le cnam

Apprentissage, réseaux de neurones et modèles graphiques

Chapitre 6 : Apprentissage et généralisation

Michel Crucianu

http://lecnam.net

http://cedric.cnam.fr/~crucianm/ml.html

le cnam

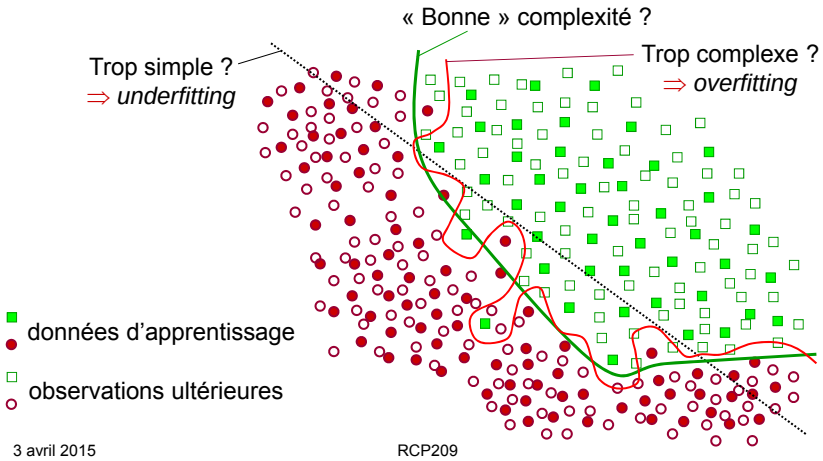
Contenu du chapitre

- Définition du cadre
- Composantes du risque
- Capacité et cohérence
- Capacité et contrôle de la généralisation
- Mesures de la capacité
- Bornes de généralisation

le cnam

Apprentissage et généralisation

- Objectif : erreur faible au-delà des données d'apprentissage



3 avril 2015 RCP209 3

le cnam

Définition du cadre

- Espace d'entrée $\mathcal{X}$, espace de sortie $\mathcal{Y}$
- Variables aléatoires $(X,Y) \in \mathcal{X} \times \mathcal{Y}$ suivant la loi inconnue P
- Données $D_n = \{(X_i, Y_i)\}_{1 \le i \le n}$ correspondant à des tirages indépendants et identiquement distribués (iid), suivant P
- Objectif : trouver une fonction $f : \mathcal{X} \to \mathcal{Y}$, $f \in \mathcal{F}$, qui prédit Y à partir de X en minimisant le risque espéré (théorique)

$$R(f) = E_P[L(X, Y, f)]$$

- P étant inconnue, on ne peut pas évaluer $R(f)$!

→ mais on peut mesurer à partir des observations D_n le risque empirique

$$R_{D_n}(f) = \frac{1}{n}\sum_{i=1}^{n} L(X_i, Y_i, f)$$

3 avril 2015 RCP209 4

le cnam

Définition du cadre (2)

- $L(X,Y,f)$ est une fonction de perte, par exemple
  - Perte quadratique pour la régression : $L(\mathbf{x},y,f)=[f(\mathbf{x})-y]^2$
  - Perte pour la discrimination entre 2 classes : $L(\mathbf{x},y,f)=\mathbf{1}_{f(\mathbf{x})\neq y}$
- Alternatives pour chercher f
  - Minimisation du risque empirique (MRE) : $f^*_{D_n}=\arg\min_{f\in\mathcal{F}} R_{D_n}(f)$
  - Régularisation : pénaliser la complexité à travers $\mathrm{reg}(f)$ $f^*_{D_n}=\arg\min_{f\in\mathcal{F}}\left[R_{D_n}(f)+\mathrm{reg}(f)\right]$
  - Minimisation du risque structurel : séquence $\mathcal{F}_d, d\in\mathbb{N}$, de modèles de « capacité » d croissante, dans chaque famille minimiser le risque empirique, ensuite pénaliser la capacité $f^*_{D_n}=\arg\min_{f\in\mathcal{F}_d,\, d\in\mathbb{N}}\left[R_{D_n}(f)+\mathrm{pen}(n,d)\right]$

le cnam

Composantes du risque

- Si f^* est la fonction de $\mathcal{F}$ qui minimise le risque espéré, $R(f^*)=\inf_{f\in\mathcal{F}} R(f)$, et $R^*=\inf_{\text{tout } f \text{ possible}} R(f)$ alors

$$R(f^*_{D_n}) = \underbrace{R^*}_{\text{risque résiduel}} + \underbrace{\left[R(f^*)-R^*\right]}_{\text{erreur d'approximation}} + \underbrace{\left[R(f^*_{D_n})-R(f^*)\right]}_{\text{erreur d'estimation}}$$

