

# Apprentissage: cours 1

## Introduction au cadre de l'apprentissage supervisé

Guillaume Obozinski

13 février 2013

## 1 Introduction générale

### Notations

- $x$  : observation scalaire (ou quelconque si  $\mathbf{x}$  n'est pas utilisé).
- $\mathbf{x}$  : observation vectorielle
- $X$  variable aléatoire scalaire ou vectorielle selon le contexte.
- $\mathbf{X}$  : matrice d'observations

### 1.1 But du cours

Le but du cours n'est pas seulement d'exposer une théorie de l'apprentissage, mais aussi d'introduire un certain nombre de concepts et techniques de mathématiques appliquées qui sont pertinents dans toutes les disciplines où in fine on résout un problème du monde réel avec des données (traitement du signal, statistique, optimisation). Entre autres : problèmes mal posés, optimisation, analyse de données matricielles, traitement du signal, statistiques, etc.

### 1.2 Qu'est-ce que l'apprentissage ?

**But :** *Prédire* une donnée de sortie  $y$  à partir d'une donnée d'entrée  $x$ , ou bien plus généralement produire la *meilleure décision*  $\delta$  à partir d'une donnée d'entrée  $x$  et en vue d'une donnée  $y$  qui n'est pas connue au moment où la décision est prise.

Exemples :

- Classifier automatiquement des images de chiffres manuscrit en leur associant le chiffre écrit. Dans ce cas, pour une image représentée par les niveaux de gris de ses pixels  $x \in [0, 1]^p \subset \mathbb{R}^p$  pour  $p$  pixels, et  $y \in \{0, \dots, 9\}$ .
- A partir de l'image caméra de la trajectoire d'une balle de ping-pong, déterminer les paramètres de commandes de la dynamique d'un robot, pour renvoyer la balle dans les limites du terrain adverse.  $x$  l'image de la balle,  $y$  les paramètres de la cinétique de la balle,  $\delta$  le paramètre de commande
- Etant donné une paire protéine + molécule déterminer si une réaction chimique a lieu.  $x \in \mathbb{R}^p$  et  $y \in \{0, 1\}$ .
- A partir de l'enregistrement d'un morceau de musique, séparer les différents instruments.

**Difficulté :**  $Y$  n'est pas une fonction déterministe de  $X$ .

- Il peut y avoir du bruit e.g.  $Y = f(X) + \varepsilon$ .
- Plus généralement,  $Y = f(X, Z)$  où  $Z$  n'est pas observé.

On peut difficilement faire bien systématiquement.

**Approches possibles :**

- Essayer de faire bien dans le pire cas → approches théorie de jeux, stratégie minimax au coup par coup.
- Essayer de faire bien en moyenne. → objectif de l'apprentissage

**Idée :** Modéliser  $X$  et  $Y$  comme des variables aléatoires. → La meilleure décision “en moyenne” peut être prise à partir de  $\mathbb{P}(Y = \cdot | X = x)$ . Mais

- On ne la connaît pas. Faire un modèle simple n'est pas possible.
- $X$  et éventuellement  $Y$  sont des objets de *grande dimension* → le problème de déterminer  $\mathbb{P}(Y = \cdot | X = x)$  est a priori beaucoup plus difficile que le problème initial.

**Information disponible :** des observations de  $X : (x_1, \dots, x_n)$ , ou des observations de  $(X, Y) : (x_1, y_1), \dots, (x_n, y_n)$ .

**Idée :** Apprendre! Utiliser une stratégie qui marche pour un ensemble d'observations existante et *qui puisse se généraliser aux autres observations*

**Formalisation :**

Soit  $D_n := \{(X_1, Y_1), \dots, (X_n, Y_n)\}$  un ensemble de *données d'entraînement*. Les  $X_i$  sont des *variables d'entrées* à valeur dans un ensemble  $\mathcal{X}$ . De même les  $Y_i$  sont des variables sortie à valeur dans un ensemble  $\mathcal{Y}$ . On appelle les  $X_i$  *descripteurs, covariables, régresseurs*, (en anglais *features*) et les  $Y_i$  *étiquettes* (en anglais *labels*).

On fera l'hypothèse dans ce cours, comme souvent en statistiques, que ces données sont i.i.d., c'est-à-dire *indépendantes et identiquement distribuées*.

En pratique :

- les données d'entraînement ne sont pas indépendantes
  - les données de test ne sont pas de la même distribution, souvent pour des raisons de *non-stationarité*.
- Une règle d'apprentissage est une fonction

$$\mathcal{A} : \begin{array}{l} \bigcup_{n \in \mathbb{N}} (\mathcal{X} \times \mathcal{Y})^n \rightarrow \mathcal{F} \\ D_n \quad \quad \quad \mapsto \hat{f} \end{array}$$

La fonction  $\hat{f}$  est construite en vue d'être utilisée pour prédire  $X$  et  $Y$  où  $(X, Y)$  est une paire de *données de test*.

