AS met l'accent sur les modèles et leurs interprétabilité, précision et incertitude.
- La distinction est devenue de plus en plus floue,

Code postaux

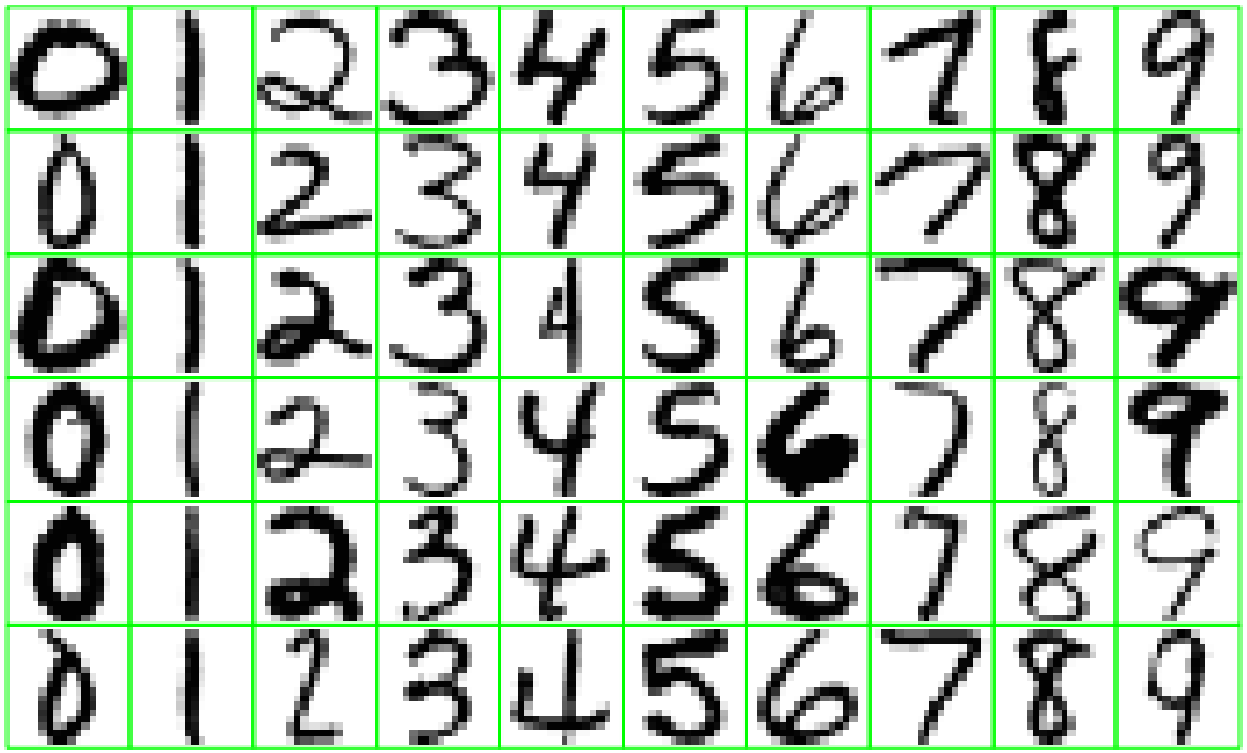


Figure – Numéros scannés sur des codes postaux

Données textuelles

Verbatim of the remarks made by Mario Draghi Speech by Mario Draghi, President of the European Central Bank at the Global Investment Conference in London 26 July 2012
I asked myself what sort of message I want to give to you ; [...]
The first message I would like to send, is that the euro is much, much stronger, the euro area is much, much stronger than people acknowledge today. Not only if you look over the last 10 years but also if you look at it now, you see that as far as inflation, employment, productivity, the euro area has done either like or better than US or Japan. [...]
But there is another message I want to tell you. Within our mandate, the ECB is ready to do whatever it takes to preserve the euro. And believe me, it will be enough. [...]

Spam ou ham

Objectif : classer les emails dans les spam ou non.

- ham : y = 0, l'email n'est pas pas désiré
- span : y = 1, l'email n'est pas désiré

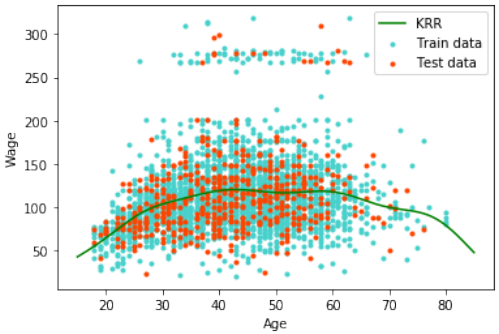
On possède un training set des n emails pour lesquels on sait si ce sont des spam ou non.

Spam ou ham

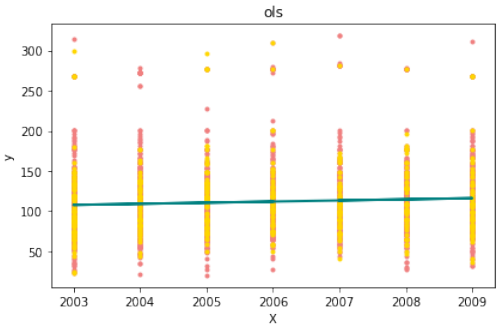
Une méthode possible consiste à crérer une matrice de termes avec la fréquence de ces derniers dans la collection de texte.

- x_{ij} correspond à la fréquence du mot j dans le document i.
- Les termes buy, cheap, jackpot, win apparaissent plus fréquemment dans les spam.

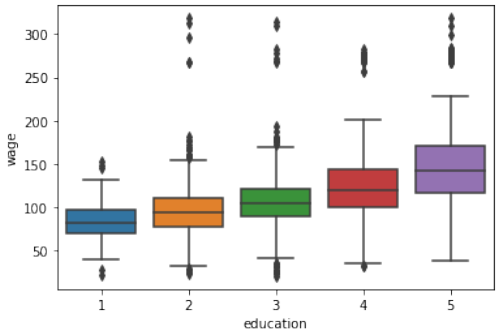
More money ? Get educated



(a) Relation Non-linéaire



(b) Relation linéaire



(c) Boxplot

Figure – Salaire, Age, Éducation : réplication

Espérance conditionnelle et fonction de perte

Soit $X \in \mathbb{R}^p$ et $Y \in \mathbb{R}$ avec une distribution jointe $Pr(X, Y)$. On cherche une fonction $f(X)$ afin de prédire Y à partir des inputs X.

Definition
Fonction de perte quadratique

$$L(e) = L(\text{observé} - \text{prédite})$$

avec e, l'erreur de prévision et

- $L(0) = 0$
- $L(e)$ est une fonction continue qui croit avec $|e|$

On s'intéresse à la fonction d'erreur quadratique :
$L(e) = (Y - f(X))^2$

Espérance conditionnelle et fonction de perte

Comment choisir la fonction f ? Minimiser l'erreur de prédiction espérée :

$$\mathrm{EPE(f)} = \mathrm{E(Y - f(X))^2} = \mathrm{E_X E_{Y|X}([Y - f(X)]^2 \, |X)}$$

La fonction f qui minimise l'EPE satisfait :

$$\mathrm{f(x)} = \mathrm{argmin_c E_{Y|X}([Y - c]^2 | X = x)}$$

La solution est tout simplement la fonction de régression E(Y|X), soit la moyenne conditionnelle.

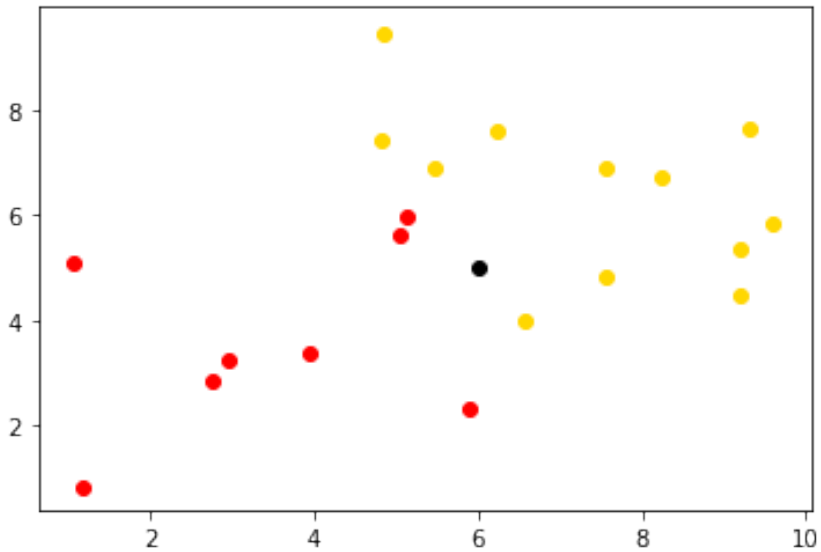
Espérance conditionnelle et fonction de perte

Avec un output G qui est une variable catégorielle,

- la fonction de perte est définie par une matrice $K \times K$ notée L, avec $K = \text{card}(\mathcal{G})$
- $L = 0$ sur la diagonale, non-négatif ailleurs. $L(k, l)$ est le cout d'avoir classé un observation dans G_l au lieu de G_k.
- $EPE = E[L(G, \hat{G}(X))]$
- $\hat{G}(x) = \text{argmin}_{g \in \mathcal{G}} \sum_{k=1}^{K} L(\mathcal{G}_k, g) Pr(\mathcal{G}_k | X = x)$

