Data Science

Apprentissage statistique

Hachem Kadri

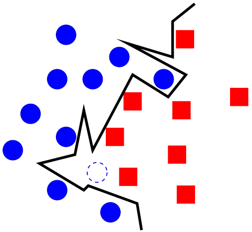
hachem.kadri@univ-amu.fr

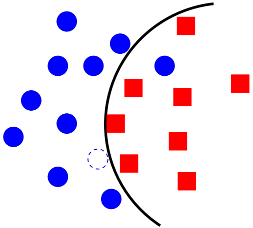
2019-2020

Performance de généralisation

- compromis adéquation aux données d'apprentissage et complexité

  $\longrightarrow$ le modèle ne doit pas être trop complexe pour se généraliser aux données de test (futures)





Apprentissage supervisé / non-supervisé / semi-supervisé

- **Apprentissage supervisé:** apprentissage avec instructeur

  $\longrightarrow$ Données: n exemples d'apprentissage étiquetés $\{(\mathbf{x_1}, \mathbf{y_1}), \ldots, (\mathbf{x_n}, \mathbf{y_n})\}$

- **Apprentissage non-supervisé:** apprentissage sans instructeur

  $\longrightarrow$ Données: m exemples non-étiquetés $\{\mathbf{x_1}, \ldots, \mathbf{x_m}\}$

- **Apprentissage semi-supervisé:**

  $\longrightarrow$ Données: un ensemble de n examples étiquetés et m exemples non-étiquetés $\{(\mathbf{x_1}, \mathbf{y_1}), \ldots, (\mathbf{x_n}, \mathbf{y_n})\} \bigcup \{\mathbf{x_1}, \ldots, \mathbf{x_m}\}$

Apprentissage statistique

- $X \times Y$ espace probabilisé avec une mesure de probabilité P
- $L : Y, Y \to [0, \infty)$ une fonction coût ou perte

La fonction risque (ou erreur) : espérance mathématique de la fonction de perte.

$$R(f) = \int L(y, f(x)) dP(x, y)$$

Problème : Apprendre $f : X \to Y$

Apprentissage statistique

$$R(f) = \int L(y, f(x)) dP(x, y)$$

Problème : Apprendre $f : X \to Y$

$$\min_{f : X \to Y} R(f)$$

à partir d'un échantillon S : un ensemble fini d'exemples

$$\{(x_1, y_1), \ldots, (x_l, y_l)\}$$

i.i.d. selon P.

étant donné un échantillon $S = \{(x_1, y_1), \ldots, (x_l, y_l)\}$, trouver une fonction f qui minimise le risque $R(f)$

Remarques

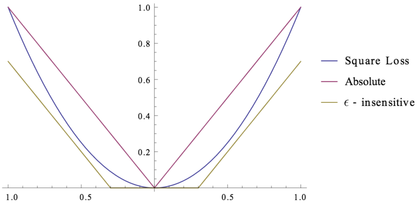
1. Le modèle est non déterministe :
   - le problème cible peut être réellement non déterministe ;
   - le problème peut être bruité ;
   - l'espace de descriptions peut ne décrire qu'incomplètement une situation complexe.
2. Le problème est non déterministe mais on en cherche une solution déterministe.
3. Le modèle est non paramétrique : aucun modèle spécifique de génération de données n'est présupposé ; aucune contrainte sur l'ensemble des fonctions que l'on doit considérer ni sur le type de dépendances entre fonctions et paramètres.
4. D'autres fonctions de pertes peuvent être considérées. En particulier, on peut envisager des coûts différents selon les erreurs commises.

