Apprentissage statistique avancé

Cours 1 : Apprentissage supervisé

Franck Iutzeler

Département de mathématiques
Université de Toulouse

Sommaire

Sommaire

Un problème d'apprentissage statistique

- Un jeu de données populaire en apprentissage : les passagers du Titanic

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S
...	...	...	...	...	...	...	...	...	...	...	...	...
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.0000	NaN	S

- Comment prédire la survie d'une personne à partir de ses informations ?
  - On modélise les informations des individus comme des réalisation d'une variable aléatoire
  - Idem pour la survie
  - Inférence statistique = trouver un lien entre la survie et les informations

Sommaire

Retour sur les passagers du Titanic

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S
...	...	...	...	...	...	...	...	...	...	...	...	...
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.0000	NaN	S

- Nous étudions les ***caractères*** ***PassengerId, Survived, ...***
  - Ce sont des ***variables aléatoires***
  - Nous les noterons ***en capitale*** $X^1, X^2, ..., X^k$
  - Le caractère à prédire (ici la survie) est souvent noté Y
- Nous observons des ***réalisations*** sur un ***échantillon de population***
  - Ce sont des valeurs (pas aléatoires) correspondants aux n individus de l'échantillon
  - Nous les noterons ***en minuscule*** $x_1^1, ..., x_n^1$ pour la variable X^1, etc.

Variables qualitatives et quantitatives

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S
...	...	...	...	...	...	...	...	...	...	...	...	...
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.0000	NaN	S

- Les différentes valeurs possibles d'un caractère sont appelées ***modalités***
- On peut distinguer les cas
  - ***Qualitatif*** : ***nominal*** *Survived, Name, Sex, Ticket, Cabin* ou ***ordinal*** *PClass, Embarked*
  - ***Quantitatif*** : ***discret*** *Age, SibSp, Parch* ou ***continu*** *Fare*
- Certaines réalisations peuvent être ***manquantes***

Sommaire

Analyse de la variable *Survived*

- Oublions dans un premier temps le reste des données

Survived	0	1
effectif	549	342
fréquence	0.62	0.38

- Que peut-on dire?

Estimation de la probabilité de survie

	Survived
0	0
1	1
2	1
3	1
4	0
...	...
886	0
887	1
888	0
889	1
890	0

Survived	0	1
effectif	549	342
fréquence	0.62	0.38

- On ***modélise*** la survie de l'individu i par une variable aléatoire $Y_i \sim \mathcal{B}(p)$
  Les $Y_1, .., Y_n$ sont indépendants et de même loi, on parle d'***échantillon aléatoire***
- On ***estime*** la valeur de p (inconnue) à l'aide d'un ***estimateur***
  Ici, l'estimateur naturel est la fréquence empirique : $F_n = \frac{1}{n}\sum_{i=1}^{n} Y_i$
- À partir des réalisations, on obtient une ***estimation***
  Comme calculé plus haut, on a $f_n = \frac{1}{n}\sum_{i=1}^{n} y_i = 0.38$

Prédiction de la survie

	Survived
0	0
1	1
2	1
3	1
4	0
...	...
886	0
887	1
888	0
889	1
890	0

Survived	0	1
effectif	549	342
fréquence	0.62	0.38

- On ***prédit*** la valeur de la réalisation d'un nouvel individu (non-observé)

  On considère une nouvelle v.a. Y_{n+1} suivant la même distribution que ***Survived***
- Nous avons appris un modèle pour la variable survie : $\mathcal{B}(0.38)$

  En prédisant 0 pour y_{n+1}, on se trompe le moins souvent
- Nous faisons plusieurs types d'erreur :
  - erreur de modélisation, la loi de la survie est plus complexe que $\mathcal{B}(p)$
  - erreur d'estimation, distance entre f_n est p
  - erreur de variance, même avec la vraie loi de Y_{n+1}, cela reste une v.a.

Sommaire

Estimation

- ***Hypothèse fondamentale*** : les v.a. $(Y_1, .., Y_n)$ sont indépendantes et suivent la même loi $\mathbb{P}_Y$
- ***Modélisation*** : la loi $\mathbb{P}_Y$ fait partie d'une famille paramétrée $\mathbb{P}_\theta$ avec $\theta \in \Theta$
- ***Estimation paramétrique*** : l'ensemble des paramètres Θ est simple discret, dimension finie
- ***Estimateur*** : v.a. $g(Y_1, .., Y_n)$ qui estime θ sans biais : $\mathbb{E}[g(..)] = \theta$, consistant : $\mathbb{V}[g(..)] \to_{n\to\infty} 0$
- ***Estimation*** : réalisation $\hat{\theta} = g(y_1, .., y_n)$ sur le jeu de données
- ***Prédiction*** : la loi d'un nouvel individu est $\mathbb{P}_{\hat{\theta}}$ avec une certaine erreur

Apprentissage supervisé

Similaire à l'estimation mais en intégrant des co-variables explicatives $(X_1, .., X_n)$

- ***Hypothèse fondamentale*** : les v.a. $((X_1, Y_1), .., (X_n, Y_n))$ sont indépendantes et suivent la même loi $\mathbb{P}_{(X,Y)}$
- ***Modélisation*** : la loi *conditionnelle* $\mathbb{P}_{Y|X}$ fait partie d'une famille paramétrée $\mathbb{P}_\theta$ avec $\theta \in \Theta$
- ***Estimation paramétrique*** : l'ensemble des paramètres Θ est simple
- ***Estimateur*** : v.a. $g((X_1, Y_1), .., (X_n, Y_n))$ qui estime θ
- ***Estimation*** : réalisation $\hat{\theta} = g((x_1, y_1), .., (x_n, y_n))$ sur le jeu de données
- ***Prédiction*** : la loi de Y pour un nouvel individu sachant $X = x$ est $\mathbb{P}_{\hat{\theta}}$ avec une certaine erreur

Exemple: Les femmes et les enfants d'abord

	Survived	Sex
0	0	male
1	1	female
2	1	female
3	1	female
4	0	male
...	...	...
886	0	male
887	1	female
888	0	female
889	1	male
890	0	male

Sex : male

Survived	0	1
effectif	468	109
fréquence	0.81	0.19

Sex : female

Survived	0	1
effectif	81	233
fréquence	0.26	0.74

- On ***modélise*** la survie de l'individu i par une variable aléatoire Y_i
  $Y_i \sim \mathcal{B}(p^M)$ si $X_i = male$ et $Y_i \sim \mathcal{B}(p^F)$ si $X_i = female$
- On ***estime*** les valeur de p^M et p^F à l'aide d'un ***estimateur***
  comme précédemment
- À partir des réalisations, on obtient une ***estimation***
  Comme calculé plus haut, on a $f_n^M = 0.19$ et $f_n^F = 0.74$

Exemple: Les femmes et les enfants d'abord

	Survived	Sex
0	0	male
1	1	female
2	1	female
3	1	female
4	0	male
...	...	...
886	0	male
887	1	female
888	0	female
889	1	male
890	0	male

Sex : male

Survived	0	1
effectif	468	109
fréquence	0.81	0.19

Sex : female

Survived	0	1
effectif	81	233
fréquence	0.26	0.74

- On ***prédit*** la valeur de la réalisation d'un nouvel individu (non-observé)

  On considère une nouvelle v.a. (jointe) X_{n+1}, Y_{n+1} suivant la même distribution

  On observe la réalisation x_{n+1} de la variable X_{n+1}
- Nous avons appris un modèle pour la variable survie :
  $\mathcal{B}(0.19)$ si x_{n+1} est *male*, $\mathcal{B}(0.74)$ sinon

  On prédit 0 pour les hommes et 1 pour les femmes
- En utilisant la ***variable explicative*** ***Sex*** , on a amélioré notre prédiction

  De combien?

Sommaire

Sommaire

Qu'est-ce que l'apprentissage ?