K-nn : K-nearest neighbor

L'input consiste au k plus proche exemples du training sample. L'output dépend si y est une variable qualitative ou quantitative.



$$\hat{f}(x) = \text{Ave}(y_i | x_i \in N_k(x))$$

Qualité de l'estimation

Généralement, lorsque l'on fait de l'apprentissage supervisé, on utilise l'erreur quadratique moyenne :

$$\mathrm{MSE} = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \hat{f}(x_i) \right)^2$$

Méthode

L'échantillon est divisé en sous-échantillon :

- Training sample : $\{(x_1, y_1), \ldots, (x_n, y_n)\}$. Cela donne $\hat{f}$
- Evaluation sample : on utilise $\hat{f}$ sur de nouvelles données afin de sélectionner les hyperparamètres
- Testing sample : la précision est alors calculée

Cela nous dit comment le modèle va réagir quand il sera confronté à de nouvelles données

Cross validation

Quand il y a peu d'observation on utilise une méthode appelée cross validation :

- k-fold cross validation
- repeated cross-validation
- leave-one out crosse validation

k-fold Cross validation : soit un training sample avec n observation $\{(x_1, y_1), \ldots, (x_n, y_n)\}$, que l'on peut diviser en k sous-échantillons. On entraine le modèle sur $k-1$ folds et on évalue sur le kth.

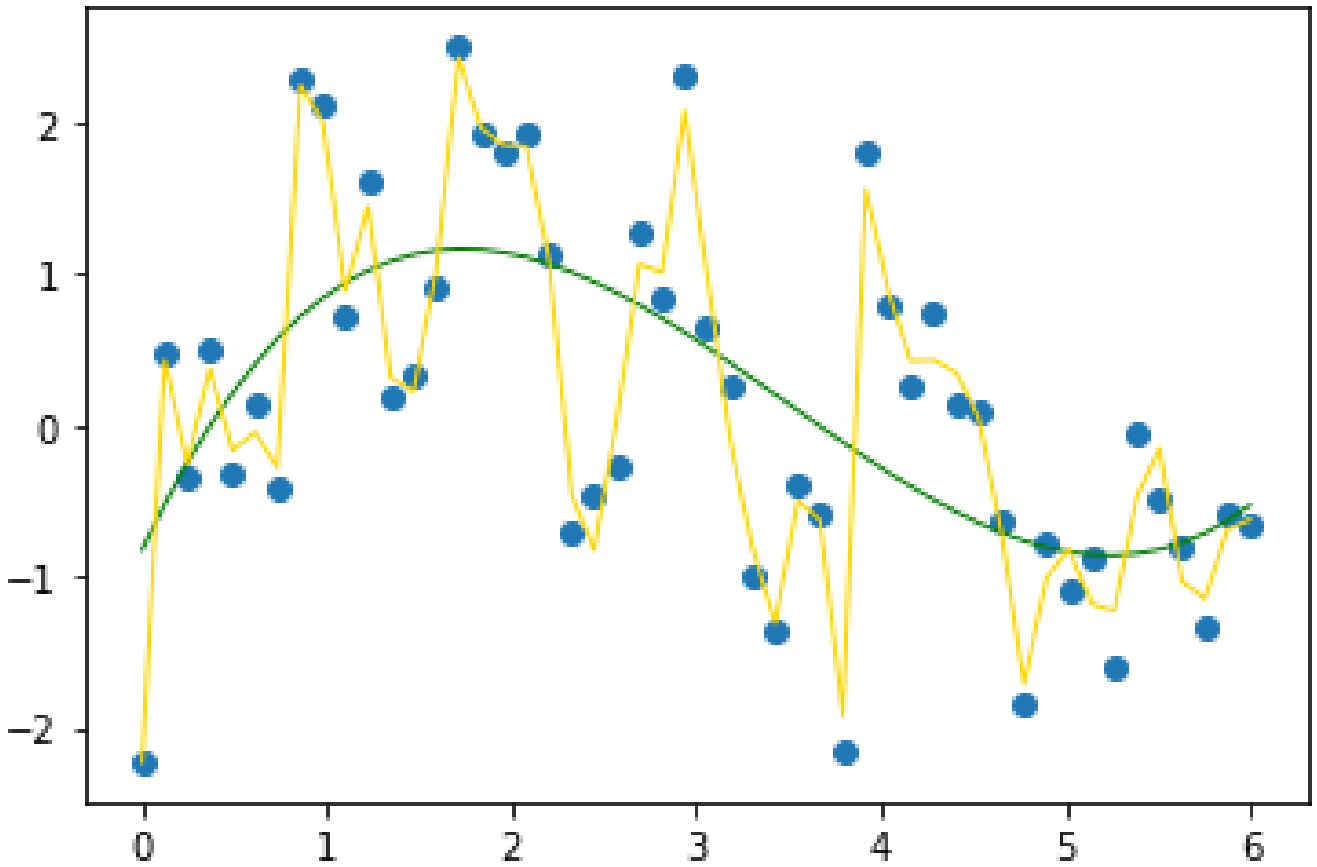
Le trade-off entre la variance et le biais

Quand on estime un ou des outputs, on cherche à faire coïncider l'estimation aux vraies valeurs. L'erreur de prédiction peut être décomposée.

$$\mathbb{E}\left[\left(y_0 - \hat{f}(x_0)\right)^2\right] = \mathbb{E}\left[y_0^2 + \hat{f}(x_0)^2 - 2y_0\hat{f}(x_0)\right]$$
$$= \mathbb{V}\text{ar}\left(\hat{f}(x_0)\right) + \text{Biais}\left[\hat{f}(x_0)\right]^2 + \mathbb{V}\text{ar}(\varepsilon)$$

- La variance représente de combien va changer $\hat{f}$ si on l'estime sur un différent training sample.
- Le biais représente l'écart entre la prévision et les vraies valeurs.

Le trade-off entre la variance et le biais



Le tradeoff

- Il faut choisir la méthode la plus flexible pour avoir un faible biais
- Il faut choisir la moins flexible pour avoir une faible variance.

Le fléau de la dimension

Berk (2008) :

> In short, higher dimensional data can be very useful when there are more associations in the data that can be exploited. But at least ideally, a large p comes with a large N . If not, what may look like a blessing can actually be a curse.

Overview

Introduction
- Machine learning, data science, AI et d'autres concepts
- Historique
- Apprentissage supervisé
- Apprentissage non supervisé
  - Présentation
  - Exercices

Apprentissage non supervisé

Apprentissage sans professeur :

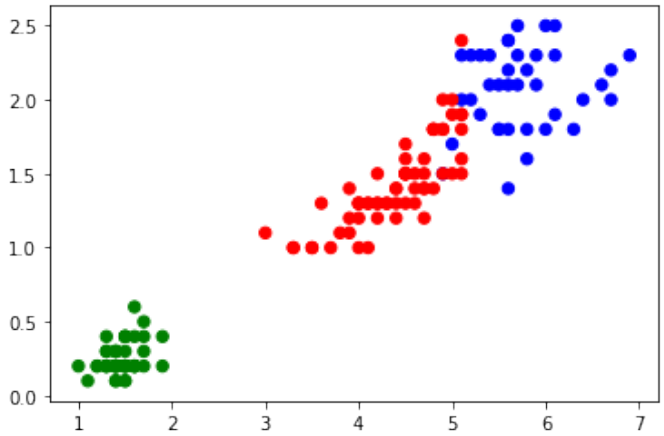
- Ensemble de N observations $(x_1, x_2, \ldots, x_N)$ d'un échantillon aléatoire avec une densité jointe $\Pr(X)$.
- On cherche alors à inférer cette densité sans l'aide d'un superviseur.
- Lien avec l'économétrie non-paramétrique.

. Comparé à l'apprentissage supervisé :

- Subjectivité, pas de data label
- Objectif différent qu'une simple prévision, pas de variable Y
- Quelle fonction de coût utiliser ?

Clustering

(a) Fleur d'iris



(b) Clustering d'Iris

Figure – Un peu de nature

Le nombre de cluster K est une hypothèse. On estime $P(K|\mathcal{D})$. Choisir la bonne valeur est l'étape de la sélection du modèle. On cherche esnuite à déterminer quel fleur appartient à quel cluster.

Facteurs latents

Avec le Big Data, p a énomément augmenté (le nombre de variables). Comment réduire ce nombre ?

