Apprentissage: cours 1
Introduction au cadre de l'apprentissage supervisé

Simon Lacoste-Julien

18 septembre 2015

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 supervisé (idée générale) ?

But : À partir de *données d'entraînement*, on veut apprendre une *loi de prédiction* pour : *prédire* une donnée de sortie y à partir d'une donnée d'entrée x, ou bien plus généralement produire *la meilleure action* a 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 $y \in \{0, \ldots, 9\}$.
- A partir d'un vidéo 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 le vidéo de la trajectoire de la balle, y les paramètres de la cinétique de la balle, a le paramètre de commande
- Étant donné une paire protéine + molécule déterminer si une réaction chimique a lieu. $x \in \mathbb{R}^p$ et $y \in \{0, 1\}$.
- À partir de l'enregistrement d'un morceau de musique, séparer les différents instruments : x le signal audio ; y la liste de signaux audio pour 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.
- La fonction f peut être très compliquée ; et est inconnue.

On peut difficilement faire bien systématiquement.

Approches possibles :
— Essayer de faire bien dans le pire cas $\rightarrow$ approches théorie de jeux, stratégie minimax au coup par coup.
— Essayer de faire bien en moyenne. $\rightarrow$ objectif de l'apprentissage statistique

Idée : Modéliser X et Y comme des variables aléatoires. $\rightarrow$ La meilleure décision "en moyenne" peut être prise à partir de $\mathbb{P}(Y = \cdot|X = x)$. Mais
— On ne la connait pas. Faire un modèle simple n'est pas possible.
— X et éventuellement Y sont des objets de *grande dimension* $\rightarrow$ 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 : $(x_1, \ldots, x_n)$, ou des observations de (X, Y) : $(x_1, y_1), \ldots, (x_n, y_n)$.

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

Formalisation :
Soit $D_n := \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ 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* (cette hypothèse sera cependant partiellement relaxée dans le dernier cours).

En pratique :
— les données d'entraînement ne sont pas totalement indépendantes
— les données de test ne sont pas exactement de la même distribution, souvent pour des raisons de *non-stationnarité*.

Une *règle d'apprentissage de prédiction* est une fonction $\mathscr{A}$ qui associe à des données d'entraînements D_n une fonction de prédiction $\widehat{f}_n$ (le chapeau sur f est pour indiquer que c'est une *estimation* de fonction) :

$$\begin{array}{rccc} \mathscr{A} : & \bigcup_{n\in\mathbb{N}} (\mathcal{X} \times \mathcal{Y})^n & \rightarrow & \mathcal{F} \\ & D_n & \mapsto & \widehat{f}_n \end{array}$$

La fonction estimée $\widehat{f}_n$ est construite en vue d'être utilisée pour prédire Y à partir d'un nouveau X où (X, Y) est une paire de *données de test*, c'est à dire pas nécessairement observée dans les données d'entraînement.

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

2 Apprentissage supervisé

2.1 Approches 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 : méthode à vecteurs supports (SVM), régression logistique, ...
— Réseaux de neurones
— Méthodes de modélisation probabilistes (modèle graphiques, méthodes bayésiennes)
— Apprentissage séquentiel

2.2 Formalisme de la théorie de la décision (version apprentissage)

La théorie de la décision peut formaliser 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 les paires de données de test possible, distribuées selon une loi P ; en notation : $(X, Y) \sim P_{X,Y}$.

— Soit $\mathcal{A}$ un ensemble d'actions, de décisions ou de prédictions possibles.