Apprentissage (*machine learning*) : discipline visant à la construction de règles d'inférence et de décision pour le traitement automatique des données

- Cela ne dit pas :
  - À quelles données on a accès
  - Quelles sont nos hypothèses
  - Pour quel objectif
- Problèmes très différents → communautés et vocabulaires très différents

Trois grandes "familles"

- ***Apprentissage supervisé :*** A partir d'un échantillon d'apprentissage $\mathcal{D}_n = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, inférer la relation entre x et y.
  Synonymes: discrimination, reconnaissance de formes (pattern recognition)
  Voc: x_i = caractéristique = feature = variable explicative
- ***Apprentissage non supervisé :*** A partir d'un échantillon d'apprentissage $\mathcal{D}_n = \{x_1, \ldots, x_n\} \subset \mathcal{X}$, inférer des propriétés de $\mathcal{X}$, par exemple partitionner $\mathcal{X}$ en classes pertinentes (clustering), réduction de dimension.
  Voc: parfois appelé 'Classification' en français (jamais en anglais)
- ***Apprentissage séquentiel :*** A chaque date n, prendre une décision à l'aide des données passées $\mathcal{D}_n$, gagner une certaine récompense à l'instant $n+1$ calculée à partir de la nouvelle donnée, et ainsi de suite.

Différentes approches

- ***Apprentissage statistique :*** modélisation probabiliste des données à des fins *instrumentales*
- Apprentissage séq. robuste (théorie des jeux, optim. convexe séq.)
- Approche structurelle, logique et logique floue

Dans tous les cas:

- science relativement *récente*
- à la frontière entre *mathématiques* et *informatique* (et intelligence artificielle)
- en *évolution rapide et constante* avec les technologies
  - nouveaux moyens (de calcul)
  - nouveaux problèmes...

Sommaire

Processus

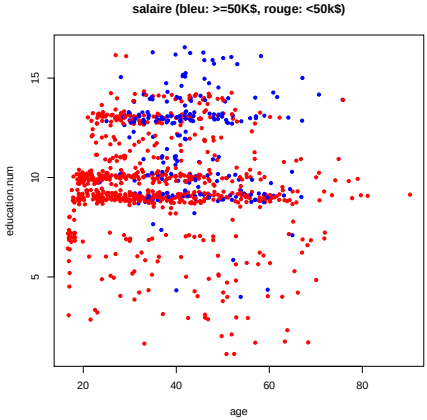
1. Définir le problème à résoudre (objectif final)
2. Acquérir des données (récolter, rassembler, attention au cadre légal)
3. Analyser et explorer les données (représentations graphiques, corrélations, biais)
4. Préparer et nettoyer les données (ré-encodage, remplissage, équilibrage, standardisation)
5. Ingénierie de caractéristiques (nouvelles variables plus exploitables ou efficaces)
6. Définir un modèle d'apprentissage (hypothèses, formulation, métrique)
7. Entraîner et optimiser le modèle (minimisation du risque, gestion des paramètres)
8. Évaluer la performance (sur un jeu de données indépendant)
9. Expliquer et déployer (quelles sont les variables importantes, code efficace)

Processus

1. Définir le problème à résoudre (objectif final)
   - Classification (variable à prédire discrète) ou Régression (continue)

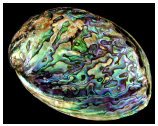
United States™ Census Bureau

Objectif: prédire qui gagne plus de 50k$ à partir de données de recensement.

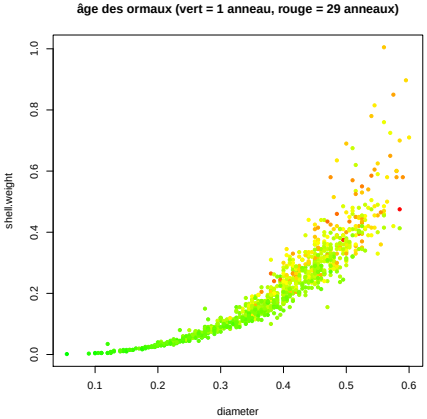


Processus

1. Définir le problème à résoudre (objectif final)
   - Classification (variable à prédire discrète) ou Régression (continue)



Prédire l'âge d'un ormeau (abalone) à partir de sa taille, son poids, etc.



Processus

2. Acquérir des données (récolter, rassembler, attention au cadre légal)

- Déjà présentes ou à collecter

De nombreux jeux de données sont en accès libre (INSEE, data.gouv.fr,data.toulouse-metropole.fr)

Top 500 des films les plus empruntés à la Bibliothèque de Toulouse

Informations | Tableau | Export | API | Commentaires (0)

Depuis 2011, listes des 500 films les plus empruntés dans toutes les bibliothèques du réseau de la Bibliothèque de Toulouse (BMVR), triés par nombre de prêts enregistrés cette année-là.

Identifiant du jeu de données	top-500-des-films-les-plus-empruntes-a-la-bibliotheque-de-toulouse
Téléchargements	59 211
Thèmes	Culture, Statistiques
Mots clés	Bibliothèque, Médiathèque, BM, BMVR, DVDs, Best-seller, Statistiques
Licence	Licence Ouverte v2.0 (Etalab)
Langue	Français
Modifié	18 janvier 2023 16:57
Producteur	Mairie de Toulouse
Villes concernées	Toulouse
Territoire	Toulouse Métropole
Dernier traitement	18 janvier 2023 16:57 (métadonnées) 18 janvier 2023 16:57 (données)

DCAT

Créé	10 juin 2015
Publié	25 juin 2015
Créateur	Bibliothèque de Toulouse
Fréquence de mise à jour	Annuelle
Période de temps	An

Processus

2. Acquérir des données (récolter, rassembler, attention au cadre légal)

- Déjà présentes ou à collecter

De nombreux jeux de données sont en accès libre (INSEE, data.gouv.fr,data.toulouse-metropole.fr)