- Projection sur un sous-espace d'une plus petit dimension (ACP)
- Facteurs latents

Conclusion

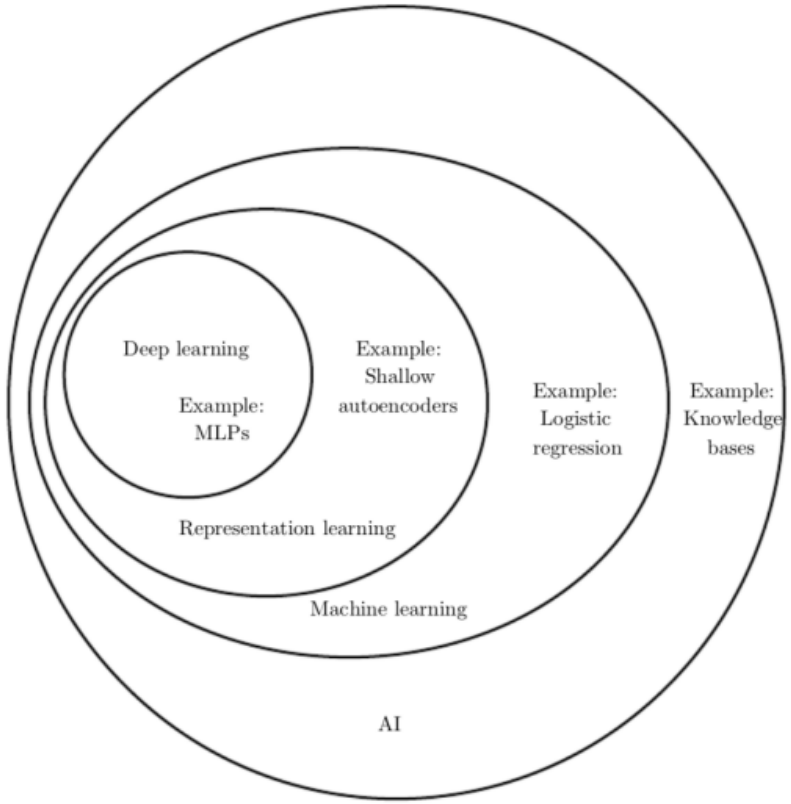
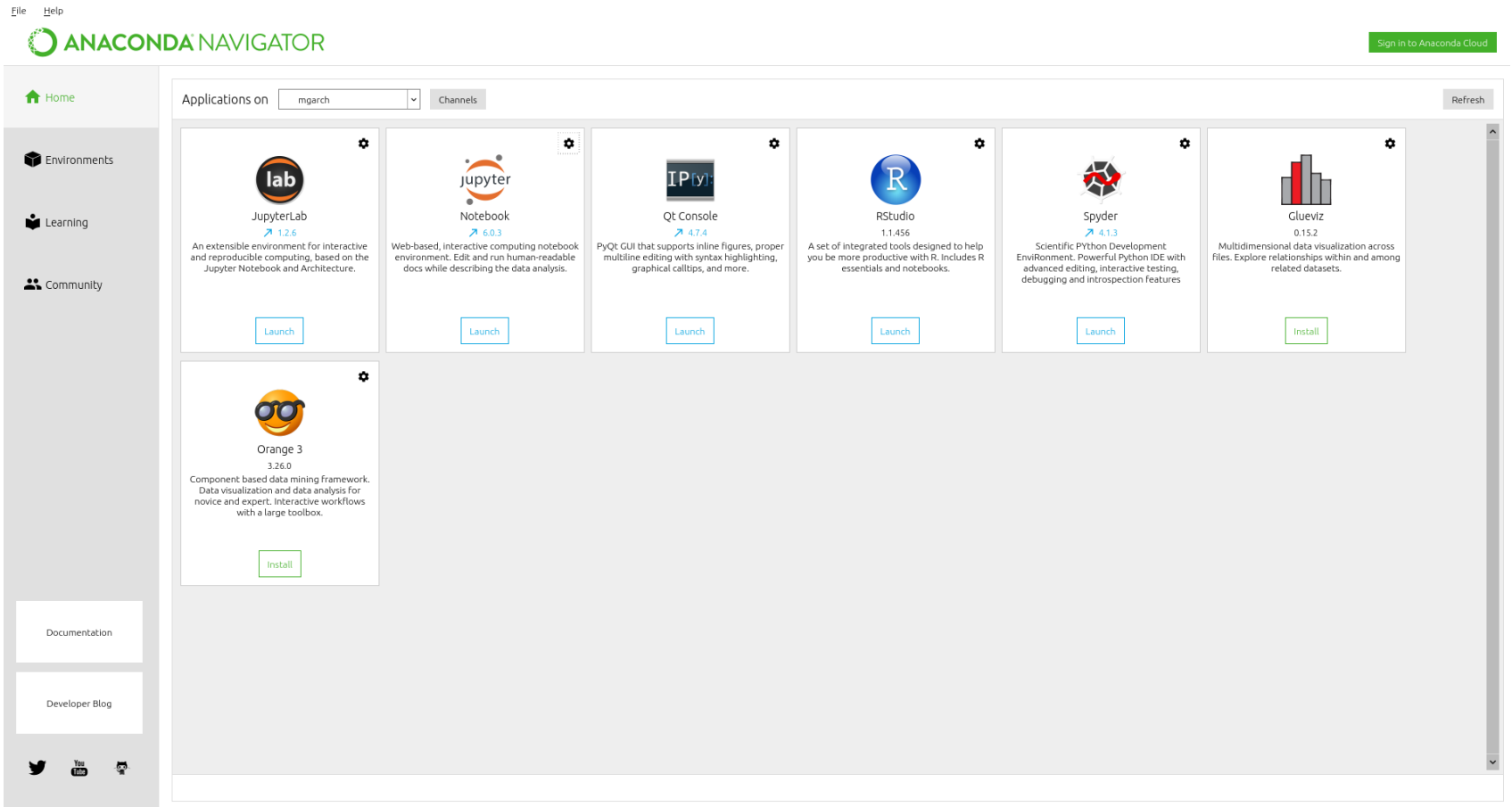


Figure – Source : Goodfellow et al. (2016).

Python avec Anaconda



Exercice 1

Comparer le modèle de classification de régression linéaire et du k-nearest neighbor sur les données des codes postaux. In particular avec $\{\mathrm{k} = 1, 3, 5, 7\mathrm{et}15\}$. Calculez l'erreur d'entrainement et l'erreur de test. Les données sont disponibles à l'adresse `https://web.stanford.edu/~hastie/ElemStatLearn/`.

Overview

Introduction

Régression linéaire pénalisée
- La régression linéaire
- La régression Ridge
- La régression Lasso

Classification

Données textuelles

Overview

Régression linéaire pénalisée
La régression linéaire
Introduction
Le modèle
La régression Ridge
La régression Lasso

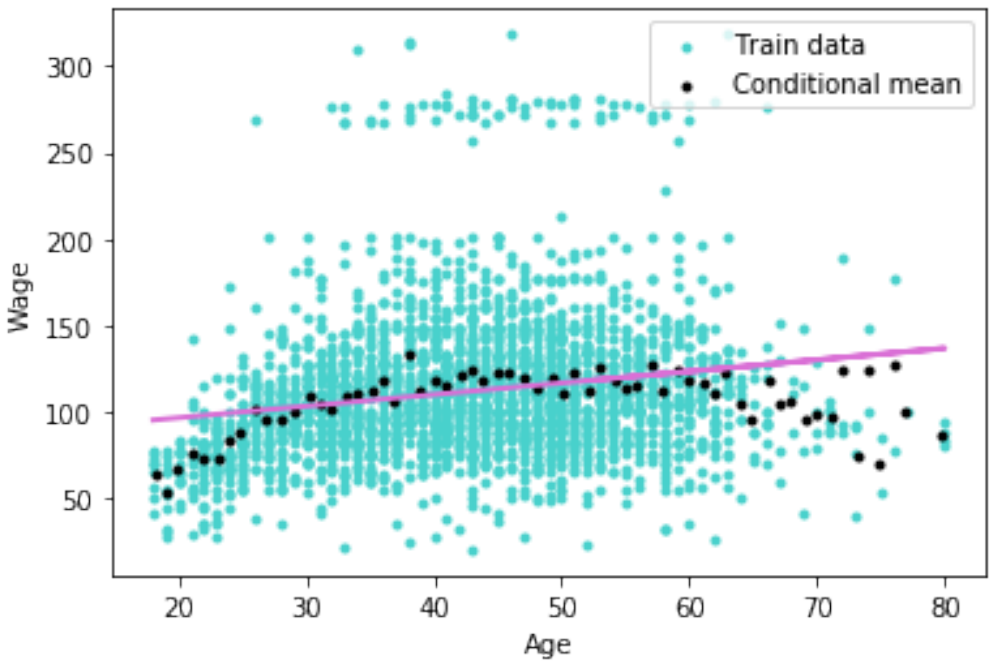
Deux références

- Berk (2008). Statistical learning from a regression perspective
- James et al. (2013). An introduction to statistical learning

Petits rappels

- Les modèles de régression ne font pas partis de la classe des modèle dit génératif
- L'objectif est de comprendre "as far as possible with the available data how the conditional distribution of some response y varies across subpopulations determined by the possible values of the predictors" (Cook and Weisberg).
- Il n'y a pas d'affirmation causale.