  - Risque résiduel (Bayes) : non nul en présence de bruit (la relation entre X et Y n'est pas une fonction)
  - Erreur d'approximation : nulle seulement si R^* peut être atteint par une fonction de $\mathcal{F}$ ($\mathcal{F}$ contient la « vraie » dépendance)
  - Erreur d'estimation : la fonction de $\mathcal{F}$ qui minimise le risque empirique n'est pas nécessairement celle qui minimise le risque espéré
  - Elargir $\mathcal{F}$ $\rightarrow$ erreur d'approximation ↘ , erreur d'estimation ↗

le cnam

Cohérence de la MRE

- Quand la taille de l'échantillon d'apprentissage augmente, l'erreur d'apprentissage converge-t-elle (en probabilité) vers le risque espéré ? (*consistency*)

$$\forall \varepsilon > 0, \quad \lim_{n\to\infty} P\left[\left| R_{D_n}\left(f^*_{D_n}\right) - \inf_{f\in\mathcal{F}} R(f) \right| > \varepsilon \right] = 0$$

et

$$\forall \varepsilon > 0, \quad \lim_{n\to\infty} P\left[R\left(f^*_{D_n}\right) - \inf_{f\in\mathcal{F}} R(f) > \varepsilon \right] = 0$$

  - Si la taille de l'échantillon augmente, alors doivent diminuer
    - L'écart entre risque empirique optimal et risque espéré optimal
    - L'écart entre risque espéré de la fonction qui minimise le risque empirique et le risque espéré optimal
  - Note : le risque espéré optimal ne dépend pas de l'échantillon

le cnam

Cohérence de la MRE (2)

- Condition nécessaire et suffisante pour la cohérence :

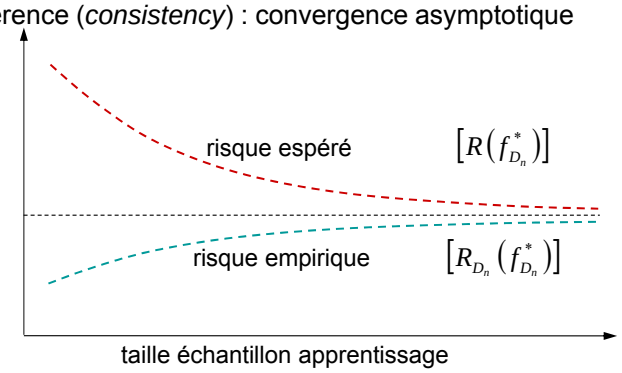
$$\forall \varepsilon > 0, \quad \lim_{n\to\infty} P\left[\sup_{f\in\mathcal{F}} \left(R(f) - R_{D_n}(f) \right) > \varepsilon \right] = 0$$

  - Lorsque la taille de l'échantillon augmente, la valeur du risque empirique ne peut pas rester éloignée de la valeur du risque espéré (quelle que soit la fonction de $\mathcal{F}$)
  - Condition pas nécessairement satisfaite pour toute famille de fonctions : famille $\mathcal{F}$ de « capacité » infinie $\Rightarrow$ l'écart entre les deux risques peut rester élevé « souvent » (il n'y a pas convergence en probabilité) quelle que soit la taille de l'échantillon !

le cnam

Risque et taille de l'échantillon

- Le risque empirique doit avoir tendance à augmenter avec la taille de l'échantillon d'apprentissage, le risque espéré doit avoir tendance à diminuer
- Cohérence (*consistency*) : convergence asymptotique



3 avril 2015 RCP209 9

le cnam

Contrôle de la généralisation

- Pour un échantillon d'apprentissage de taille finie, peut-on borner la différence entre l'erreur d'apprentissage et le risque espéré ?