	ANNEE	Nbre de prêts	TITRE	AUTEUR	Editeur	Indice	BIB	COTE	Cat 1	Cat 2
1	2 013	522	Rebelle	Andrews, Mark	Marne-la-Vallée : The Walt Disney C...	A REBE	CABANIS	A REBE	E	DVDFIC
2	2 014	458	Le livre de la jungle	Reitherman, Wolfgang	Paris : Warner Home Video, 1999.	A LIVR	CABANIS	A LIVR	E	DVDFIC
3	2 012	429	Harry Potter et la chambre des secr...	Columbus, Chris	Paris : Warner Home Video, 2002.	HARR	CABANIS	HARRY	E	DVDFIC
4	2 012	429	Raiponce	Greno, Nathan	Marne-la-Vallée : The Walt Disney C...	A RAIP	CABANIS	A RAIP	E	DVDFIC
5	2 014	427	La petite sirène	Musker, John	Burbank, Calif. : Buena Vista Home ...	A PETI	CABANIS	A PETI	E	DVDFIC
6	2 014	419	Rebelle	Andrews, Mark	Marne-la-Vallée : The Walt Disney C...	A REBE	CABANIS	A REBE	E	DVDFIC
7	2 015	419	Le livre de la jungle	Reitherman, Wolfgang	Paris : Warner Home Video, 1999.	A LIVR	CABANIS	A LIVR	E	DVDFIC
8	2 012	414	Harry Potter et la coupe de feu	Newell, Mike	Paris : Warner Home Video, 2006	F HARR	CABANIS	HARR	E	DVDFIC
9	2 016	405	Mon voisin Totoro	Miyazaki, Hayao	[S.l. : Buena Vista Home Interaisime...	A MONV	CABANIS	Anim MONV	E	DVDFIC
10	2 016	404	Le livre de la jungle	Reitherman, Wolfgang	Paris : Warner Home Video, 1999.	A LIVR	CABANIS	A LIVR	E	DVDFIC
11	2 015	400	La petite sirène	Musker, John	Burbank, Calif. : Buena Vista Home ...	A PETI	CABANIS	A PETI	E	DVDFIC
12	2 012	400	Harry Potter et le prince de sang-mêlé	Yates, David	Neuilly-sur-Seine : Warner Home Vid...	HARR	CABANIS	HARR	E	DVDFIC
13	2 012	398	Ponyo sur la falaise	Miyazaki, Hayao	Marne-la-Vallée : Buena Vista Home...	A PONY	CABANIS	A PONY	E	DVDFIC
14	2 012	397	Charlie et la chocolaterie	Burton, Tim	Neuilly-sur-Seine : Warner Home Vid...	CHAR	CABANIS	P CHAR	E	DVDFIC
15	2 014	396	Le Hobbit : un voyage inattendu	Jackson, Peter	Neuilly-sur-Seine : Warner Home Vid...	RTF TOLK	CABANIS	HOBB	A	DVDFIC
16	2 016	395	Vice-versa	Docter, Pete	Marne-la-Vallée : The Walt Disney C...	Anim VICE	CABANIS	Anim VICE	E	DVDFIC
17	2 016	395	Kiki la petite sorcière	Miyazaki, Hayao	[S.l. : Buena Vista Home Interaisime...	A KIKI	CABANIS	Anim KIKI	E	DVDFIC
18	2 015	391	Mon voisin Totoro	Miyazaki, Hayao	[S.l. : Buena Vista Home Interaisime...	A MONV	CABANIS	Anim MONV	E	DVDFIC
19	2 013	391	Le Chat Potté	Miller, Chris	s.l. : DreamWorks France, 2012	A CHAT	CABANIS	A CHAT	E	DVDFIC
20	2 016	389	La belle au bois dormant	Geronimi, Clyde	Marne-la-Vallée : Buena Vista Home...	A BELL	CABANIS	Anim BELL	E	DVDFIC
21	2 015	386	La belle au bois dormant	Geronimi, Clyde	Marne-la-Vallée : Buena Vista Home...	Anim BELL	CABANIS	Anim BELL	E	DVDFIC
22	2 013	385	Avengers	Whedon, Joss	Marne-la-Vallée : The Walt Disney C...	SF AVEN	CABANIS	SF AVEN	A	DVDFIC
23	2 012	384	Kirikou et la sorcière	Ocelot, Michel	Paris : France Télévision Distribution...	A KIRI	CABANIS	Anim KIRI	E	DVDFIC
24	2 015	381	Rebelle	Andrews, Mark	Marne-la-Vallée : The Walt Disney C...	A REBE	CABANIS	A REBE	E	DVDFIC
25	2 012	380	Harry Potter et l'ordre du Phénix	Yates, David	Neuilly-sur-Seine : Warner Home Vid...	HARR	CABANIS	F HARR	E	DVDFIC
26	2 016	379	La princesse et la grenouille	Clements, Ron	Marne-la-Vallée : The Walt Disney C...	A PRIN	CABANIS	Anim PRIN	E	DVDFIC
27	2 013	379	La belle et le clochard	Luske, Hamilton	[S.l. : Buena Vista Home Interaisime...	A BELL	CABANIS	A BELL	E	DVDFIC

Processus

2. Acquérir des données (récolter, rassembler, attention au cadre légal)
   - Déjà présentes ou à collecter

Définition d'un protocole
Attention : le RGPD engage la personne qui récolte les données à informer les personnes concernées par opt-in sur :

- Pour quelles utilisations sont stockées les données
- Combien de temps et où elles seront stockées

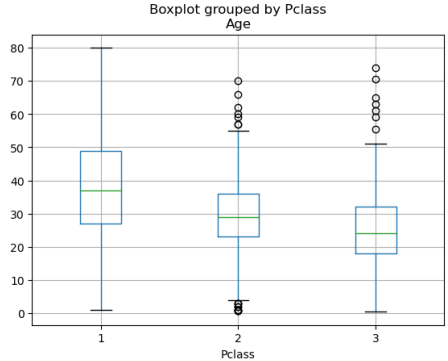
n° question	texte question	code	n° colonne Excel / nom variable	
1	Êtes-vous un homme ou une femme ?	Homme, Femme	A	sexe
2	Vivez-vous seul(e), en couple ou en famille ?	Seul, Couple, Famille	B	situation
3	Quelle est votre consommation quotidienne moyenne de tasses de thé ?	en nombre de tasses (entier)	C	the
4	Quelle est votre consommation quotidienne moyenne de tasses de café ?	en nombre de tasses (entier)	D	cafe
5	Quelle est votre taille ?	en centimètre (cm)	E	taille
6	Quel est votre poids ?	en kilogramme (kg)	F	poids
7	Quel est votre âge ?	en année	G	age
8	Quelle est votre consommation de viande ?	Jamais : jamais 1 fois/sem : 1 fois/semaine 2 fois/sem : 2 fois/semaine 3 fois/sem : 3 fois/semaine 4 fois/sem : 4 fois/semaine 1 fois/j : 1 fois/jour	H	viande
9	Quelle est votre consommation de poisson ?	0 : jamais 1 : moins d'1 fois/semaine 2 : 1 fois/semaine 3 : 2 ou 3 fois/semaine 4 : 4 à 6 fois/semaine 5 : tous les jours	I	poisson
10	Quelle est votre consommation de fruits ou légumes crus ?	0 : jamais 1 : moins d'1 fois/semaine 2 : 1 fois/semaine 3 : 2 ou 3 fois/semaine 4 : 4 à 6 fois/semaine 5 : tous les jours	J	fruit_crus

Processus

3. Analyser et explorer les données (représentations graphiques, corrélations, biais)

- Quelles sont les variables utiles/biaisées?

Explorer graphiquement, corrélations
Exemple avec le jeu de données du Titanic

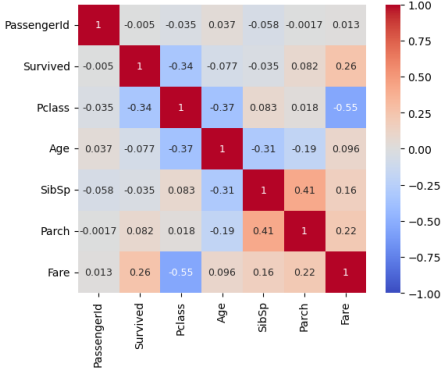


Processus

3. Analyser et explorer les données (représentations graphiques, corrélations, biais)

- Quelles sont les variables utiles/biaisées?

Explorer graphiquement, corrélations
Exemple avec le jeu de données du Titanic

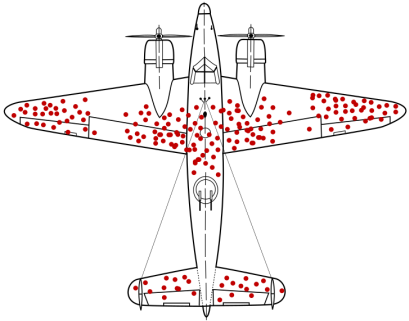


Processus

3. Analyser et explorer les données (représentations graphiques, corrélations, biais)

- Quelles sont les variables utiles/biaisées?

Attention au biais dans la sélection des individus, création des variables, etc. Ces biais vont se retrouver dans la modèle final.