Exemple


Train data
Conditional mean
Wage
Age
300
250
200
150
100
50
20
30
40
50
60
70
80

Quelques enseignements

- une analyse de régression ne nécessite pas automatiquement de résultats arithmétiques.
- des distributions conditionnelles entières peuvent être examinées.
- le choix des variables est un sujet ou une décision politique.
- lorsque l'inférence statistique est utilisée, sa validité dépendra fondamentalement de la manière dont les données ont été générées.
- Level 1 : description des données
- Level 2 : Inférence statistique
- Level 3 : Inférence causale

Le modèle linéaire et les moindres carrés

Soit un vector d'input $X' = (X_1, X_2, \ldots, X_p)$, on prédit l'ouput Y via le modèle

$$\hat{Y} = \hat{\beta}_0 + \sum_{j=1}^{p} X_j \hat{\beta}_j$$

ou

$$Y = X\hat{\beta}$$

X est de taille $N \times (p+1)$, Y est de taille $N \times 1$.

Le modèle peut aussi s'écrire :

$$\mathrm{y_i} = \beta_0 + \beta_1 \mathrm{x_{1i}} + \cdots + \beta_\mathrm{p} \mathrm{x_{pi}} + \varepsilon_\mathrm{i},\ \varepsilon_\mathrm{i} \sim \mathrm{NIID}(0, \sigma^2)$$

La nature définit les valeurs de X, multiplie chaque input par son coefficient, fait la somme puis ajoute une perturbation.

- quand les conditions de premier ordre sont remplies, l'estimation OLS est non biaisée.

Estimation

Soit $e_i = y_i - \hat{y}_i$ le i-ème résidu. On définit la sommes des résidus au carré :

$$\text{RSS} = \sum_{i=1}^{n} e_i^2$$

On cherche donc à minimiser cette somme.

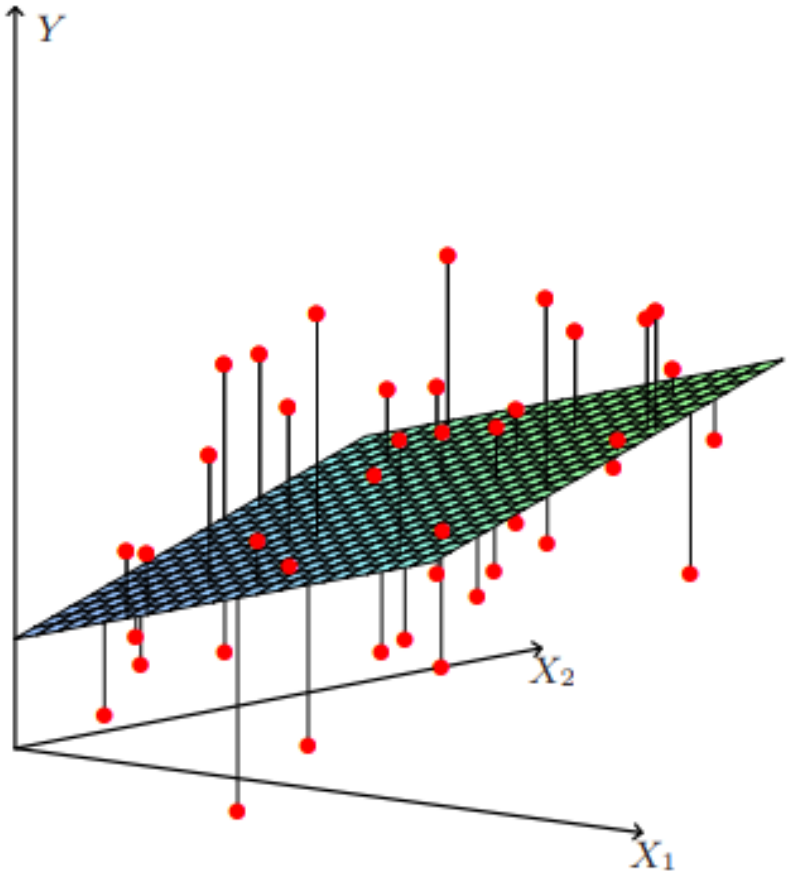


Figure – Plan minimisant la RSS, X $\in \mathbb{R}^2$. Source : ESL 2ème édition

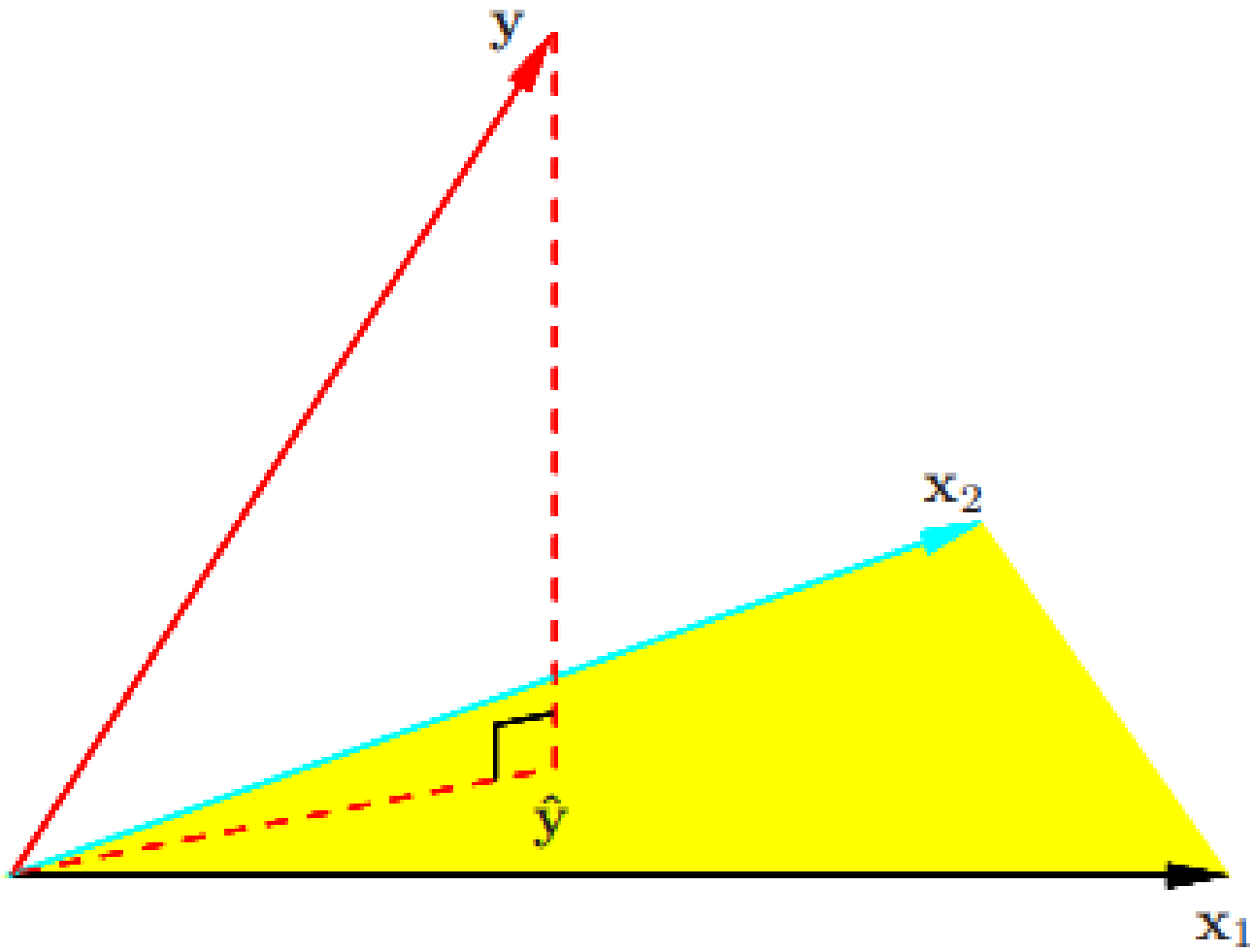


Figure – Géométrie de dimension N des OLS avec p = 2. Le vecteur y est une projection orthogonal sur l'hyperplan généré par x_1 et x_2.

Propriétés

Les estimateurs $\hat{\beta}_0$ et $\hat{\beta}_1$ sont des estimations ponctuels. Via OLS, ils sont :

- sans biais $\mathbb{E}(\hat{\beta}_0) = \beta_0$, $\mathbb{E}(\hat{\beta}_1) = \beta_1$
- variance minimale
- convergents

Estimation sous forme matricielle

Sous forme matricielle, on a :

$$\mathrm{y} = \mathrm{X}\beta + \varepsilon$$

Soit la résolution suivante :

$$\min \mathrm{RSS}_\beta = (\mathrm{y} - \mathrm{X}\beta)' (\mathrm{y} - \mathrm{X}\beta)$$

donc

$$\frac{\partial \mathrm{RSS}}{\partial \beta} = -2\mathrm{X}' (\mathrm{y} - \mathrm{X}\beta) = 0, \ \frac{\partial^2 \mathrm{RSS}}{\beta \beta'} = 2(\mathrm{X}'\mathrm{X}).$$

$(\mathrm{X}'\mathrm{X})^{-1}$ doit donc exister.

Propriétés suite

On a :

- $\mathbb{V}\mathrm{ar}(\hat{\beta}) = (\mathrm{X}'\mathrm{X})^{-1}\sigma^2$
- $\hat{\beta} \sim \mathcal{N}(\beta, (\mathrm{X}'\mathrm{X})^{-1}\sigma^2)$
- $\frac{\sum_{\mathrm{i}=1}^{\mathrm{n}} \varepsilon_{\mathrm{i}}^2}{\sigma^2} \sim \chi_{\mathrm{n}-\mathrm{p}}$
- on peut donc facilement créer des statistiques de test.