— Soit $\ell : \mathcal{A} \times \mathcal{Y} \to \mathbb{R}$ 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* $f : \mathcal{X} \to \mathcal{A}$ (et $\mathcal{F}$ un sous-ensemble de fonctions de $\mathcal{X}$ vers $\mathcal{A}$)

Pour un problème de décision défini par $(X, Y) \sim P_{X,Y}$, $\mathcal{A}$ et ℓ, on définit le *risque* $\mathcal{R}(f)$ (au sens de Vapnik [1]) pour une fonction de prédiction f tel que :

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

L'espérance est bien évidemment prise par rapport à la loi jointe sur $(X, Y) \sim P_{X,Y}$. 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 (sur $\mathcal{F}$) du risque, ce prédicteur $f^\star$ est appelé *fonction cible*, *fonction oracle* (avec un $\mathcal{F}$ souvent implicitement compris par le contexte ; par exemple, toutes les fonctions mesurables de $\mathcal{X}$ vers $\mathcal{A}$) et on définit :

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

On appelle *risque conditionnel* et on le note $\mathcal{R}(a \mid X) := \mathbb{E}[\ell(a, Y) \mid X]$. Si $\inf_{a \in \mathcal{A}} \mathcal{R}(a \mid X)$ 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 :

$$f^\star(X) := \operatorname{argmin}_{a \in \mathcal{A}} \mathcal{R}(a \mid X).$$

Si le prédicteur de Bayes est dans $\mathcal{F}$, il est alors identique à 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

$$\mathcal{R}\left(\widehat{f}_n\right) - \mathcal{R}(f^\star) = \mathbb{E}[\ell(\widehat{f}_n(X), Y) \mid D_n] - \inf_{f \in \mathcal{F}} \mathbb{E}[\ell(f(X), Y)]$$

Remarque (Risque d'un prédicteur issu d'une règle d'apprentissage). Comme $\widehat{f}_n = \mathscr{A}(D_n)$ où D_n est aléatoire, le risque de $\widehat{f}_n$ est une **variable aléatoire** :

$$\mathcal{R}\left(\widehat{f}_n\right) = \mathbb{E}[\ell(\widehat{f}_n(X), Y) \mid D_n]$$

Définition 2. (Risque fréquentiste) On peut évaluer un algorithme d'apprentissage en analysant *l'espérance* du risque (en faisant la moyenne sur les données d'entraînement D_n possibles), qui s'appelle le *risque fréquentiste* pour $\mathscr{A}$:

$$\mathcal{R}^F(\mathscr{A}) := \mathbb{E}_{D_n \sim P^{\otimes n}}[\mathcal{R}\left(\widehat{f}_n\right)] \quad (\widehat{f}_n = \mathscr{A}(D_n)).$$

Nous pouvons aussi analyser la probabilité du risque de $\widehat{f}_n$ d'être inférieur à une valeur donnée ; nous allons revisiter ces concepts plus en détail dans le cours 6 sur la théorie.

1. **Attention !** La notion de 'risque' en apprentissage statistique est subtilement différente de la notion de risque pour les *estimateurs* dans la théorie de décision en statistique traditionnelle – nous y reviendrons dans le cours sur la théorie. Voir aussi le 'risque fréquentiste' défini plus bas

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 : $f^\star(X) = \mathbb{E}[Y|X]$

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 : $f^\star(X) = \text{argmax}_k \mathbb{P}(Y = k|X)$

Autres sortes d'apprentissage

À l'exception de l'estimation de densité, nous n'allons pas couvrir ces autres approches dans ce cours, mais elles font aussi partie du domaine de l'apprentissage !

— **Apprentissage non-supervisée :**
  — **Estimation de densité** : nous observons seulement des x sans étiquette y ; l'action est une distribution sur $\mathcal{X}$: $\mathcal{A} = \{p_\theta | \theta \in \Theta\}$; la perte est normalement $\ell(\theta, x) := -\log p_\theta(x)$.
  — Le regroupement (clustering) : identifier des groupes dans x. La réduction de dimensions. Etc.
— **Apprentissage séquentiel** : les données (x_i, y_i) n'arrivent pas i.i.d., et sont peut-être mêmes générées par un adversaire !
— **Apprentissage par renforcement** : modélise une séquence de problèmes d'apprentissages avec une fonction de récompense dans le temps ; très populaire pour les robots et les agents en IA.