Distinguer *phase d'apprentissage* et *phase de test*

## 2 Apprentissage supervisé

### 2.1 Approche algorithmiques en apprentissage supervisé

- Méthodes par moyennage local : k-ppv, Nadaraya-Watson, fenêtres de Parzen, arbres de décisions
- Méthodes par minimisation du risque empirique : modèle linéaire, méthodes à noyaux.
- Réseaux de neurones
- Méthodes de modélisation probabilistes (modèle graphiques, méthodes bayésiennes)
- Approche PAC-bayésienne
- Méthodes bayésiennes
- Apprentissage séquentiel

### 2.2 Formalisme de la théorie de la décision

La théorie de la décision formalise les critères qui vont nous permettre d'évaluer la qualité de la décision que nous allons essayer d'apprendre.

- Soit  $(X, Y)$  les variables aléatoire formant une paire de données de test.
- Soit  $\mathcal{A}$  un ensemble d'actions, de décisions ou de prédictions possibles.

- Soit  $\ell : \mathcal{A} \times \mathcal{Y} \rightarrow \mathbb{R}$  une fonction de perte, ou fonction de coût. La fonction de perte spécifie le prix à payer  $\ell(a, y)$  pour avoir pris la décision  $a$  quand la variable de sortie prend la valeur  $y$ .
- Soit une fonction de prédiction  $f : \mathcal{X} \rightarrow \mathcal{A}$

On définit le *risque* associé au problème de décision défini par  $X, Y$  et  $\ell$ , la quantité

$$\mathcal{R}(f) = \mathbb{E}[\ell(f(X), Y)]$$

Au vu de cette définition, un prédicteur optimal est un prédicteur pour lequel le risque est minimal. S'il existe un prédicteur atteignant l'infimum du risque, ce prédicteur  $f^*$  est appelé *fonction cible*, *fonction oracle* et on a

$$f^* = \operatorname{argmin}_{f \in \mathcal{F}} \mathcal{R}(f).$$

On appelle risque conditionnel et on note  $\mathcal{R}(a | X) = \mathbb{E}[\ell(Y, a) | X]$ . Si  $\inf_{a \in \mathcal{A}} \mathcal{R}(a | X)$  est atteint dans  $\mathcal{A}$  pour presque tout  $X$ , on définit le prédicteur de Bayes comme le prédicteur qui minimise le risque conditionnel au sens où :

$$f^*(X) = \operatorname{argmin}_{a \in \mathcal{A}} \mathcal{R}(a | X).$$

Si le prédicteur de Bayes est dans  $\mathcal{F}$  c'est la fonction cible.

**Définition 1.** (Excès de risque) On appelle *excès de risque* la différence entre le risque du prédicteur considéré et le risque de la fonction cible. C'est la quantité

$$\mathcal{R}(\hat{f}) - \mathcal{R}(f^*) = \mathbb{E}[\ell(\hat{f}(X), Y)] - \inf_{f \in \mathcal{F}} \mathbb{E}[\ell(f(X), Y)]$$

**Remarque (Risque d'un prédicteur issu d'une règle d'apprentissage).** Comme le prédicteur de l'ensemble d'apprentissage qui est aléatoire, le risque reste une quantité aléatoire :

$$\mathcal{R}(\hat{f}) = \mathbb{E}[\ell(\hat{f}(X), Y) | D_n]$$

## Exemples

*Exemple 1.* (régression au sens des moindres carrés : perte quadratique)

Cas où  $\mathcal{A} = \mathcal{Y} = \mathbb{R}$ .

- perte :  $\ell(a, y) = \frac{1}{2}(a - y)^2$
- risque :  $\mathcal{R}(f) = \frac{1}{2}\mathbb{E}[(f(X) - Y)^2]$
- fonction cible :  $f^*(X) = \mathbb{E}[Y|X]$

*Exemple 2.* (classification à  $K$ -classes : perte 0-1)

Cas où  $\mathcal{A} = \mathcal{Y} = \{0, \dots, K - 1\}$ .

- perte :  $\ell(a, y) = 1_{\{a \neq y\}}$
- risque :  $\mathcal{R}(f) = \mathbb{P}(f(X) \neq Y)$
- fonction cible :  $f^*(X) = \operatorname{argmax}_k \mathbb{P}(Y = k | X)$

## Application de la régression à la classification binaire

On considère un problème de classification binaire ( $K = 2$ ) avec  $(X, Y)$  une paire de données d'entrée et sortie avec  $X$  à valeur dans  $\mathcal{X} = \mathbb{R}^p$  et  $Y$  à valeurs dans  $\mathcal{Y} = \{0, 1\}$ . Comme  $\{0, 1\} \subset \mathbb{R}$ , on peut aussi considérer le problème de régression au sens des moindres carrés de  $Y$  sur  $X$ . Pour la régression la fonction cible est  $\eta^*(x) = \mathbb{E}[Y|X = x] = \mathbb{P}(Y = 1|X = x)$ . Pour la classification la fonction cible est  $g^*(x) = \operatorname{argmax}_{y \in \mathcal{Y}} \mathbb{P}(Y = y|X = x) = 1_{\{\eta^*(x) \geq \frac{1}{2}\}}$ .