Theorem (Gausse-Markov)

L'estimateur OLS est celui qui a la plus petite variance parmi les estimateurs non biaisés.

Overview

Régression linéaire pénalisée
- La régression linéaire
- La régression Ridge
- La régression Lasso

Principe

On oublie le principe du sans-biais.

- Précision de la prévision : les OLS donnent une estimation sans biais mais généralement avec une large variance.
- Avec beaucoup de variables explicatives, on veut généralement sélectionner un sous-ensemble qui a les effets les plus important.

La régression Ridge (et Lasso) utilise la méthode de contraction des coefficients (shrinkage coefficients) qui introduit un petit biais mais diminue drastiquement la variance.

Équation de régression

$$\hat{\beta}^{r} = \text{argmin}_{\beta_0,\beta} \left\{ \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 \right\}$$

$$\text{subject to} \sum_{j=1}^{p} \beta_j^2 \leq t$$

La contrainte peut se réécrire $||\beta_j||_2^2 \leq t$ (norme l_2) et le problème :

$$\hat{\beta}^{r} = \text{argmin}_{\beta_0,\beta} \left\{ \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 \right\} + \lambda \sum_{j=1}^{p} \beta_j^2$$

Intuition

- il y un lien entre λ et t.
- lorsque les variables sont corrélées, un effet positif d'une variable peut-être annulé par un large effet négatif sur une variable corrélée.
- β_0 n'est pas pénalisé !
- En pratique, on "standardise" X si elles ont différentes unités de mesure. De fait, on a $\hat{\beta}_0 = \bar{y} - \sum_{j=1}^{p} \bar{x}_j \hat{\beta}_j$

Exercice

Soit le problème de minimisation Ridge défini précédemment. Montrer que ce problème est équivalent à

$$\hat{\beta}^{c} = \text{argmin}_{\beta}^{c} \left\{ \sum_{i=1}^{N} \left[y_i - \beta_0^c - \sum_{j=1}^{p} (x_{ij} - \bar{x}_j) \beta_j^c \right]^2 + \lambda \sum_{j=1}^{p} \beta_j^{c2} \right\}$$

Forme matricielle

Le programme de minimisation d'une régression Ridge peut s'écrire aussi s'écrire sous forme matricielle :

$$\mathrm{RSS}(\lambda) = (\mathrm{y} - \mathrm{X}\beta)' (\mathrm{y} - \mathrm{X}\beta) + \lambda\beta'\beta$$

avec comme solution

$$\hat{\beta}^{\mathrm{r}} = \left(\mathrm{X}'\mathrm{X} + \lambda\mathrm{I}\right)^{-1} \mathrm{X}'\mathrm{y}$$

La solution ajoute une constante positive à la diagonale de $\mathrm{X}'\mathrm{X}$ avant l'inversion. On retire le problème de non singularité.

SVD

Décomposition en valeurs singulières (SVD) de la matrice X de taille $N \times p$: $X = UDV'$.

- U est de taille $N \times p$, orthogonale. La combinaison linéaire des colonnes de U engendre l'espace colonne de X.
- V est de taille $p \times p$. La combinaison linéaire des colonnes de V engendre l'espace ligne de X.
- D est une matrice diagonale $p \times p$ avec éléments $d_1 \geq d_2 \geq \times \geq d_p$, les valeurs singulières de X (racine des valeurs propres).

SVD

$$\begin{aligned} \mathrm{X}'\mathrm{X} &= \mathrm{V}\mathrm{D}\mathrm{U}'\mathrm{U}\mathrm{D}\mathrm{V}' \\ &= \mathrm{V}\mathrm{D}^2\mathrm{V}' \end{aligned}$$

En remplaçant dans l'estimateur de la régression Ridge, on a :

$$\begin{aligned} \hat{\beta}^{\mathrm{r}} &= \left(\mathrm{X}'\mathrm{X} + \lambda \mathrm{I}\right)^{-1} \mathrm{X}'\mathrm{y} \\ &= \left(\mathrm{V}\mathrm{D}^2\mathrm{V}' + \lambda \mathrm{V}\mathrm{V}'\right)^{-1} \mathrm{V}\mathrm{D}\mathrm{U}'\mathrm{y} \\ &= \left(\mathrm{V}\left(\mathrm{D}^2 + \lambda \mathrm{I}\right)\mathrm{V}'\right)^{-1} \mathrm{V}\mathrm{D}\mathrm{U}'\mathrm{y} \end{aligned}$$

SVD

De plus,

$$\begin{aligned}\widehat{\mathrm{y}}^{\mathrm{r}} &= \mathrm{X}\hat{\beta}^{\mathrm{r}} \\ &= \mathrm{UD}\left(\mathrm{D}^2 + \lambda \mathrm{I}\right)^{-1} \mathrm{DU}'\mathrm{y}\end{aligned}$$

ou encore

$$\widehat{\mathrm{y}}^{\mathrm{r}} = \sum_{\mathrm{j}=1}^{\mathrm{p}} \mathrm{u}_{\mathrm{j}} \frac{\mathrm{d}_{\mathrm{j}}^2}{\mathrm{d}_{\mathrm{j}}^2 + \lambda} \mathrm{u}_{\mathrm{j}}'\mathrm{y}$$

SVD

Soit $S = X'X/N = VD^2V'/N$ la matrice variance-covariance.

- Les colonnes de V sont des vecteurs propres aussi appelés composants principales de X

Overview

Régression linéaire pénalisée
- La régression linéaire
- La régression Ridge
- La régression Lasso
  - Le LASSO en pratique

Le LASSO - Équation de régression

$$\hat{\beta}^{r} = \text{argmin}_{\beta_0,\beta} \left\{ \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 \right\}$$

$$\text{subject to} \sum_{j=1}^{p} |\beta_j| \leq t$$

La contrainte peut se réécrire $||\beta_j||_1 \leq t$ (norme l_1) et le modèle :

$$y_i = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j + \lambda \sum_{j=1}^{p} |\beta_j|$$

La base de donnée

FRED-MD Database : Lien vers la base

FRED-MD and FRED-QD are large macroeconomic databases designed for the empirical analysis of "big data."

Dictionnaire de la base

Appendix

The column TCODE denotes the following data transformation for a series x: (1) no transformation; (2) Δx_t; (3) $\Delta^2 x_t$; (4) $log(x_t)$; (5) $\Delta \log(x_t)$; (6) $\Delta^2 \log(x_t)$; (7) $\Delta(x_t/x_{t-1} - 1.0)$. The FRED column gives mnemonics in FRED followed by a short description. The comparable series in Global Insight is given in the column GSI.

Some series require adjustments to the raw data available in FRED. We tag these variables with an asterisk to indicate that they been adjusted and thus differ from the series from the source. A summary of the adjustments is detailed in the paper https://research.stlouisfed.org/wp/2015/2015- 012.pdf

Group 1: Output and income

	id	tcode	fred	description	gsi	gsi:description
1	1	5	RPI	Real Personal Income	M_14386177	PI
2	2	5	W875RX1	Real personal income ex transfer receipts	M_145256755	PI less transfers
3	6	5	INDPRO	IP Index	M_116460980	IP: total
4	7	5	IPFPNSS	IP: Final Products and Nonindustrial Supplies	M_116460981	IP: products
5	8	5	IPFINAL	IP: Final Products (Market Group)	M_116461268	IP: final prod
6	9	5	IPCONGD	IP: Consumer Goods	M_116460982	IP: cons gds
7	10	5	IPDCONGD	IP: Durable Consumer Goods	M_116460983	IP: cons dble
8	11	5	IPNCONGD	IP: Nondurable Consumer Goods	M_116460988	IP: cons nondble
9	12	5	IPBUSEQ	IP: Business Equipment	M_116460995	IP: bus eqpt
10	13	5	IPMAT	IP: Materials	M_116461002	IP: matls
11	14	5	IPDMAT	IP: Durable Materials	M_116461004	IP: dble matls
12	15	5	IPNMAT	IP: Nondurable Materials	M_116461008	IP: nondble matls
13	16	5	IPMANSICS	IP: Manufacturing (SIC)	M_116461013	IP: mfg
14	17	5	IPB51222s	IP: Residential Utilities	M_116461276	IP: res util
15	18	5	IPFUELS	IP: Fuels	M_116461275	IP: fuels
16	19	1	NAPMPI	ISM Manufacturing: Production Index	M_110157212	NAPM prodn
17	20	2	CUMFNS	Capacity Utilization: Manufacturing	M_116461602	Cap util

Objectifs

1. Prévoir l'inflation
2. Modélisation multivariée
3. Sélectionner les variables explicatives

On va utiliser un modèle VAR (Vector AutoRegressive) avec :

- des variables issues de la théorie économique (inflation, chômage, Taux d'intérêt, Taux de change)
- des variable issues du LASSO

Enseignement

- Le LASSO retient d'autres variables
- A à peu près le même pouvoir de prévision que le modèle économique
- On aurait du cleaner un peu la base de données (Outliers, transformation, ...)