Processus

4. Préparer et nettoyer les données (ré-encodage, remplissage, équilibrage, standardisation)

- Données manquantes
  - Causes possibles: données censurées, défaut de mesure, pas de raison
  - Remédiations classiques
    - imputation de "0" / "NR"
    - imputation de la moyenne, médiane ou mode de la variable (ou autre)
  - Le choix n'est pas innocent et peut biaiser le résultat, à réfléchir
  - Si un individu a "trop" de données manquantes, il peut être judicieux de l'enlever
- Recodage de variables existantes (en numérique)
  - Variables qualitatives nominales → une variable numérique par modalité "0/1" ***one-hot encoding***
  - Variables qualitatives ordinales → une variable numérique "0/1/2/3"

optionel Variables quantitatives continues → discrete par intervalle ***binning***

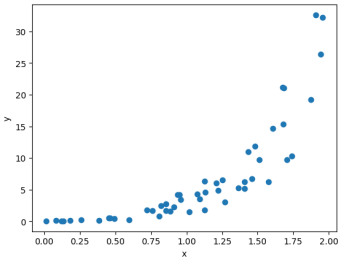
conseillé ***Normaliser*** les variables quantitatives pour avoir une moyenne de 0 et une variance de 1

- Toujours garder une trace des modifications opérées
  - Les mêmes traitements devront être fait pour chaque nouvel individu

Processus

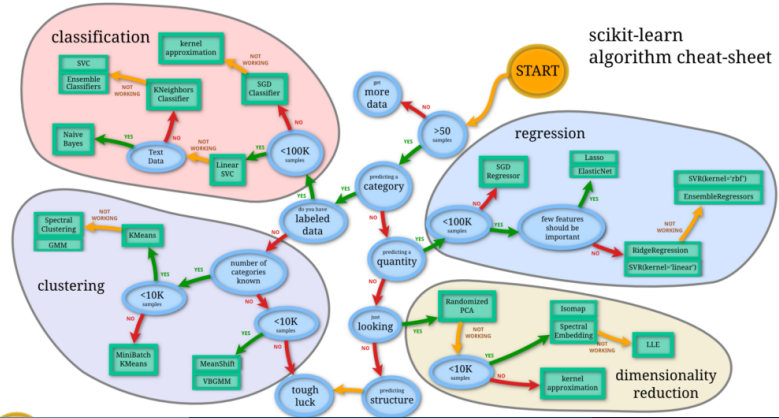
5. Ingénierie de caractéristiques (nouvelles variables plus exploitables ou efficaces)

- A partir de visualisations (et de réflexions), former des caractéristiques intéressantes
- Le coefficient de corrélation de Pearson cherche un lien *linéaire* !
  - $r(x, y) = 0.79$
  - $r(x, \sqrt[4]{y}) = 0.96$
- C'est également le cas de beaucoup de modèles d'apprentissage
- Regrouper une variable quantitative en classes peut être intéressant (Age $\longrightarrow$ enfant, ado, jeune adulte, adulte, personne agée)



Processus

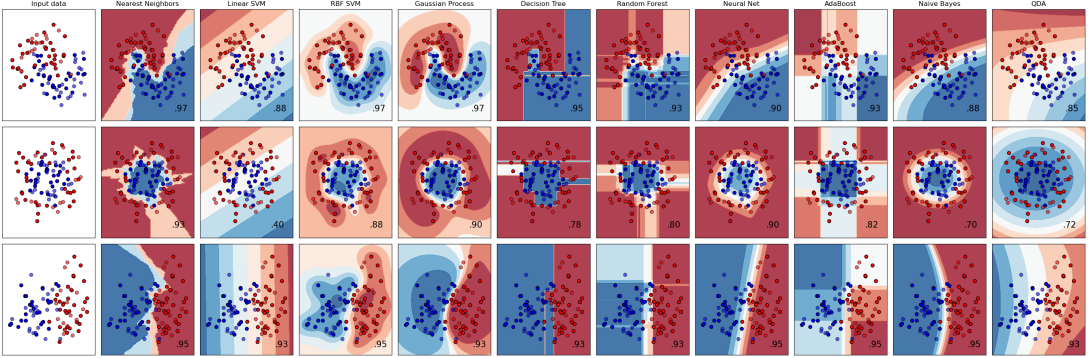
6. Définir un modèle d'apprentissage (hypothèses, formulation, métrique)
   - Lister les modèles modèles possibles, et trier les plus raisonnables (objectif, linéarité, etc.)
   - Par exemple, sur le site de scikit-learn, on trouve



Processus

6. Définir un modèle d'apprentissage (hypothèses, formulation, métrique)

- Lister les modèles modèles possibles, et trier les plus raisonnables (objectif, linéarité, etc.)
- Par exemple, sur le site de scikit-learn, on trouve



Processus

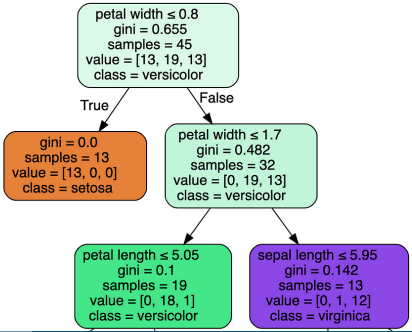
7. Entraîner et optimiser le modèle (minimisation du risque, gestion des paramètres)
8. Évaluer la performance (sur un jeu de données indépendant)

- Deux étapes intimement liées
- Bibliothèques d'apprentissage ou d'optimisation (il est rarement utile de coder l'entrainement)
- Quand arrêter l'entrainement ?
  - Quand la perte du modèle atteint une certaine précision ? (ex. 'tol' dans le Lasso de sklearn)
  - Quand l'erreur sur un jeu de test/validation cesse de diminuer ?
  - Après un certain nombre d'iterations ('max_iter') ou de passes sur les données (epochs)
- Il faut donc choisir une métrique d'évaluation avant d'optimiser son modèle

Processus

9. Expliquer et déployer (quelles sont les variables importantes, code efficace)

- "Vendre" le modèle et le rendre utilisable
  - Importance des variables (selon les modèles)
  - Fonction de prédiction (le coût d'une prédiction change selon le modèle)
  - Le pipeline peut être coûteux si beaucoup de pré-traitement (pub en ligne = Réponse en $\approx$ 10ms)
  - L'explicabilité peut être une demande du client



Résumé

Données *échantillon d'apprentissage* $(x_k, y_k)_{1 \leq k \leq n}$ dans $\mathcal{X} \times \mathcal{Y}$, constitué d'observations que l'on suppose représentatives et sans lien entre elles

Objectif prédire les valeurs de $y \in \mathcal{Y}$ associées à chaque valeur possible de $x \in \mathcal{X}$

Formalisation du problème Classification si $\mathcal{Y}$ discret (binaire ou multi-classe), Régression sinon

Étude préliminaire qualitative allure des distributions (graphiques), présence de données atypiques, corrélations et cohérence, transformations éventuelles des données (normalisation), description multidimensionnelle (PCA), classification (clustering)

Apprentissage Cœur du sujet, faisant appel à des bibliothèques

Règle de décision à partir de l'échantillon d'apprentissage, construire $f_n : \mathcal{X} \to \mathcal{Y}$ associant, à chaque entrée possible x une valeur de y prédite

Sommaire

Apprentissage et (vs. ?) Statistique

Modèle linéaire Gaussien : $y_k = \beta^\top x_k + \varepsilon_k \qquad (k = 1, \dots, n)$

- ***Hypothèses :*** Lien linéaire entre x et y + "bruit" $(\varepsilon_k)_{k=1}^n$ i.i.d. Gaussien centré
- ***Intéret :*** modèle simple $\longrightarrow$ on peut calculer la paramètre β qui maximise la vraisemblance
  - $y_k | x_k \sim \mathcal{N}(\beta^\top x_k, \sigma^2)$
  - On cherche le paramètre β qui maximise la densité de $f(y; \beta | x)$

$$\begin{aligned}\hat{\beta} &= \arg\sup_\beta \frac{1}{\sqrt{2\pi}\sigma} e^{-(y_1 - \beta^\top x_1)^2/(2\sigma^2)} \times .. \times \frac{1}{\sqrt{2\pi}\sigma} e^{-(y_n - \beta^\top x_n)^2/(2\sigma^2)} \\ &= \arg\inf_\beta \sum_{k=1}^n (y_k - \beta^\top x_k)^2\end{aligned}$$

  - Moindres carrés ordinaires
- Nous pouvons trouver le paramètre "optimal" pour la vraisemblance statistique

Apprentissage et (vs. ?) Statistique

Modèle linéaire Gaussien : $y_k = \beta^\top x_k + \varepsilon_k \qquad (k = 1, \dots, n)$

- ***Hypothèses :*** Lien linéaire entre x et y + "bruit" $(\varepsilon_k)_{k=1}^n$ i.i.d. Gaussien centré
- ***Situation en apprentissage :*** Les données $(x_k, y_k)_{k=1}^n$ sont des réalisations de v.a. iid de même loi $(X, Y) \sim P_{(X,Y)}$
- ***Limites du modèle Gaussien :*** En apprentissage, P existe mais n'est pas caractérisé
  - Pas de "vrai modèle", de "vraies valeurs du paramètre"
  - Un modèle linéaire peut être intéressant en dehors du cadre Gaussien
  - Tous les coups sont permis, seul critère = efficacité prédictive.