Fonctions coût pour la régression

$$L : Y, Y \to [0, \infty)$$

$L(y, f(x))$: coût ou perte de la prédiction de $f(x)$ à la place de y

- perte quadratique : $L(y, y') = (y - y')^2$
- perte valeur absolue : $L(y, y') = |y - y'|$
- perte ϵ-sensitive : $L(y, y') = \max(|y - y'| - \epsilon, 0)$

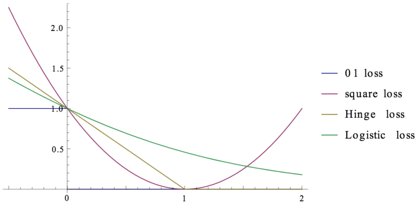


Fonctions coût pour la classification

$$L : Y, Y \rightarrow [0, \infty)$$

$L(y, f(x))$: coût ou perte de la prédiction de $f(x)$ à la place de y

- perte 0-1 : $L(y, y') = \Theta(-yy')$, avec $\Theta(a) = 1$, si $a > 0$ et 0 sinon
- perte quadratique : $L(y, y') = (y - y')^2$
- perte hinge : $L(y, y') = \max(1 - yy', 0)$
- perte logistique : $L(y, y') = \log(1 + \exp(-yy'))$



Cas de la classification

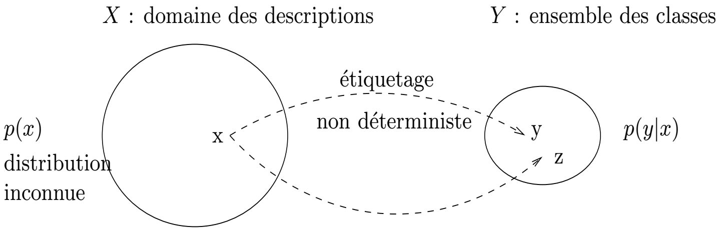
Classifieur : $f : X \to Y$, avec Y un ensemble fini de classes
Fonction de perte (loss function)

$$L(y, f(x)) = \begin{cases} 0 \text{ si } y = f(x) \\ 1 \text{ sinon.} \end{cases}$$

La fonction risque

$$R(f) = \int L(y, f(x)) dP(x, y) = \int_{y \neq f(x)} dP(x, y) = P(y \neq f(x)).$$

Cas de la classification



$$S = \{(x_1, y_1), \ldots, (x_l, y_l)\} \text{ tiré selon } p(x, y) = p(x)p(y|x)$$

Objectif : trouver $f : X \to Y$ dont l'erreur $R(f) = P(y \neq f(x))$ soit la plus petite possible.

Quelques règles de classification

- La règle majoritaire: pour toute nouvelle instance, retourner la classe y_{maj} majoritaire, c'est-à-dire pour laquelle $P(y)$ est maximum : pour tout $x \in X$,

$$f_{maj}(x) = ArgMax_y P(y) = y_{maj} \text{ et } R(f_{maj}) = 1 - P(y_{maj}).$$

- La règle du maximum de vraisemblance (*maximum likelihood*): retourner pour chaque instance x la classe y pour laquelle x est la valeur la plus observée.

$$f_{mv}(x) = ArgMax_y P(x|y).$$

- La règle de Bayes: retourner pour chaque instance x, la classe y dont l'observation est la plus probable, ayant observé x.

$$f_B(x) = ArgMax_y P(y|x).$$

Exemple

Une banque souhaite proposer une offre commerciale à certains de ses clients : une carte permettant de régler de manière sécurisée des achats sur Internet. Comment cibler les clients le plus susceptibles d'êtres intéressés ? Lorsqu'elle leur demande d'indiquer leur coordonnées, certains indiquent spontanément une adresse e-mail : c'est peut-être un critère sur lequel baser le mailing. Un sondage réalisé sur un échantillon supposé représentatif de sa clientèle, indique que

- 40% sont intéressés dont 80% ayant indiqué leur e-mail,
- 60% ne sont pas intéressés dont 40% ayant indiqué leur e-mail.

Modèle :

$X = \{email, \overline{email}\}$, $Y = \{interesse, \overline{interesse}\}$,

$P(email) = 0.8 \times 0.4 + 0.6 \times 0.4 = 0.56$, $P(interesse|email) = 4/7$ et $P(interesse|\overline{email}) = 8/44 = 2/11$.