Overview

Introduction

Régression linéaire pénalisée

Classification
- Introduction
- Linear discriminant analysis (LDA)
- Régression logit
- Support Vector Machine

Données textuelles

Overview

Classification
Introduction
Linear discriminant analysis (LDA)
Régression logit
Support Vector Machine

Introduction

- Dans un problème de classification, on veut assigner une classe à une réponse quantitative.
- On va donc estimer la probabilité de chaque catégorie puis assigner la classe basé sur un critère.
- Exemples : hausse des prix ou baisse des prix, inflation ou désinflation, ...

La régression linéaire est une mauvaise idée

- On cherche à prédire le niveau d'étude selon des caractéristiques sociales :

$$\mathrm{G} = \begin{cases} 1 & \text{si bac} \\ 2 & \text{si licence} \\ 3 & \text{si master} \\ 4 & \text{si doctorat} \end{cases}$$

- Cela suppose le même écart entre les différents niveaux
- qu'il existe un ordre entre les différents niveaux d'éducation

Variable qualitative

On s'intéresse aux méthodes linéaires pour la classification.

- G(x) est à valeur discrète dans G
- Les règles de décisions sont linéaires

Soit deux classes avec :

$$\Pr(G = 1|X = x) = \frac{\exp(\beta_0 + \beta' x)}{1 + \exp(\beta_0 + \beta' x)}$$

$$\Pr(G = 2|X = x) = \frac{1}{1 + \exp(\beta_0 + \beta' x)}$$

alors,

$$\log \frac{\Pr(G = 1|X = x)}{\Pr(G = 2|X = x)} = \beta_0 + x\beta.$$

Critère de décision : $\{x|\beta_0 + \beta' x = 0\}$

Overview

Classification
- Introduction
- Linear discriminant analysis (LDA)
- Régression logit
- Support Vector Machine

Intuition

On cherche à modéliser la distribution des X dans chaque classes, puis on utilise le théorème de Bayes pour estimer $\Pr(G = k|X = x)$:

$$\Pr(G = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}$$

- π_k probabilité a priori que l'observation appartient à la classe k.
- $f_k(x) = \Pr(X = x|G = k)$, la densité de X dans la classe k

Pour p = 1

On fait l'hypothèse que la densité de x sur k est Gaussienne :

$$f_k(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_k)^2\right)$$

donc

$$p_k(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_k)^2\right)}{\sum_{l=1}^{K} \pi_l \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_l)^2\right)}$$

Après quelques modifications, on peut montrer que l'on va assigner l'observation à la classe pour laquelle, la mesure suivante est la plus grande :

$$\delta_k(x) = x\frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k)$$

En pratique

$$\hat{\mu}_k = \frac{1}{n_k} \sum_{i:y_i=k} x_i$$

$$\hat{\sigma}^2 = \frac{1}{n-K} \sum_{k=1}^{K} \sum_{i:y_i=k} (x_i - \hat{\mu}_k)^2$$

$$\hat{\pi}_k = \frac{n_k}{n}$$

Pour p > 1

On a une distribution Gaussienne multivariée :

$$\mathrm{f_k(x)} = \frac{1}{(2\pi)^{\mathrm{p}/2} \mid \mathrm{_k}|^{1/2}} \exp\left(-\frac{1}{2}\left(\mathrm{x} - \mu_\mathrm{k}\right)' \ {}^{-1}_{\mathrm{k}} \left(\mathrm{x} - \mu_\mathrm{k}\right)\right)$$

On peut montrer que :

$$\delta_\mathrm{k}(\mathrm{x}) = -\frac{1}{2}\mu'_\mathrm{k} \ {}^{-1}\mu_\mathrm{k} + \mathrm{x}' \ {}^{-1}\mu_\mathrm{k} + \log(\pi_\mathrm{k})$$

Exemple

Prévision du défault de paiement :

Table – Table de confusion

	Prédit		
	Non	Oui	Erreur
Non	4529	158	0,03
Oui	989	324	0.25
	0.18	0.34	0.19

Exercice

Supposons que nous souhaitons prédire si une action donnée émettra un dividende cette année sur la base du profit X de l'année passée.
Nous examinons un grand nombre d'entreprises : la valeur moyenne de X pour les entreprises ayant émis une dividende était de $\bar{X} = 10$, tandis que la moyenne de ceux qui ne l'ont pas fait était de $\bar{X} = 0$. La variance de X pour ces deux ensembles d'entreprises était de $\hat{\sigma}^2 = 36$.
Enfin, 80% des entreprises ont émis des dividendes. En supposant que X suit une distribution normale, prédisez la probabilité qu'une entreprise émette un dividende cette année étant donné que son pourcentage de profit était de $X = 4$.

Overview

Classification
- Introduction
- Linear discriminant analysis (LDA)
- Régression logit
- Support Vector Machine

Régression logit

- P(y|x) doit être modélisée comme une fonction qui est à valeur entre 0 and 1 quelque soit les valeurs de x.
- On fait alors l'hypothèse d'une variable latente y$*$ (non observée) à valeur entre $-\infty$ et $+\infty$.
- Il existe une propension latente à faire ou non une action qui génère la variable observée.
- On observe pas y$*$, mais à partir d'un certain point, un changement de y$*$ change le résultat que l'observe.

Régression logit

On a alors :

$$y_i* = x_i\beta + \varepsilon_i$$

et

$$G = \begin{cases} 1 & \text{si } y* > \tau \\ 0 & \text{si } y* \leq \tau \end{cases}$$

Régression logit

La régression logistic modélise la probabilité à postériori des K classes via une fonction linéaire des prédicteurs :

$$\begin{aligned}
\log \frac{\Pr(\mathrm{G}=1|\mathrm{X}=\mathrm{x})}{\Pr(\mathrm{G}=\mathrm{K}|\mathrm{X}=\mathrm{x})} &= \beta_{10} + \beta_1'\mathrm{x} \\
\log \frac{\Pr(\mathrm{G}=2|\mathrm{X}=\mathrm{x})}{\Pr(\mathrm{G}=\mathrm{K}|\mathrm{X}=\mathrm{x})} &= \beta_{20} + \beta_2'\mathrm{x} \\
&\vdots \\
\log \frac{\Pr(\mathrm{G}=\mathrm{K}-1|\mathrm{X}=\mathrm{x})}{\Pr(\mathrm{G}=\mathrm{K}|\mathrm{X}=\mathrm{x})} &= \beta_{(\mathrm{K}-1)0} + \beta_{(\mathrm{K}-1)1}'\mathrm{x}.
\end{aligned}$$

Odds Ratio : rapport de deux probabilités complémentaires, la probabilité p de survenue d'un événement (risque), divisé par la probabilité $(1 - \mathrm{p})$ que cet événement ne survienne pas (non risque, c'est-à-dire sans l'événement).

Estimation

On utilise généralement la méthode du maximum de vraisemblance :

$$\mathrm{l}(\theta) = \sum_{\mathrm{i}=1}^{\mathrm{N}} \log \mathrm{p}_{\mathrm{g_i}}(\mathrm{x_i}; \theta)$$

Pour le cas bimodal :

$$\begin{aligned}\mathrm{l}(\beta) &= \sum_{\mathrm{i}=1}^{\mathrm{N}} \{\mathrm{y_i} \log \mathrm{p}(\mathrm{x_i}; \beta) + (1 - \mathrm{y_i}) \log(1 - \mathrm{p}(\mathrm{x_i}; \beta))\} \\ &= \sum_{\mathrm{i}=1}^{\mathrm{N}} \left\{\mathrm{y_i} \beta' \mathrm{x_i} - \log(1 + \exp(\beta' \mathrm{x_i}))\right\}\end{aligned}$$

Maximisation

On pose la dérivée égale à 0 :

$$\frac{\partial \mathrm{l}(\beta)}{\partial \beta} = \sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{x_i}(\mathrm{y_i} - \mathrm{p}(\mathrm{x_i}; \beta)) = 0$$

La hessienne est égale à :

$$\frac{\partial^2 \mathrm{l}(\beta)}{\partial \beta \partial \beta'} = -\sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{x_i x_i'} \mathrm{p}(\mathrm{x_i}; \beta)(1 - \mathrm{p}(\mathrm{x_i}; \beta)).$$

Méthode de Newton

Algorithme :

$$\beta_{\text{new}} = \beta_{\text{old}} - \frac{\partial^2 \text{l}(\beta)}{\partial\beta\partial\beta'}^{-1} \frac{\partial \text{l}(\beta)}{\partial\beta}$$

Explication

LASSO Logit

La pénalité L_1 utilisé par le peut aussi être utilisé pour sélectionner les variables dans le cadre de la régression logit :

$$\max_{\beta_0,\beta} \left\{ \sum_{i=1}^{N} \left[y_i \left(\beta_0 + \beta' x_i \right) - \log \left(1 + \exp \left(\beta_0 + \beta' x_i \right) \right) \right] + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

Overview

Classification
- Introduction
- Linear discriminant analysis (LDA)
- Régression logit
- Support Vector Machine

SVM

Berk (2008) : SVM can be seen as a worthy competitor to random forests and boosting .