Apprentissage et (vs. ?) Statistique

Modèle logistique : $\mathbb{P}[y_k = 1] = \frac{1}{1+\exp(-\beta^\top x_k)} \qquad (k = 1, \dots, n)$

- ***Hypothèses :*** y Bernoulli de probabilité dépendant de $\beta^\top x$
- ***Intéret :*** modèle simple $\longrightarrow$ on peut calculer la paramètre β qui maximise la vraisemblance
  - $y_k | x_k \sim \mathcal{B}(\frac{1}{1+\exp(-\beta^\top x_k)})$
  - On cherche le paramètre β qui maximise la densité de $f(y; \beta | x)$

$$\hat{\beta} = \arg\inf_{\beta} \sum_{k=1}^{n} \left(-y_k \log\left(\frac{1}{1+\exp(-\beta^\top x_k)} \right) - (1-y_k) \log\left(1 - \frac{1}{1+\exp(-\beta^\top x_k)} \right) \right)$$

  - Regression logistique
- Mêmes défauts et avantages que dans le cas précédent

Données simulées

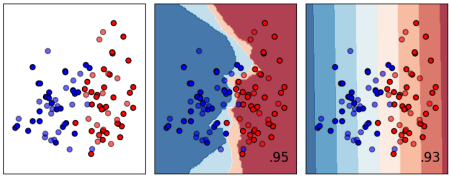
- Il est courant d'utiliser des générateurs de données en apprentissage
  - On se donne un *modèle génératif* (on spécifie une loi jointe P) sklearn: make blobs, moons, etc.
  - On entraîne notre modèle sur des données générées $(x_k, y_k)_{k=1}^n$ (réalisations de v.a. iid de même P)
- Avantages : on contrôle P et on peut générer autant de données que l'on veut
  - On peut parfois calculer la règle de décision optimale
  - On peut estimer des taux d'erreurs *sur* P par la loi des grands nombres
- Utile pour comparer des prédicteurs

Estimation par Monte-Carlo, loi des grands nombres : Soit P une mesure de probabilité sur $\mathbb{R}^p$ et (X_k) une suite de v.a. iid. de loi P. Alors, pour toute fonction f continue sur $\mathbb{R}^p$ (plus généralement mesurable), pourvu que l'espérance existe

$$\frac{1}{n}\sum_{k=1}^{n} f(X_k) \underset{p.s.}{\longrightarrow} \mathbb{E}_{X\sim P}[f(X)] = \int_{x\in\mathbb{R}^p} f(x)\,\mathrm{d}P(x)$$

Paramétrique vs. Non-paramétrique

- Un modèle est paramétrique si la prédiction est de la forme $f_\theta(x)$ pour $\theta \subset \mathbb{R}^d$
- Théoriquement, quand un modèle (paramétrique) est vrai il est optimal de l'utiliser
  - Théorème de Gauss-Markov: parmi les estimateurs sans biais, celui des moindres carrés est de variance minimale
  - ***mais*** on peut avoir intérêt à sacrifier du biais contre de la variance !
- Même quand il y en a un 'vrai' modèle, on n'a pas forcément intérêt à l'utiliser
- Des approches ***non-paramétriques*** peuvent avoir une efficacité proche :
  - exemple : k-NN versus régression logistique



- ... et ils peuvent être sont plus robustes et plus simples à entraîner qu'un modèle paramétrique en très grande dimension !

Pas de meilleure méthode

- Plus ou moins adaptée au problème, nature des données, relation entre descripteurs et variable expliquée...
- Apprendre les qualités et les défauts de chaque méthode
- Apprendre à expérimenter pour identifier la plus pertinentes pour un problème donné
- Estimer de la performances des méthodes est central (mais pas toujours évident)

Sommaire

Sommaire

Fonction de perte

Données *Échantillon d'apprentissage* $(x_k, y_k)_{1 \leq k \leq n}$ dans $\mathcal{X} \times \mathcal{Y}$
Hypothèse : observations iid. selon la loi P sur $\mathcal{X} \times \mathcal{Y}$

Vocabulaire X est une variable explicative, Y est une variable à expliquer

Objectif prédire les valeurs de $y \in \mathcal{Y}$ associées à chaque valeur possible de $x \in \mathcal{X}$
Classification ($\mathcal{Y} = \{0, 1\}$, $\mathcal{Y} = \{bleu, rouge, vert\}$) ou Regression ($\mathcal{Y} = \mathbb{R}$, $\mathcal{Y} = [a, b]$)

Règle de décision/prédicteur fonction $f_n : \mathcal{X} \to \mathcal{Y}$ qui prédit une valeur $\hat{y}$ pour une entrée $x \in \mathcal{X}$

- ***Fonction de perte (loss)*** $L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ mesure l'erreur entre prédiction et vérité
  $L(y, y) = 0$, erreur de prédiction pour le point (x, y) : $L(\hat{y}, y)$ où $\hat{y} = f_n(x)$
  Exemples typiques :
  - Régression : $L(\hat{y}, y) = (\hat{y} - y)^2$ (ou valeur absolue, ...)
  - Classification : $L(\hat{y}, y) = 0$ si $\hat{y} = y$ et 1 sinon (ou perte logistique, ...)

Risques

- ***Risque empirique*** (de la règle de décision f) : perte moyenne sur l'échantillon

  Accessible, base de l'entraînement des modèles d'apprentissage supervisé

  Implicitement lié à la vraisemblance par indépendance

$$\hat{R}_n(f) = \frac{1}{n} \sum_{k=1}^{n} L(f(x_k), y_k)$$

- (vrai) ***Risque*** (de la règle de décision f) : espérance de la perte contre la loi des données

  Lois inconnue à priori, objectif ultime de l'apprentissage

  Ne dépend pas directement de l'échantillon, mais si $f = f_n$ dépend de l'échantillon, alors oui

$$R(f) = \mathbb{E}_{(X,Y)\sim P}\left[L(f(X), Y)\right] = \int L(f(x), y) \mathrm{d}P(x, y)$$

Sommaire

Exercice

- Cas simple : $\mathcal{X} = \emptyset$! (Pas de variable explicative)
- Données $(y_k)_{1 \leq k \leq n}$ iid. suivant une loi P sur $\mathbb{R}$ de variance finie (variable continue)
- Règle de prédiction : $\hat{y} \in \mathbb{R}$ et perte quadratique

1) Quelle est la règle optimale pour le risque empirique ?

2) Quelle est la règle optimale pour le vrai risque ?

3) Comparer

Exercice

- Cas simple : $\mathcal{X} = \emptyset$! (Pas de variable explicative)
- Données $(y_k)_{1 \leq k \leq n}$ iid. suivant une loi P sur $\mathbb{R}$ de variance finie (variable continue)
- Règle de prédiction : $\hat{y} \in \mathbb{R}$ et perte quadratique

1) Quelle est la règle optimale pour le risque empirique ?

$$\min_{\hat{y} \in \mathbb{R}} \hat{R}_n(\hat{y}) = \min_{\hat{y} \in \mathbb{R}} \frac{1}{n} \sum_{k=1}^{n} (\hat{y} - y_k)^2$$

2) Quelle est la règle optimale pour le vrai risque ?

$$\min_{\hat{y} \in \mathbb{R}} R(\hat{y}) = \min_{\hat{y} \in \mathbb{R}} \mathbb{E}_{Y \sim P} \left[(\hat{y} - Y)^2 \right]$$

3) Comparer

Règle optimale : Erreur de Bayes – Régression

Théorème : Règle de Bayes pour la régression

Soit un couple de v.a. (X, Y) distribués selon une lois P de variance finie sur $\mathcal{X} \times \mathcal{Y}$. Avec $\mathcal{Y} = \mathbb{R}$ et la perte quadratique, la meilleure règle possible (au sens du risque) est la *règle de Bayes* :
pour tout $x \in \mathcal{X}$

$$f^*(x) = \mathbb{E}[Y|X = x] = \min_{y \in \mathbb{R}} \mathbb{E}\left[(Y - y)^2 | X = x\right]$$

Problème: on ne connaît *jamais* P, donc on ne peut pas la calculer

Attention: le risque de la règle de Bayes n'est pas nul ! On ne peut pas tout prédire parfaitement

Exercice

- Cas simple : $\mathcal{X} = \emptyset$! (Pas de variable explicative)
- Données $(y_k)_{1 \leq k \leq n}$ iid. suivant une loi P sur $\{0, 1\}$ (variable binaire)
- Règle de prédiction : $\hat{y} \in \{0, 1\}$ et perte $0/1$

1) Quelle est la règle optimale pour le risque empirique ?

