Problème d'apprentissage

- Données
  - $D_n = (Z_1, Z_2, \ldots, Z_n)$ générées par la "nature"
- IID: "independent and identically distributed"
  - tirés de la même distribution INCONNUE $p(Z)$
  - de manière indépendante

Problème d'apprentissage

- Les trois problèmes considérés
  - classification: $Z = (X, Y) \in \mathbb{R}^d \times \{1, \ldots, N\}$ ($\mathbb{R}^d \times \{-1, 1\}$)
  - régression: $Z = (X, Y) \in \mathbb{R}^d \times \mathbb{R}$
  - estimation de densité: $Z \in \mathbb{R}^d$
- Ensemble de fonctions F (solutions possibles), $f \in F$:
  - classification: $f : \mathbb{R}^d \to \{-1, 1\}$
  - régression: $f : \mathbb{R}^d \to \mathbb{R}$
  - estimation de densité: F contient des fonctions de densité

Problème d'apprentissage

- La perte $L(Z, f)$
  - mesurer la qualité d'une solution sur un point de donnée
  - classification: $L(Z, f) = L\big((X, Y), f\big) = I_{\{f(X) \neq Y\}}$
  - régression: $L(Z, f) = L\big((X, Y), f\big) = (f(X) - Y)^2$
  - estimation de densité: $L(Z, f) = -\log f(Z)$

Problème d'apprentissage

- Le risque
  - perte espérée, erreur de généralisation:

$$R(f) = E[L(Z, f)] = \int L(Z, f)p(Z)dz$$

  - classification: probabilité de mauvaise classification
  - régression: erreur quadratique espérée
  - estimation de densité: "distance" de l'estimation de la vrai densité

Problème d'apprentissage

- Le problème de l'induction
  - trouver $f \in F$ qui minimise le risque $R(f)$
  - $p(Z)$ est INCONNU!!!

Problème d'apprentissage

- Erreurs d'estimation et d'approximation, capacité
  - la sortie de notre algorithme d'apprentissage: $\widehat{f}(D_n) = \widehat{f}_n$
  - la meilleure fonction dans F :

$$f_F^* = \arg\min_{f \in F} R(f)$$

  - la meilleure fonction possible (la décision/l'erreur de Bayes):

$$f^* = \arg\min R(f)$$

  - $R(\widehat{f}_n) - R(f^*) = (R(\widehat{f}_n) - R(f_F^*)) + (R(f_F^*) - R(f^*))$

Problème d'apprentissage

- Le risque empirique
  - perte moyenne, erreur empirique:

$$\widehat{R}(f, D_n) = \widehat{R}(f) = \frac{1}{n}\sum_{i=1}^{n} L(Z_i, f)$$

  - $\widehat{R}(f, D_n)$ est un estimateur bruité mais non-biaisé de $R(f)$:

$$E_{D_n}[\widehat{R}(f, D_n)] = R(f)$$

  - classification: la fréquence/le taux des points mal classifiés
  - régression: erreur quadratique moyenne
  - estimation de densité: la vraisemblance négative pour que les points soient générés par f

Problème d'apprentissage

- Le principe de minimisation du risque empirique
  - trouver $f \in F$ qui minimise $\widehat{R}(f)$:

$$\widehat{f}^*(D_n) = \widehat{f}_n^* = \arg\min_{f \in F} \widehat{R}(f, D_n)$$

  - Lemme (Vapnik & Chervonenkis)

$$R(\widehat{f}_n^*) - R(f_F^*) \leq 2 \max_{f \in F} |\widehat{R}(f) - R(f)|$$

  - question théorique: est-ce que $\widehat{R}(f)$ approche bien $R(f)$ ($f \in F$)?
  - réponse: oui, si la capacité (complexité) de F est petite

Problème d'apprentissage

- Estimer l'erreur de généralisation
  - problème: $\widehat{R}(\widehat{f}_n^*)$ est un estimateur biaisé de $R(\widehat{f}_n^*)$
  - ensemble de test: $D'_m = (Z'_1, Z'_2, \ldots, Z'_m)$
  - erreur de test:

$$\widehat{R}(\widehat{f}_n^*, D'_m) = \frac{1}{m} \sum_{i=1}^{m} L(Z'_i, \widehat{f}_n^*)$$

  - si D'_m est iid et indépendant de D_n, alors $\widehat{R}(\widehat{f}_n^*, D'_m)$ est un estimateur non-biaisé de $R(\widehat{f}_n^*)$:

$$\lim_{m \to \infty} \widehat{R}(\widehat{f}_n^*, D'_m) = R(\widehat{f}_n^*)$$

Problème d'apprentissage

- Problèmes de l'apprentissage
  - algorithmique: choisir F tel que $\widehat{f}_n^*$ est facile à trouver
  - théorique 1: choisir F tel que $\widehat{R}(\widehat{f}_n^*)$ est près de $R(\widehat{f}_n^*)$
  - théorique 2: choisir F tel que $R(\widehat{f}_n^*)$ est près de $R(f^*)$
- Courbes d'apprentissage
- Le principe de minimisation du risque structurel

Théorie de Bayes

- Classification bayesienne
  - classification, deux classes $\{1, 0\}$, une variable X
  - probabilités a-postériori:

$$\begin{aligned} p_+(x) &= P(Y=1|X=x) = \mu(x) = E[Y|X=x] \\ p_-(x) &= P(Y=0|X=x) \end{aligned}$$

  - décision optimale:

$$f^*(x) = \begin{cases} 1 & \text{si } p_+(x) \geq p_-(x) \\ 0 & \text{si } p_+(x) < p_-(x) \end{cases} = \begin{cases} 1 & \text{si } \mu(x) \geq \frac{1}{2} \\ 0 & \text{si } \mu(x) < \frac{1}{2} \end{cases}$$

Théorie de Bayes

- Théorème
  - pour une fonction de décision $f : \mathbb{R} \mapsto \{0, 1\}$ quelconque:

$$P(f^*(X) \neq Y) \leq P(f(X) \neq Y)$$

  - $f^*(x)$: décision de Bayes
  - $R^* = R(f^*) = P(f^*(X) \neq Y)$: erreur/risque de Bayes

Théorie de Bayes

- Le théorème de Bayes
  - probabilités a-priori: $P(Y = C_i)$
  - probabilités conditionnelles de classe: $p(X|Y = C_i)$
  - théorème:

$$p(C_i|X) = \frac{p(X|C_i)p(C_i)}{p(X)}$$

ou

$$p(X) = \sum_{i=1}^{N} p(X|C_i)p(C_i)$$

Théorie de Bayes

- Décision de Bayes:

$$f^*(x) = C_i \text{ si } \frac{p(x|C_i)p(C_i)}{p(x)} \geq \frac{p(x|C_j)p(C_j)}{p(x)} \text{ pour tout } j = 1, \ldots, N$$

ou

$$f^*(x) = C_i \text{ si } p(x|C_i)p(C_i) \geq p(x|C_j)p(C_j) \text{ pour tout } j = 1, \ldots, N$$

  - avantage: $P(Y = C_i)$ et $p(X|Y = C_i)$ sont plus faciles à estimer que $P(Y = C_i|X = x)$ directement