Exemple (suite)

- La majorité des clients n'est pas intéressée par l'offre. Donc,

$$f_{maj}(x) = \overline{interesse} \text{ et } R(f_{maj}) = 0.4.$$

- Comme $P(email|interesse) = 0.8 > P(email|\overline{interesse})$, $f_{mv}(email) = interesse$.

  Et comme
  $P(\overline{email}|interesse) = 1 - P(email|interesse) < P(\overline{email}|\overline{interesse})$, $f_{mv}(\overline{email}) = \overline{interesse}$.

$$R(f_{mv}) = 0.4 \times 0.2 + 0.6 \times 0.4 = 0.24.$$

- Comme $P(interesse|email) = 32/56 > 1/2$ et $P(interesse|\overline{email}) = 8/44$, la règle de Bayes conduit au même classifieur que la règle du maximum de vraisemblance.

Optimalité de la règle de Bayes

Théorème : La règle de décision de Bayes est la règle de risque minimal.

- Le risque de Bayes n'est nul que pour des problèmes déterministes.
- **Pb.** Le règle de décision de Bayes est le plus souvent inaccessible !

Cas de la régression

La variable y prend des valeurs continues.
Fonction de perte : l'*écart quadratique* défini par

$$L(y, f(x)) = (y - f(x))^2.$$

Le risque ou l'erreur d'une fonction f : l'*écart quadratique moyen* défini par :

$$R(f) = \int_{X \times Y} (y - f(x))^2 dP(x, y).$$

Théorème : La fonction $\overline{f}$, moyenne des valeurs observables en x, définie par

$$\overline{f}(x) = \int_Y y dP(y|x)$$

est la fonction de régression de risque minimal.

Cas de l'estimation de densité

- On dispose de réalisations indépendantes $x_1, \ldots, x_l$ de X.
- On cherche à estimer $P(x)$ pour tout x.
- On cherche une fonction $P' : X \to [0,1]$ qui approche P (ou sa densité dans le cas continu) au mieux.
- Fonction de perte : $L(x,y) = -\log y$
- Fonction de risque :

$$R(P') = \sum_{x \in X} -P(x) \cdot \log P'(x); R(f) = \int_X -\log f(x) dP(x).$$

Théorème : $R(P')$ est minimal pour $P' = P$ (cas discret).

L'apprentissage en pratique

- On dispose d'un échantillon S qu'on suppose i.i.d.
- On recherche une fonction f de *classification*, de *régression* ou de *densité* dont le risque $R(f)$ soit le plus faible possible.
- Il existe toujours une meilleure solution f_{min} ... inaccessible !

L'apprentissage en pratique

- Dans la pratique, on cherche une solution dans un ensemble de fonctions $\mathcal{F}$ particulier : *arbres de décision*, *réseaux de neurones*, *fonctions linéaires*, *modèles de Markov cachés*, etc.

- Soit f_{opt} une fonction de risque minimal dans $\mathcal{F}$

- L'ensemble $\mathcal{F}$ doit avoir deux qualités :

  1. contenir des fonctions f_{opt} dont le risque n'est pas trop éloigné de $R(f_{min})$
  2. permettre d'approcher un classifieur de risque minimal f_{opt} au moyen des informations dont on dispose.

Principe de minimisation du risque empirique

- Une idée naturelle : sélectionner une fonction dans $\mathcal{F}$ qui décrit au mieux les données de l'échantillon d'apprentissage.

- Le *risque empirique* $R_{emp}(f)$ d'une fonction f sur l'échantillon $S = \{(x_1, y_1), \ldots (x_l, y_l)\}$ est la moyenne de la fonction de perte calculée sur S :

$$R_{emp}(f) = \frac{1}{l} \sum_{i=1}^{l} L(y_i, f(x_i)).$$

- $R_{emp}(f)$ est une estimation du risque réel $R(f)$ de f.