2) Quelle est la règle optimale pour le vrai risque ?

3) Comparer

Exercice

- Cas simple : $\mathcal{X} = \emptyset$! (Pas de variable explicative)
- Données $(y_k)_{1 \leq k \leq n}$ iid. suivant une loi P sur $\{0, 1\}$ (variable binaire)
- Règle de prédiction : $\hat{y} \in \{0, 1\}$ et perte $0/1$

1) Quelle est la règle optimale pour le risque empirique ?

$$\min_{\hat{y} \in \mathbb{R}} \hat{R}_n(\hat{y}) = \min_{\hat{y} \in \mathbb{R}} \frac{1}{n} \sum_{k=1}^{n} \mathbb{1}_{\hat{y} \neq y_k}$$

2) Quelle est la règle optimale pour le vrai risque ?

$$\min_{\hat{y} \in \mathbb{R}} R(\hat{y}) = \min_{\hat{y} \in \mathbb{R}} \mathbb{E}_{y \sim P} \mathbb{1}_{\hat{y} \neq y} = \min_{\hat{y} \in \mathbb{R}} \mathbb{P}_{y \sim P}[\hat{y} \neq y]$$

3) Comparer

Règle optimale : Erreur de Bayes – Classification

Théorème : Règle de Bayes pour la classification

Soit un couple de v.a. (X, Y) distribués selon une lois P de variance finie sur $\mathcal{X} \times \mathcal{Y}$. Avec $\mathcal{Y} = \{0, 1\}$ et la perte 0/1, la meilleure règle possible (au sens du risque) est la *règle de Bayes* : pour tout $x \in \mathcal{X}$, avec $\eta(x) = \mathbb{P}[Y = 1|X = x]$,

$$f^*(x) = \mathbb{1}\{\eta(x) > 1/2\} = \min_{y \in \{0,1\}} \mathbb{P}\left[Y \neq y | X = x\right]$$

Problème: on ne connaît *jamais* P, donc on ne peut pas la calculer

Attention: le risque de la règle de Bayes n'est pas nul ! On ne peut pas tout prédire parfaitement

Sommaire

Prise en compte des données

Données: *échantillon d'apprentissage* $(x_k, y_k)_{1\leq k\leq n}$ iid selon P sur $\mathcal{X}\times\mathcal{Y}$

Règle de décision: à partir de l'échantillon d'apprentissage, construire $f_n : \mathcal{X} \to \mathcal{X}$

Attention Formellement, un algorithme d'apprentissage est *une fonction des données et de x*

$$h_n\colon (x, x_1, y_1, \ldots, x_n, y_n) \mapsto y \in \mathcal{Y}$$

que l'on note de manière concise $f_n\colon x \mapsto h_n(x, x_1, y_1, \ldots, x_n, y_n)$

Modèle : Préciser la fonction h_n (`estimator` pour scikit-learn)
Entrainer: Étant donné $(x_k, y_k)_{1\leq k\leq n}$, construire $f_n\colon x \mapsto h_n(x, x_1, y_1, \ldots, x_n, y_n)$ (`fit`)
Prédire: Après entraînement, étant donné $x \in \mathcal{X}$, évaluer $f_n(x) = h_n(x, x_1, y_1, \ldots, x_n, y_n)$ (`predict`)

Exemples

Échantillon d'entraînement dans $(x_k, y_k)_{1 \leq k \leq n}$ dans $\mathbb{R}^d \times \mathbb{R}$

Soient

- h_n est l'algorithme des k plus proches voisins
- h'_n est l'algorithme de régression linéaire

Que se passe du point de vue calcul numérique pour ces deux algorithmes

- A l'entraînement?
- A la prédiction?

Minimisation du risque en espérance

- En apprentissage supervisé, les règles de décision f_n dépendent des données
  - Les données sont des ***réalisation d'un processus aléatoire***
  - L'***aléa des données*** n'est pas pris en compte dans le risque
- Il est donc intéressant d'analyser l'espérance du risque sur les données d'entraînement
- On va se concentrer sur le cas régression ($\mathcal{Y} = \mathbb{R}$ et perte quadratique)

$$
\begin{aligned}
R_n(f_n) &= \mathbb{E}_{(X_1,Y_1),\dots,(X_n,Y_n)}\left[\mathbb{E}_{(X,Y)}\left[(Y - f_n(X))^2\right]\right] \\
&= \mathbb{E}_{(X_1,Y_1),\dots,(X_n,Y_n)}\left[\mathbb{E}_{(X,Y)}\left[(Y - h_n(X, X_1, Y_1, \dots, X_n, Y_n))^2\right]\right]
\end{aligned}
$$

Décomposition du risque : étape 1

Pour un $x \in \mathcal{X}$ et un échantillon fixés, le (vrai) risque de la règle apprise f_n se décompose

$$\begin{aligned}
\mathbb{E}_Y\left[(Y - f_n(X))^2 \,|X = x\right] &= \mathbb{E}_Y\left[(Y - f^*(X) + f^*(X) - f_n(X))^2 \,|X = x\right] \\
&= \mathbb{E}_Y\left[(Y - f^*(X))^2 \,|X = x\right] + \mathbb{E}_Y\left[(f^*(X) - f_n(X))^2 \,|X = x\right] \\
&\quad + 2\mathbb{E}_Y\left[(Y - f^*(X))(f^*(X) - f_n(X))|X = x\right] \\
&= \underbrace{\mathbb{E}_Y\left[(Y - f^*(X))^2 \,|X = x\right]}_{\sigma^2(x)} + (f^*(x) - f_n(x))^2
\end{aligned}$$

où $f^*(x) = \mathbb{E}[Y|X = x]$ est la décision de Bayes et $\sigma^2(x)$ est l'erreur de Bayes au point x

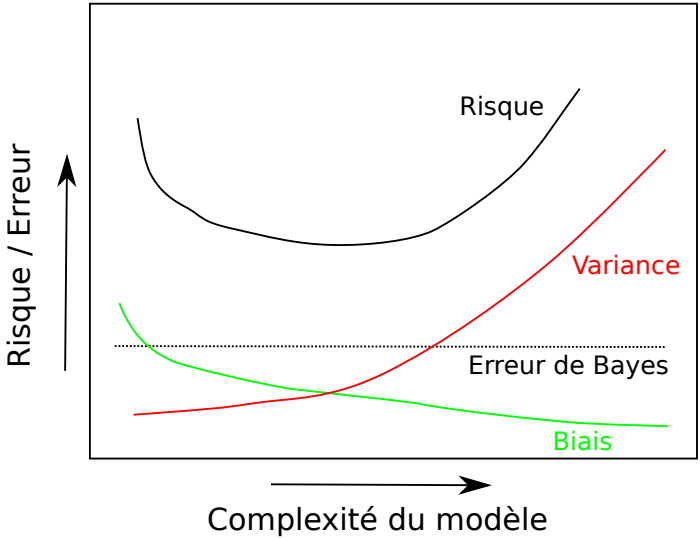
Décomposition du risque : étape 2

Prenons l'espérance sur le tirage de l'échantillon (noté $\mathbb{E}_n$) pour un $x \in \mathcal{X}$ fixé

$$\begin{aligned} R_n(f_n; x) = \mathbb{E}_{(X_1,Y_1),\dots,(X_n,Y_n)} \Big[\mathbb{E}_Y \left[(Y - f_n(X))^2 \,|X = x \right] \Big] \\ &= \sigma^2(x) + \mathbb{E}_n \Big[\left(f^*(x) - f_n(x) \right)^2 \Big] \\ &= \sigma^2(x) + \mathrm{Var}_n \left[f^*(x) - f_n(x) \right] + \Big(\mathbb{E}_n \Big[f^*(x) - f_n(x) \Big] \Big)^2 \\ &= \sigma^2(x) + \mathrm{Var}_n \left[f_n(x) \right] + \Big(\mathrm{Biais}_n \Big[f_n(x) \Big] \Big)^2 \end{aligned}$$