Comme l'objectif de la régression est d'apprendre à prédire la probabilité conditionnelle de  $Y$  sachant  $X$  il paraît naturel de se servir de ce prédicteur pour construire un prédicteur de classification.

**Théorème 1.** Soit  $\eta : \mathbb{R}^n \rightarrow \mathbb{R}$  un prédicteur pour la régression des moindres carrés et  $g_\eta : \mathbb{R}^n \rightarrow \{0, 1\}$  défini par  $g_\eta(x) = 1_{\{\eta(x) \geq \frac{1}{2}\}}$ . Alors le risque pour la régression de  $\eta$  noté  $\mathcal{R}^{\text{reg}}(\eta) := \mathbb{E}[(\eta(X) - Y)^2]$  et le risque pour la classification de  $g_\eta$  noté  $\mathcal{R}^{0-1}(g_\eta) := \mathbb{P}(g_\eta(X) \neq Y)$  sont liés par la relation

$$\mathcal{R}^{0-1}(g_\eta) - \mathcal{R}^{0-1}(g^*) \leq 2\sqrt{\mathcal{R}^{\text{reg}}(\eta) - \mathcal{R}^{\text{reg}}(\eta^*)}.$$

## 2.3 Minimisation du risque empirique

Idée : estimer le risque grâce à l'ensemble d'apprentissage disponible, i.e remplacer la distribution de probabilité  $P_{X,Y}$  par la *distribution empirique*  $P_n = \frac{1}{n} \sum_{i=1}^n \delta_{(x_i, y_i)}$ .

On définit donc le *risque empirique*

$$\widehat{\mathcal{R}}_n(f) = \mathbb{E}_n[\ell(f(X), Y)] = \frac{1}{n} \sum_{i=1}^n \ell(f(x_i), y_i)$$

Le prédicteur de Bayes pour le risque empirique est très inintéressant puisqu'il n'est défini qu'aux  $x_i$  déjà vu. On se restreint donc à une *famille de prédicteurs* ou *espace d'hypothèses*  $S \subset \mathcal{F}$ .

Le *principe de minimisation du risque empirique*

$$\min_{f \in S} \widehat{\mathcal{R}}_n(f) = \min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \ell(f(x_i), y_i).$$

*Exemple 3.* (La régression linéaire)

On considère le cas  $\mathcal{X} = \mathbb{R}^p$ ,  $\mathcal{Y} = \mathbb{R}$  et  $\ell$  est la perte quadratique. On se restreint à des fonctions linéaires de la forme  $f_{\mathbf{w}} : \mathbf{x} \mapsto \mathbf{w}^\top \mathbf{x}$ . L'espace d'hypothèse est donc  $S = \{f_{\mathbf{w}} \mid \mathbf{w} \in \mathbb{R}^p\}$ . On a alors :

$$\widehat{\mathcal{R}}_n(f_{\mathbf{w}}) = \frac{1}{2n} \sum_{i=1}^n (y_i - \mathbf{w}^\top \mathbf{x}_i)^2 = \frac{1}{2n} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2$$

avec  $\mathbf{y}^\top = (y_1, \dots, y_n) \in \mathbb{R}^n$  le vecteur de sorties,  $\mathbf{X} \in \mathbb{R}^{n \times p}$  la *matrice de design*.

Le problème  $\min_{\mathbf{w} \in \mathbb{R}^p} \widehat{\mathcal{R}}_n(f_{\mathbf{w}})$  est résolu par *les équations normales*

$$\mathbf{X}^\top \mathbf{X} \mathbf{w} - \mathbf{X}^\top \mathbf{y} = 0.$$

Problème :  $\mathbf{X}^\top \mathbf{X}$  n'est pas inversible quand  $p > n$ , le prédicteur n'est pas unique.

Si  $\mathbf{X}^\top \mathbf{X}$  est inversible, alors

$$\widehat{f}_S(\mathbf{x}') \mapsto \mathbf{x}'^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$

**Définition 2.** (Consistance par rapport à une loi  $P$ ) Pour des données d'entraînement et de test i.i.d de loi  $P$ , on dit que l'algorithme d'apprentissage est consistant si le prédicteur qu'il définit satisfait

$$\lim_{n \rightarrow \infty} \mathbb{E}[\mathcal{R}(\widehat{f})] - \mathcal{R}(f^*) = 0.$$

**Définition 3.** (Consistance universelle) On dit qu'un algorithme d'apprentissage est universellement consistant s'il est consistant pour toute loi  $P$ .