- Généralisation du Maximal Margin Classifier
- les classes ne doivent pas forcément être séparée par une frontière linéaire.

Maximal Margin Classifier

Definition

Un hyperplan dans un espace à p-dimension est un sous-espace de dimension $p - 1$.

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p = 0$$

Si un point $X = (X_1, X_2, \ldots, X_p)'$ satisfait cette condition alors il appartient à l'hyperplan.
Si un point ne satisfait pas cette équation, il est soit d'un côté soit de l'autre :

$$\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p > 0$$
$$\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p < 0$$

Hyperplan

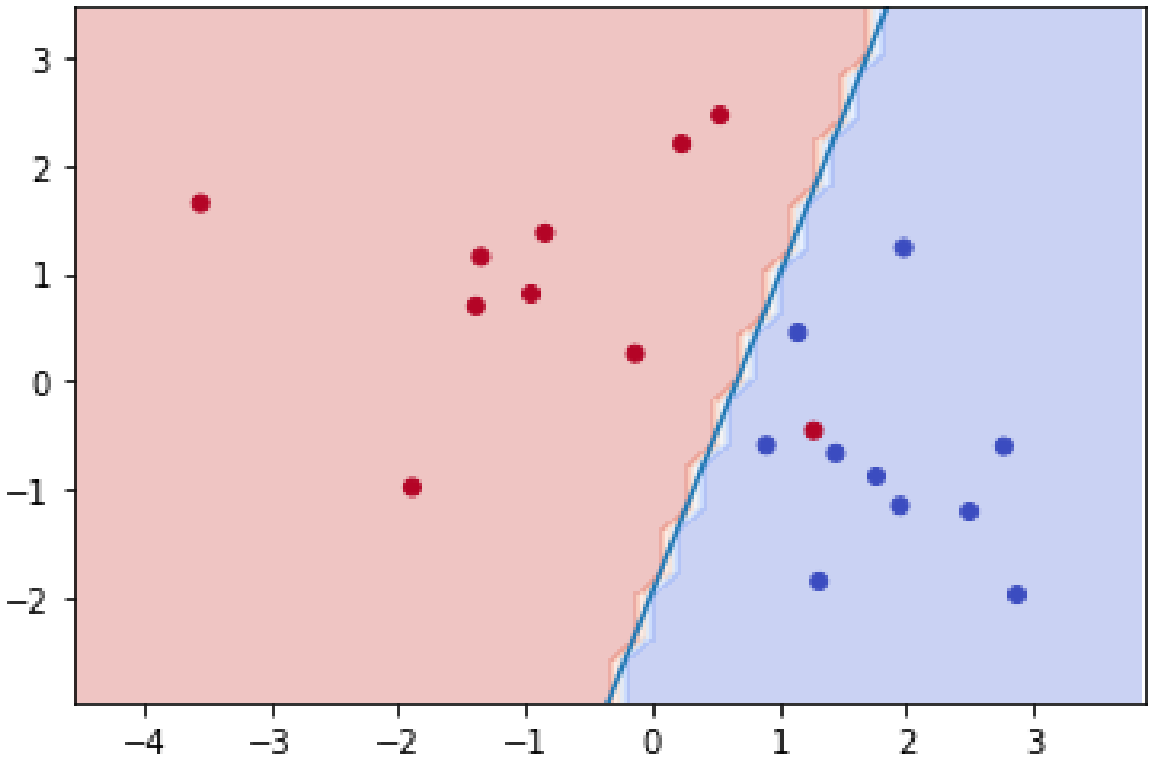


Figure – $x_2 = -0.75 + 2.94x_1$

Maximal Margin Classifier

Soit une matrice n × p X de n observations (training) dans un espace de p-dimension :

$$x_1 = \begin{pmatrix} x_{11} & \vdots & x_{1p} \end{pmatrix}, \ldots, x_n = \begin{pmatrix} x_{n1} & \vdots & x_{np} \end{pmatrix}$$

- Ces observations sont réparties en deux classes : $\{-1, 1\}$
- L'objectif est de développer un classifieur sur les données de training qui va correctement classé les données tests.
- On va utiliser la séparation d'hyperplan.

Maximal Margin Classifier

- Il existe une infinité d'hyperplans parfaitement séparateurs.
- Comme d'habitude, nous aimerions décider parmi l'ensemble possible, quel est le choix optimal, par rapport à un critère.
- Une solution consiste à calculer la distance de chaque observation à un hyperplan séparateur donné. La distance la plus petite s'appelle la marge.
- L'objectif est de sélectionner l'hyperplan séparateur pour lequel la marge est la plus éloignée des observations, c'est-à-dire de sélectionner l'hyperplan à marge maximale.

Maximal Margin Classifier

Trouver un hyperplan qui sépare les deux classes n'est pas compliqué, mais il peut y en avoir une infinité : c'est un problème d'optimisation.

$$\max_{\beta_0,\beta_1,\ldots,\beta_p,M} M$$

$$sc \sum_{i=1}^{p} \beta_i^2 = 1$$

$$y_i(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}) \geq M, \ \forall i = 1, \ldots, n$$

La contrainte définit une marge vide autour de la limite de décision linéaire d'épaisseur $1/||\beta||$. Il peut aussi ne pas y avoir de solution.

SVC

Pour résoudre ce soucis, on introduit le Support Vector Classifier :

$$\max_{\beta_0,\beta_1,\ldots,\beta_p,M} M$$

$$sc \sum_{i=1}^{p} \beta_i^2 = 1$$

$$y_i(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}) \geq M(1 - \xi_i),\ \forall i = 1, \ldots, n$$

avec $\xi_i > 0$ et $\sum_{i=1}^{n} \xi_i \leq C$, C est un paramètre à définir.

SVC

La variable ξ_i est appelée slackvariable :

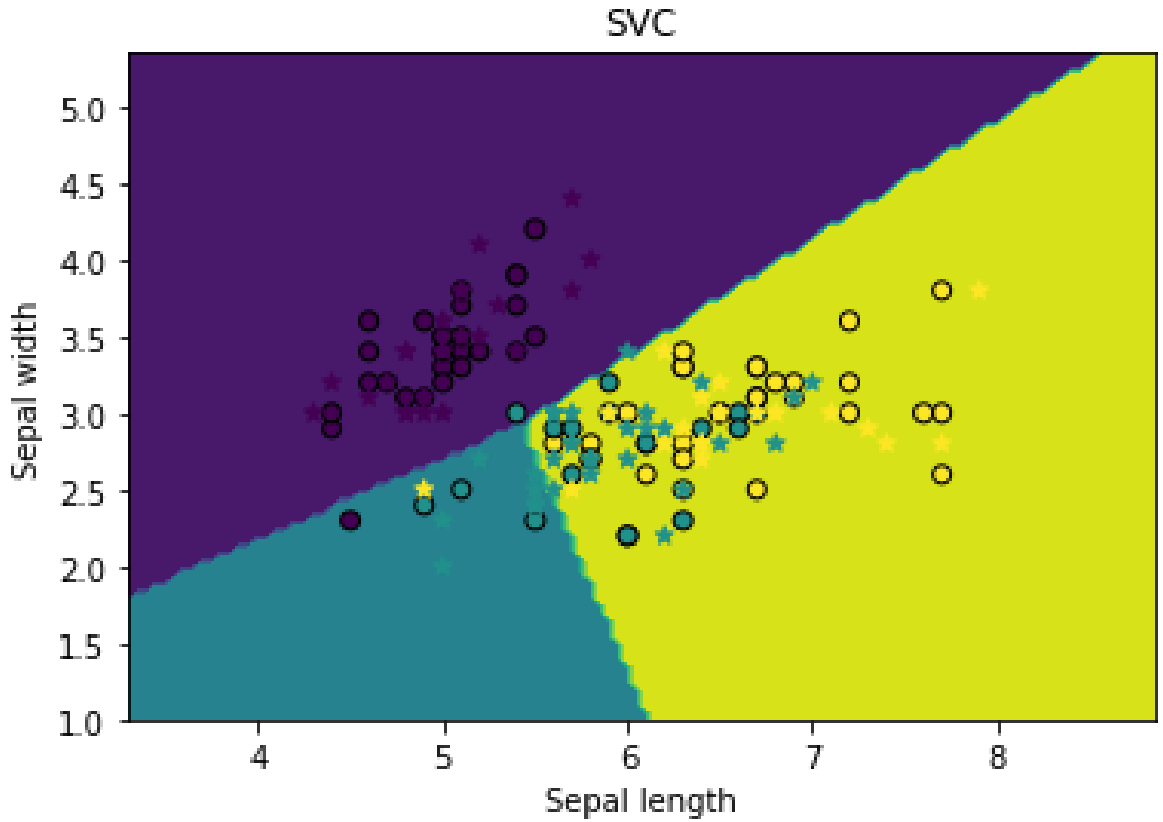
- si $\xi_i = 0$, l'observation est du bon côté de de la marge
- si $\xi_i > 0$, c'est le contraire

Le paramètre C reflète si la contrainte est forte ou non. Est-ce que l'on est plus ou moins permissif :

- $C = 0$, on autorise aucune observation à être du mauvais côté.
- $C > 0$, pas plus de C observation peuvent être du mauvais côté.