- ***Biais*** quantifie la différence entre f_n et f^* en moyenne
- ***Variance*** quantifie la dispersion f_n due au caractère aléatoire de l'échantillon
- Le risque en espérance est toujours supérieur à $\sigma^2(x)$ (erreur de Bayes)
- Plus le modèle est complexe (plus de "paramètres" à estimer) plus le biais devient petit mais la variance augmente ! C'est le compromis ***biais-variance*** (plus complexe ≠ meilleur)

Compromis biais-variance

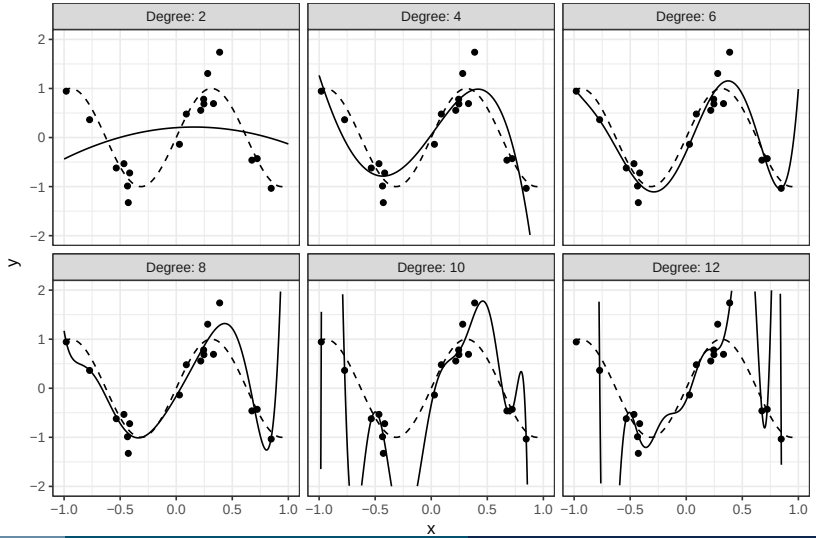


Exemple: régression polynomiale

échantillon 1

$$\min_{P \in \mathbb{R}_d[x]} \sum_{i=1}^{n} (P(x_i) - y_i)^2$$

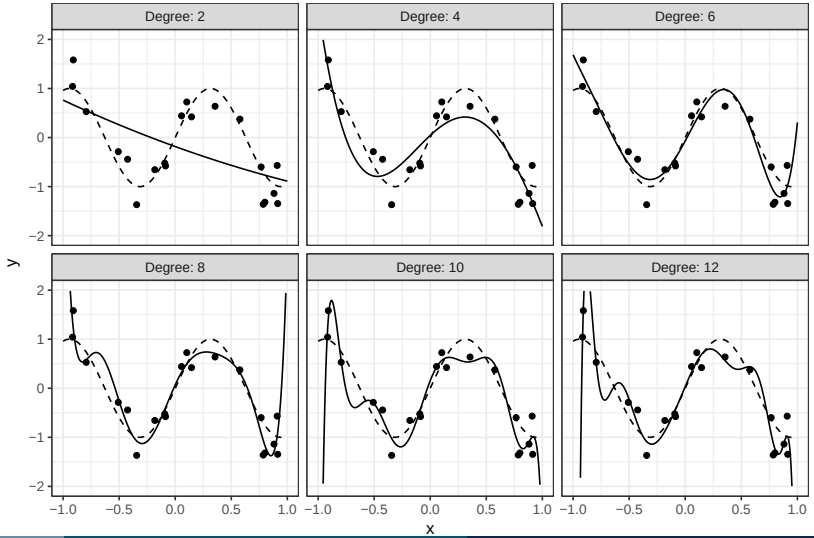
Comment résout-on ce problème ?



Exemple: régression polynomiale

échantillon 2

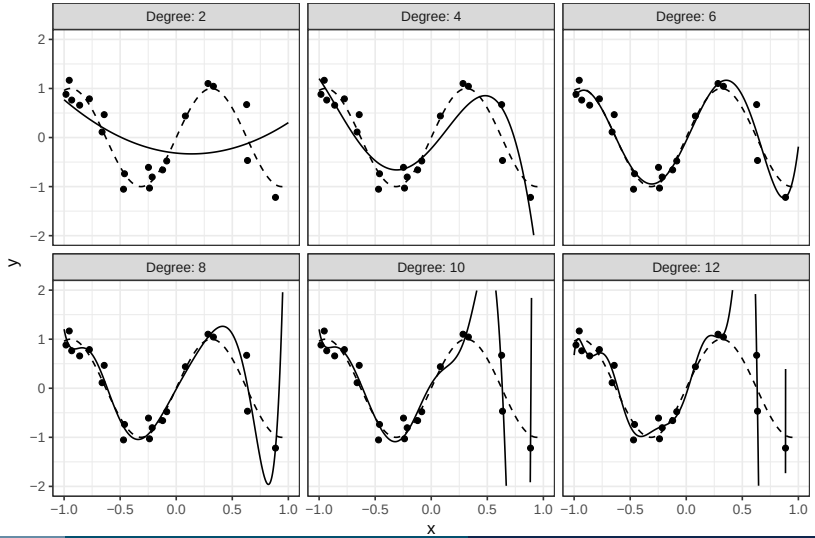
$$\min_{P \in \mathbb{R}_d[x]} \sum_{i=1}^{n} (P(x_i) - y_i)^2$$



Exemple: régression polynomiale

échantillon 3

$$\min_{P \in \mathbb{R}_d[x]} \sum_{i=1}^{n} (P(x_i) - y_i)^2$$

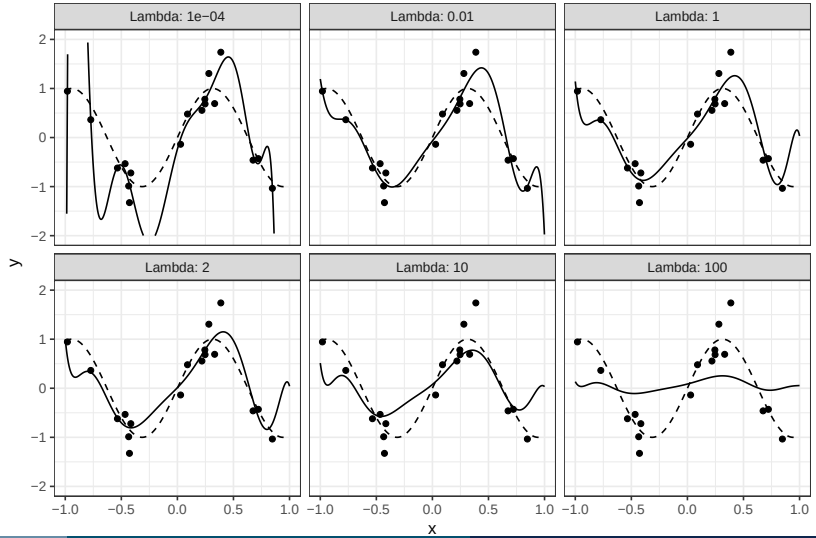


Exemple: régression polynomiale + régularisation

échantillon 1

$$\min_{P \in \mathbb{R}_{10}[x]} \sum_{i=1}^{n} (P(x_i) - y_i)^2 + \lambda \|P\|^2$$

où $\|P\|^2$ est la somme des coefficients de P au carré.
Comment résout-on ce problème ?