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* $P_n = \frac{1}{n}\sum_{i=1}^n \delta_{(x_i,y_i)}$. [2]

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* – on estime $\widehat{f}_n$ en minimisant le risque empirique :

$$\widehat{f}_n \in \arg\min_{f \in S} \widehat{\mathcal{R}}_n(f) = \arg\min_{f \in S} \underbrace{\frac{1}{n}\sum_{i=1}^n \ell(f(x_i), y_i)}_{\text{erreur d'entraînement}} .$$

Exemple 3. (La régression linéaire)
On considère le cas $\mathcal{X} = \mathbb{R}^p$, $\mathcal{Y} = \mathbb{R}$ et ℓ 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 :

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

2. Notez que $P_n \underset{n\to\infty}{\to} P_{X,Y}$ dans un sens qui peut être formalisé – au sens faible par la loi des grands nombres, et au sens fort grâce au théorème de Glivenko-Cantelli, également appelé "théorème fondamental de la statistique".

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 3. (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* pour P (et ℓ et $\mathcal{F}$ souvent implicites) si le prédicteur qu'il définit satisfait

$$\lim_{n \to \infty} \mathbb{E}\big[\mathcal{R}(\widehat{f}_n)\big] - \mathcal{R}(f^\star) = 0.$$

Définition 4. (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 diapos de l'intro ; et TP de la semaine suivante.

2.5 Décomposition du risque

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

— la fonction cible $f^\star = \operatorname{argmin}_{f \in \mathcal{F}} \mathcal{R}(f)$

— la meilleur approximation de la fonction cible dans S : $f_S^\star := \operatorname{argmin}_{f \in S} \mathcal{R}(f)$

— le prédicteur $\widehat{f}_S$ obtenu par la règle d'apprentissage à partir de données

On a la décomposition (triviale) :

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

2.6 Compromis Biais-Variance

Apparentée mais différente de la décomposition précédente, c'est une décomposition de l'*espérance du risque* (donc une décomposition du risque fréquentiste) 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 :

$$\mathbb{E}\big[\mathcal{R}(\widehat{f})\big] = \mathbb{E}[(\widehat{f}(X) - Y)^2] = \underbrace{\mathbb{E}[(\widehat{f}(X) - \mathbb{E}[\widehat{f}(X)|X])^2]}_{\substack{\textbf{variance} \text{ de } \widehat{f} \text{ causée par } D_n \\ \text{intégrée par rapport à } X}} + \underbrace{\mathbb{E}[(\mathbb{E}[\widehat{f}(X)|X] - \overbrace{\mathbb{E}[Y|X]}^{f^\star(X)})^2]}_{\textbf{biais} \text{ de } \widehat{f}} + \underbrace{\mathbb{E}[(Y - \overbrace{\mathbb{E}[Y|X]}^{f^\star(X)})^2]}_{\text{variance du “bruit”}}$$

Pour bien interpréter cette expression : par exemple, $\mathbb{E}[\widehat{f}(X)|X]$ est l'espérance de $\widehat{f}(X)$ par rapport aux variations de données d'entraînement D_n, pour un X (test) fixé (qui est indépendant de D_n). Aussi, les trois termes peuvent être vus comme intégrer les ‘variances’ et ‘biais’ par rapport à la loi sur les X : $\mathbb{E}[\mathbb{E}[\dots|X]]$.

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 $\rightarrow$ 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.

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

exemple : norme hilbertienne, norme ℓ_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 λ le *paramètre de régularisation* ou *hyperparamètre*. Il fait néanmoins choisir λ $\rightarrow$ problème analogue au problème de sélection de modèle. Nous allons couvrir cela plus en détails sur le cours de la sélection de modèle.

3.0.4 Régression ridge

Forme de régularisation la plus classique en statistiques.

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

Grâce à la régularisation, le problème est devenu fortement convexe (nous allons revoir ce concept dans le cours d'analyse convexe). Donc la solution est unique :

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

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.