Mais, tout ceci est encore linéaire !!

SVC - Iris



Résolution du SVC

Le problème peut se réécrire :

$$\min_{\beta_0,\beta} \frac{1}{2}||\beta||^2 + \mathrm{C}\sum_{\mathrm{i}=1}^{\mathrm{N}} \xi_{\mathrm{i}}$$

$$\text{sc } \xi_{\mathrm{i}} \geq 0,\ \mathrm{y_i}(\mathrm{x_i'}\beta + \beta_0) \geq 1 - \xi_{\mathrm{i}},\ \forall \mathrm{i}$$

Support Vector Machine

- On souhaite avoir une frontière non-linéaire mais sans augmenter la taille de X
- Le SVM va estimer un hyperplan séparateur qui a une dimension plus grande que l'espace engendré par les X
- Au lieu d'utiliser X, on va utiliser un noyau

SVM

Le SVC peut être représenté par :

$$f(x) = \beta_0 + \sum_{i=1}^{n} \alpha_i < x, x_i >$$

Cela demande de calculer tous les produits $< x_i, x_i' >$. Le SVM remplace ces produits par une forme fonctionnelle (un noyau) :

$$f(x) = \beta_0 + \sum_{i=1}^{n} \alpha_i K(x, x_i)$$

SVM

On peuit utiliser un noyau non linéaire comme le noyau polynomial :

$$K(x_i, x_i') = \left(1 + \sum_{j=1}^{p} x_{ij} x_{i'j} + \right)^d$$

Quand le SVC est combiné avec un noyau non linéaire, on appelle cela le Support Vector Machine.

Overview

Introduction

Régression linéaire pénalisée

Classification

Données textuelles
- Introduction
- Topic models

Overview

Données textuelles
Introduction
Topic models

Introduction

- Aug 3rd 2020 - 83k RT @realDonaldTrump : FAKE NEWS IS THE ENEMY OF THE PEOPLE !
- Oct 7th 2020 - 80k 440k THE FAKE NEWS MEDIA IS THE REAL OPPOSITION PARTY !
- Nov 15th 2017 - 25k 112k While in the Philippines I was forced to watch @CNN, which I have not done in months, and again realized how bad, and FAKE, it is. Loser !
- Oct 12th 2020 - 22k 108k Stock Market Up Big. Do I get no credit for this ? Never even mentioned by the Fake News. A New Record for Stocks and Jobs Growth. Remember, “it’s the Economy Stupid”. VOTE !!!
- Oct 11th 2017 - 12k 62k It would be really nice if the Fake News Media would report the virtually unprecedented Stock Market growth since the election.Need tax cuts

Représentation

- Représentation des documents sous forme de matrice terme-document après un pré-process

Table – Matrix terms documents

		Documents				
		1	2	3	4	5
	Fake	1	1	1	1	1
	News	1	1	0	1	1
Terms	Enemy	1	0	0	0	0
	growth	0	0	0	1	1

- Grande dimensionalité des données et sparsité de la matrice terme-document. Thèmes latents.
- Un document peut traiter de plusieurs thèmes

Objectifs

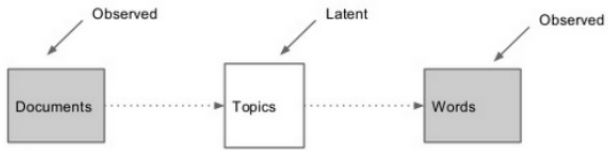


Figure – Topics latent

Figure – Soft clustering

Approches

Plusieurs méthodes :

- Algébriques : LSA (Deerwester et al., 1990), NMF (Paatero et Tapper, 1994), DicPonary learning (JenaFon et al., 2010)
- Approches géométriques : TDT (Allan et al., 1998) (Pons-Porrata et al., 2003)
- Approches probabilistes : pLSA, LDA...

Overview

Données textuelles
Introduction
Topic models

Latent Sementic Analysis

1. construction de la matrix terms-documents
2. Décomposition de la matrice par des valeurs singulières : Identifier les valeurs singulières de la matrice X afin de pouvoir la décomposer en trois matrices $X = UD0V'$ distinctes.
3. La troisième étape consiste à réduire la matrice diagonale D0 de taille $m \times m$ à une matrice diagonale D de taille $k \times k$
4. Calcul de la matrice $X*$

p-LSA

- proposé par Hofmann (SIGIR 1999), est une version probabiliste de LSA. Les poids (p. ex. prépondérance d'un thème dans un document P(t|d)) sont désormais des probabilités, donc positives et interprétables.

$$P(w|d) = \sum_t P(t|d)P(w|d) \tag{1}$$

Latent Dirichlet Allocation

LDA (Latent Dirichlet Allocation), proposé par Blei et al (NIPS 2001), est une alternative à pLSA complètement générative, inspirée par les modèles graphiques probabilistes.

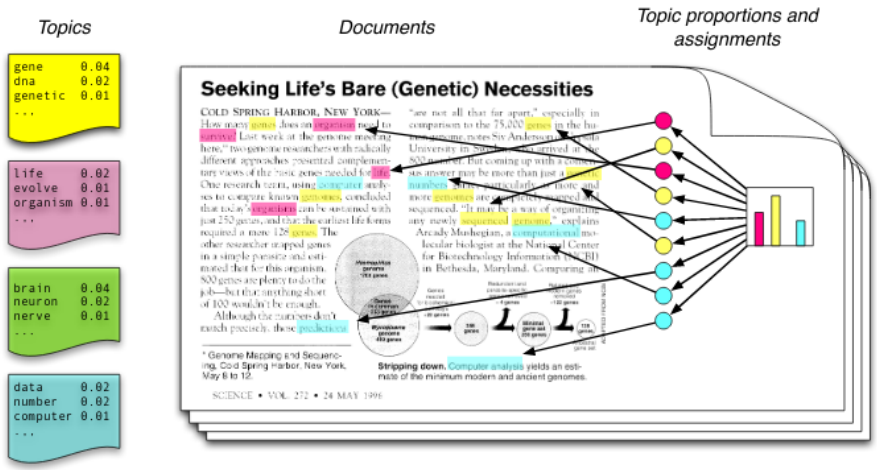


Figure – LDA représentation

Latent Dirichlet Allocation

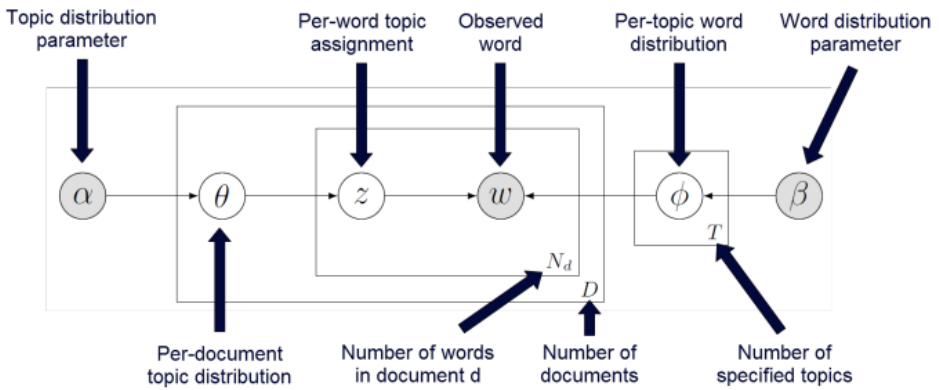


Figure – LDA représentation

- Les nœuds représentent les variables aléatoires
- Les nœuds grisés représentent les variables observés ou fixées
- La flèche entre deux nœuds indique une dépendance conditionnelle
- Les rectangles indiquent la réplication des variables

Latent Dirichlet Allocation

- Un mot w est la donnée discrète, correspondant à l'indice d'un mot dans un vocabulaire fixe de taille V. On peut considérer que w est un vecteur de taille V de composantes toutes nulles sauf pour la composante i où i est l'indice du mot choisi ($w^i = 1$).
- Un document est un N-uplet de mots, $w = (w_1, \ldots, w_N)$.
- Un corpus est une collection de D documents, $D = (w_1, \ldots, w_D)$.
- Les variables $z_{d,n}$, représentent le topic choisi pour le mot $w_{d,n}$.
- Les paramètres θ représentent la distribution de topics du document d.
- α et β définissent les distributions à priori sur θ et ϕ respectivement, où ϕ_k décrit la distribution du topic k

Latent Dirichlet Allocation

L'histoire générative permet de compléter la représentation graphique d'un Topic Model.

1. Pour chaque thème $j \in \{1, \dots, T\}$, tirer une distribution de mots ϕ_j à partir de $\text{Dirichlet}_W(\beta)$.
2. Pour chaque document $d \in \{1, \dots, D\}$:
   - Tirer une distribution de thèmes θ_d à partir de $\text{Dirichlet}_T(\alpha)$;
   - Pour chaque mot d'indice $n \in \{1, \dots, N_d\}$ dans le document d :
     - Tirer un thème $z_{d,n}$ à partir de $\text{Multinomial}_T(\theta_d)$;
     - Tirer un mot $w_{d,n}$ à partir de $\text{Multinomial}_W(\phi_{z_{d,n}})$