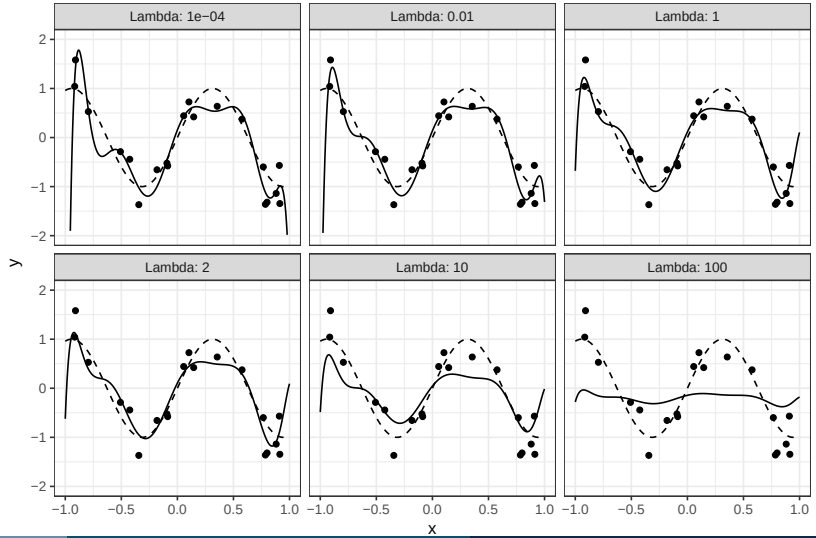


Exemple: régression polynomiale + régularisation

échantillon 2

$$\min_{P \in \mathbb{R}_{10}[x]} \sum_{i=1}^{n} (P(x_i) - y_i)^2 + \lambda \|P\|^2$$

où $\|P\|^2$ est la somme des coefficients de P au carré.
Comment résout-on ce problème ?

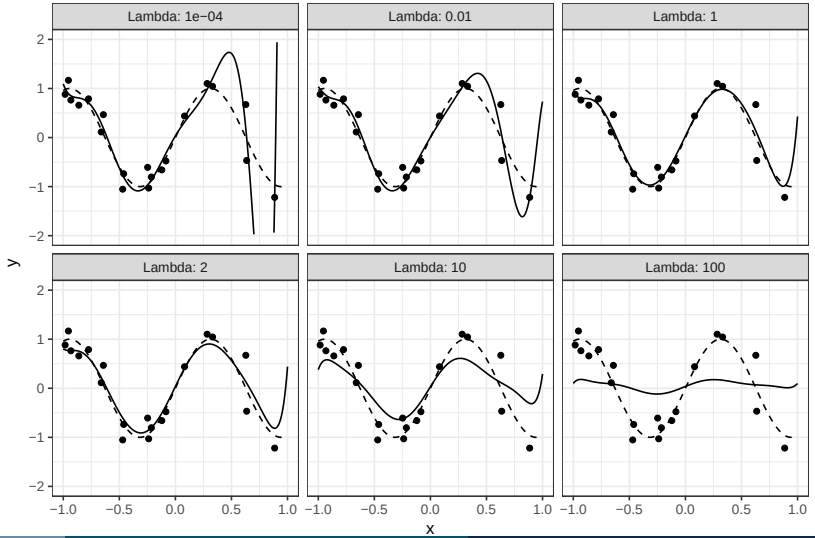


Exemple: régression polynomiale + régularisation

échantillon 3

$$\min_{P \in \mathbb{R}_{10}[x]} \sum_{i=1}^{n} (P(x_i) - y_i)^2 + \lambda \|P\|^2$$

où $\|P\|^2$ est la somme des coefficients de P au carré.
Comment résout-on ce problème ?



Exercice : k plus proches voisins

Soit $Y = f^*(X) + \varepsilon$ avec ε une variable aléatoire indépendante de moyenne nulle et de variance σ^2 .

1) Que vaut $E[Y|X = x]$? Que vaut $\sigma^2(x)$?

On fixe les $(x_i)_{1\leq i\leq n}$ et on considère un échantillon $(y_i)_{1\leq i\leq n}$ distribués selon la loi de $Y|X$ et on fixe $k \in \mathbb{N}$, la prédiction des k plus proches voisins est $f_n \colon x \mapsto \frac{1}{k}\sum_{i=1}^{k} y_{\ell_i}$ où $\ell_1, .., \ell_k$ sont les indices des k plus proches voisins de x

2) Que vaut la variance $\mathrm{Var}_n(f_n(x))$? Que vaut le biais $(\mathbb{E}_n[f_n(x) - f^*(x)])^2$?

3) Comment se comporte le risque en espérance en fonction de k ?

Exercice : k plus proches voisins

Soit $Y = f^*(X) + \varepsilon$ avec ε une variable aléatoire indépendante de moyenne nulle et de variance σ^2 .

1) Que vaut $E[Y|X = x]$? Que vaut $\sigma^2(x)$?

On fixe les $(x_i)_{1\leq i\leq n}$ et on considère un échantillon $(y_i)_{1\leq i\leq n}$ distribués selon la loi de $Y|X$ et on fixe $k \in \mathbb{N}$, la prédiction des k plus proches voisins est $f_n \colon x \mapsto \frac{1}{k}\sum_{i=1}^{k} y_{\ell_i}$ où $\ell_1, .., \ell_k$ sont les indices des k plus proches voisins de x

2) Que vaut la variance $\mathrm{Var}_n(f_n(x))$? Que vaut le biais $(\mathbb{E}_n[f_n(x) - f^*(x)])^2$?

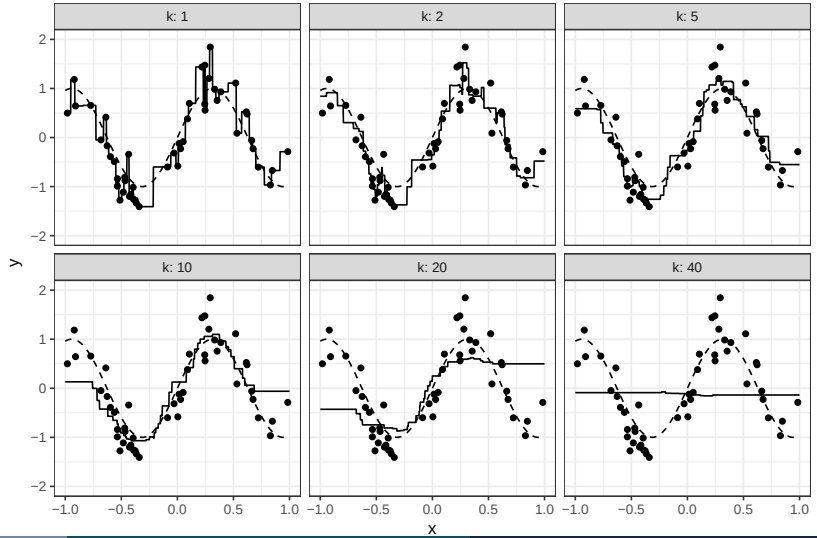
3) Comment se comporte le risque en espérance en fonction de k ?

$$R_n(f_n; x) = \sigma^2 + \left(f^*(x) - \frac{1}{k}\sum_{i=1}^{k} f^*(x_{\ell_i})\right)^2 + \frac{\sigma^2}{k},$$

Exemple : k plus proches voisins

$$f_n : x \mapsto \frac{1}{k} \sum_{i=1}^{k} y_{\ell_i},$$

où $\ell_1, .., \ell_k$ sont les indices des k plus proches voisins de x dans l'échantillon de données



Sommaire

La fin du compromis biais-variance?

- Discussion! https://www.argmin.net/p/overfitting-to-theories-of-overfitting