Principe de minimisation du risque empirique

en classification : $R_{emp}(f)$ est la moyenne du nombre d'erreurs de prédiction de f sur les éléments de S :

$$R_{emp}(f) = \frac{Card\{i | f(x_i) \neq y_i\}}{l}.$$

en régression : $R_{emp}(f)$ est la moyenne des carrés des écarts à la moyenne de f sur S :

$$R_{emp}(f) = \frac{1}{l} \sum_{i=1}^{l} (y_i - f(x_i))^2.$$

en estimation de densité : $R_{emp}(f)$ est l'opposé de la log-vraisemblance de S :

$$R_{emp}(f) = \frac{1}{l} \sum_{i=1}^{l} -\log f(x_i) = \frac{-1}{l} \log \prod_{i=1}^{l} f(x_i).$$

Principe de minimisation du risque empirique

Le *principe inductif de minimisation du risque empirique (ERM)* recommande de

trouver une fonction $f \in \mathcal{F}$ *qui minimise* $R_{emp}(f)$

- en classification, cela revient à minimiser le nombre d'erreurs commises par f sur l'échantillon ;
- en régression, on retrouve la méthode des moindres carrés ;
- en estimation de densité, on retrouve la *méthode du maximum de vraisemblance*.

Principe de minimisation du risque empirique

Soit f_{emp} une fonction minimisant le risque empirique.

$$R(f_{emp}) = R(f_{min}) + [R(f_{opt}) - R(f_{min})] + [R(f_{emp}) - R(f_{opt})]$$

$R(f_{min})$: incompressible, donne une mesure de la difficulté intrinsèque du problème, du volume de bruit qu'il comporte.

$R(f_{opt}) - R(f_{min})$: mesure l'adéquation de $\mathcal{F}$ au problème considéré.

$R(f_{emp}) - R(f_{opt})$: représente l'erreur liée au principe de minimisation du risque empirique.

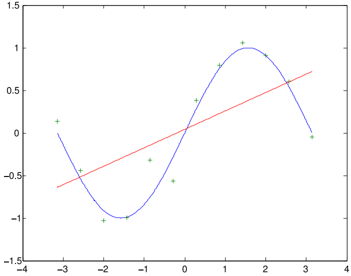
On dit que le principe ERM est *consistant* dans la classe $\mathcal{F}$ si f_{emp} tend vers f_{opt} lorsque le nombre d'exemples tend vers l'infini.

Exemple

- Estimation de la régression à l'aide de fonctions polynômes.

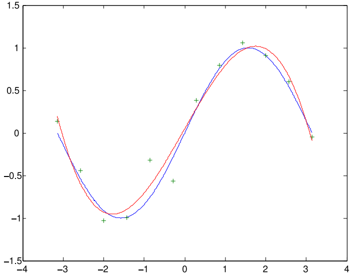
Estimation de la régression à l'aide de fonctions polynômes

- 11 points sur la courbe $x \longmapsto \sin(x)$ avec un bruit additif normal d'écart-type 0.2/
- En bleu: la courbe $x \longmapsto \sin(x)$
- En rouge: le polynôme de degré **1** qui minimise le risque empirique quadratique



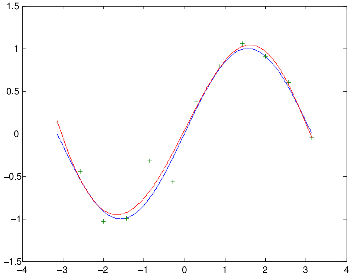
Estimation de la régression à l'aide de fonctions polynômes

- 11 points sur la courbe $x \longmapsto \sin(x)$ avec un bruit additif normal d'écart-type 0.2/
- En bleu: la courbe $x \longmapsto \sin(x)$
- En rouge: le polynôme de degré **3** qui minimise le risque empirique quadratique