## 2.4 Phénomène de surapprentissage : exemple de la régression polynomiale

Voir TP de la semaine suivante.

## 2.5 Décomposition du risque

Pour un risque  $\mathcal{R}$  défini par la donnée de v.a.  $X, Y$  et d'une fonction de perte  $\ell$ . On se donne une règle d'apprentissage dont le codomaine est l'*espace d'hypothèse*  $S \subset \mathcal{F}$  on définit :

- la fonction cible  $f^* = \operatorname{argmin}_{f \in \mathcal{F}} \mathcal{R}(f)$
- la meilleure approximation de la fonction cible dans  $S$  :  $f_S^* := \operatorname{argmin}_{f \in S} \mathcal{R}(f)$
- le prédicteur obtenu par la règle d'apprentissage à partir de données  $\widehat{f}_S$

On a la décomposition :

$$\underbrace{\mathcal{R}(\widehat{f}_S) - \mathcal{R}(f^*)}_{\text{excès de risque}} = \underbrace{\mathcal{R}(\widehat{f}_S) - \mathcal{R}(f_S^*)}_{\text{erreur d'estimation}} + \underbrace{\mathcal{R}(f_S^*) - \mathcal{R}(f^*)}_{\text{erreur d'approximation}}$$

## 2.6 Compromis Biais-Variance

Apparentée mais différente de la décomposition précédente, c'est une décomposition de l'*espérance du risque* dans le cas particulier de la perte quadratique. Cette décomposition est plus adaptée aux règles d'apprentissage autre que la minimisation du risque empirique. On a :

$$\mathbb{E}[\mathcal{R}(\hat{f})] = \mathbb{E}[(\hat{f}(X) - Y)^2] = \underbrace{\mathbb{E}[(\hat{f}(X) - \mathbb{E}[\hat{f}(X)|X])^2]}_{\text{variance de } \hat{f}} + \underbrace{\mathbb{E}[(\mathbb{E}[\hat{f}(X)|X] - \mathbb{E}[Y|X])^2]}_{\text{biais de } \hat{f}} + \underbrace{\mathbb{E}[(Y - \mathbb{E}[Y|X])^2]}_{\text{variance du "bruit"}}$$

## 3 Contrôle de la complexité

### 3.0.1 Un problème mal posé

On dit qu'un problème est bien posé au sens de Hadamard si

- Il admet une solution
- Cette solution est *unique*
- La solution dépend de façon continue des paramètres du problème dans une topologie bien choisie.

Problème de l'apprentissage comme un problème essentiellement *mal posé* car sous-contraint et disposant d'information par essence incomplète.

### 3.0.2 Le fléau de la dimension

Le conditionnement du problème se dégrade de façon exponentielle avec la dimension. Exemple de l'estimation de densité. Autre exemple dans le cours sur les méthodes de moyennage.

### 3.0.3 Espace d'hypothèse et régularisation

- Contrôle explicite de la complexité : degré du polynôme, choix de la largeur de bande pour les méthodes de lissage, choix des variables, etc → problème de choix de l'espace d'hypothèse  $S$ .
- Contrôle implicite de la complexité pour les méthodes par minimisation du risque empirique : *régularisation de Tikhonov*.

Le principe de la régularisation est de pénaliser la valeur d'une norme  $f \mapsto \|f\|$  qui "contrôle la complexité" de la fonction  $f$ .

$$\min_{f \in S} \hat{\mathcal{R}}_n(f) + \lambda \|f\|^2$$

exemple : norme hilbertienne, norme  $\ell_q$ , norme de Sobolev.

La régularisation induit un compromis entre la minimisation du risque empirique et le choix d'une fonction trop complexe. Elle a l'avantage que la complexité de la fonction ne doit pas être connue à l'avance. Le compromis est contrôlé par  $\lambda$  le *paramètre de régularisation* ou *hyperparamètre*. Il fait néanmoins choisir  $\lambda$  → problème analogue au problème de sélection de modèle.

### 3.0.4 Régression ridge

Forme de régularisation la plus classique en statistiques.

$$\min_{\mathbf{w} \in \mathbb{R}^p} \frac{1}{2n} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 + \lambda \|\mathbf{w}\|_2^2$$

Grâce à la régularisation, le problème est devenu fortement convexe. Donc la solution est unique :

$$\hat{\mathbf{w}}^{(\text{ridge})} = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^\top \mathbf{y}$$

Effet de lissage du spectre de la matrice de design. Notion de shrinkage. La régularisation a pour effet de transformer le problème en un problème bien posé au sens de Hadamard. Le paramètre de régularisation contrôle le conditionnement de la matrice.