$$R(f^*_{D_n}) = R_{D_n}(f^*_{D_n}) + \underbrace{\left[R(f^*_{D_n}) - R_{D_n}(f^*_{D_n})\right]}_{\leq ?}$$

- Bornes de généralisation : par rapport à $R_{D_n}(f^*_{D_n})$

$$R(f^*_{D_n}) \leq R_{D_n}(f^*_{D_n}) + B(n, \mathcal{F})$$

- Autres types de bornes :
  - Par rapport à $R(f^*)$: $R(f^*_{D_n}) \leq R(f^*) + B(n, \mathcal{F})$
  - Par rapport à R^* : $R(f^*_{D_n}) \leq R^* + B(n, \mathcal{F})$

3 avril 2015 RCP209 10

le cnam

Dimension de Vapnik-Chervonenkis

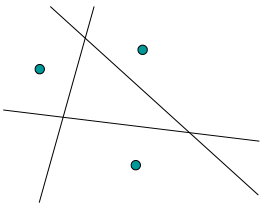
- Considérons un échantillon $\{\mathbf{x}_1,\ldots,\mathbf{x}_n\}$ de $\mathbb{R}^m$, il y a 2^n façons différentes de le séparer en 2 sous-échantillons
- **Définition** : un ensemble $\mathcal{F}$ de fonctions $f:\mathbb{R}^m \to \{-1,\ 1\}$ pulvérise (*shatters*) $\{\mathbf{x}_1,\ldots,\mathbf{x}_n\}$ si toutes les 2^n séparations peuvent être construites avec des représentants de $\mathcal{F}$
- **Définition** (VC-dimension) : l'ensemble $\mathcal{F}$ est dit de VC-dimension h s'il pulvérise au moins un échantillon de h vecteurs et aucun échantillon de $h+1$ vecteurs
- **Théorème** (cohérence de la MRE) : la Minimisation du Risque Empirique est cohérente si et seulement si la VC-dimension de $\mathcal{F}$ est finie

3 avril 2015 RCP209 11

le cnam

VC-dimension : exemple

- La VC-dimension de l'ensemble des hyperplans de $\mathbb{R}^m$ est $m+1$
- Dans $\mathbb{R}^2$:



Aucune droite ne peut séparer les points bleus des points rouges

3 avril 2015 RCP209 12

le cnam

Bornes de généralisation

- **Théorème** [Vap98] : soit $R_{D_n}(f)$ défini par $L(\mathbf{x}, y, f) = \mathbf{1}_{f(\mathbf{x}) \neq y}$; si la VC-dimension de $\mathcal{F}$ est $h < \infty$, alors pour tout $f \in \mathcal{F}$, avec une probabilité au moins égale à $1 - \delta \quad (0 < \delta < 1)$,

$$R(f) \leq R_{D_n}(f) + \underbrace{\sqrt{\frac{h\left(\log\frac{2n}{h} + 1\right) - \log\frac{\delta}{4}}{n}}}_{B(n,\mathcal{F})} \qquad \text{pour} \quad n > h$$

  - $B(n, \mathcal{F})$ diminue quand $n \uparrow$, quand $h \downarrow$ et quand $\delta \uparrow$
  - $B(n, \mathcal{F})$ ne fait pas intervenir le nombre de variables
  - $B(n, \mathcal{F})$ ne fait pas intervenir la loi conjointe P de X, Y

3 avril 2015 RCP209 13

le cnam

Contenu de la séance suivante

- SVM pour la discrimination
  - Discrimination linéaire et marge
  - Maximisation de la marge
  - Problème d'optimisation dans le cas séparable
  - Problème d'optimisation dans le cas non séparable
  - Astuce des noyaux (*kernel trick*)
  - Problème d'optimisation dans le cas non linéaire
  - Formulation alternative : ν-SVM

3 avril 2015 RCP209 14

le cnam

Références

[BBL99] O. Bousquet, S. Boucheron, G. Lugosi. Introduction to statistical learning theory. *Neural Information Processing Systems*, 1999.

[VC74] V. Vapnik, A. Chervonenkis. *Theory of Pattern Recognition*. Nauka, Moscow. 1974 (en russe).

[Vap98] V. Vapnik. *Statistical Learning Theory*. John Wiley, New York. 1998.