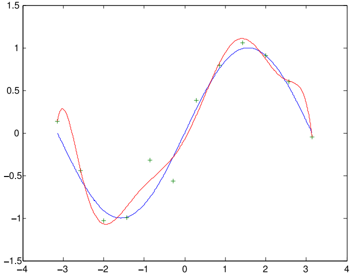
Estimation de la régression à l'aide de fonctions polynômes

- 11 points sur la courbe $x \longmapsto \sin(x)$ avec un bruit additif normal d'écart-type 0.2/
- En bleu: la courbe $x \longmapsto \sin(x)$
- En rouge: le polynôme de degré **5** qui minimise le risque empirique quadratique



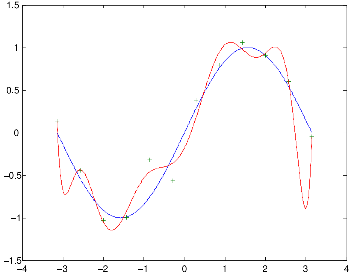
Estimation de la régression à l'aide de fonctions polynômes

- 11 points sur la courbe $x \longmapsto \sin(x)$ avec un bruit additif normal d'écart-type 0.2/
- En bleu: la courbe $x \longmapsto \sin(x)$
- En rouge: le polynôme de degré **7** qui minimise le risque empirique quadratique



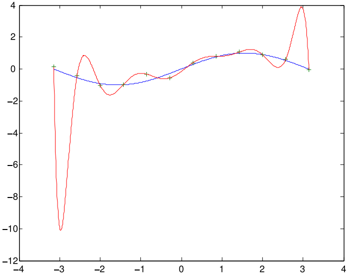
Estimation de la régression à l'aide de fonctions polynômes

- 11 points sur la courbe $x \longmapsto \sin(x)$ avec un bruit additif normal d'écart-type 0.2/
- En bleu: la courbe $x \longmapsto \sin(x)$
- En rouge: le polynôme de degré **9** qui minimise le risque empirique quadratique

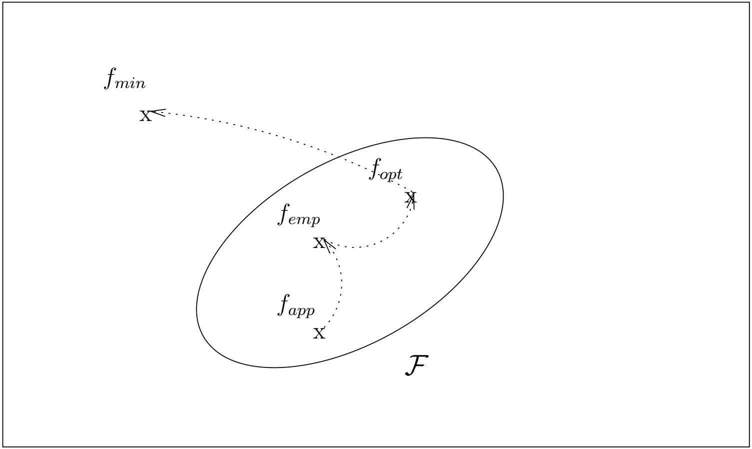


Estimation de la régression à l'aide de fonctions polynômes

- 11 points sur la courbe $x \longmapsto \sin(x)$ avec un bruit additif normal d'écart-type 0.2/
- En bleu: la courbe $x \longmapsto \sin(x)$
- En rouge: le polynôme de degré **11** qui minimise le risque empirique quadratique



Niveaux de difficultés en apprentissage



Niveaux de difficultés en apprentissage

Il y a donc au moins quatre raisons pour lesquelles une méthode d'apprentissage appliquée à un problème particulier peut ne pas donner de résultats satisfaisants :

- la *nature non déterministe du problème*,
- la trop *faible expressivité* de l'espace fonctionnel $\mathcal{F}$ choisi,
- la *non consistance du principe ERM* ou plus généralement, du principe choisi pour approcher une fonction optimale dans $\mathcal{F}$,
- la *difficulté à minimiser le risque empirique* (ou plus généralement, à mettre en application le principe choisi).