Cours de Data Science

Marc Tommasi

10 septembre 2022

Outline

Outline

Exemples de problèmes de classification supervisée

- spam : présence des mots « sex », « offer »
- papaye bonne ou pas selon couleur, moelleux, ...

Formalisation

- Données et espace de description $x \in \mathcal{X}$
- Cible $y \in \mathcal{Y}$
- Un échantillon $S \subseteq \mathcal{X} \times \mathcal{Y}$
- sortie de l'apprentissage : une règle de prédiction $f : \mathcal{X} \to \mathcal{Y}$

Vocabulaire

- espace de description : features, attributs, (voire champs)
- échantillons ou jeu de données, dataset
- composé d'instances, records, exemples, enregistrements
- cible, (étiquette ou classe dans certains cas),
- données d'entraînement
- sortie : fonction, règle(s) ou modèle

Hypothèses classiques pour la classification supervisée

- Les données de S sont générées selon une distribution de probabilités $\mathcal{D}$, fixée et inconnue.
- On suppose qu'il existe une telle fonction cible f et que les données sont générées par $\mathcal{D}$ puis étiquetées par f.
- Expression de la perte, ou erreur réelle pour une hypothèse h comme

$$L_{\mathcal{D},f}(h) = \mathcal{D}(\{x \mid h(x) \neq f(x)\}).$$

- L'apprenant ne connaît pas $\mathcal{D}$ donc
- le calcul de $L_{\mathcal{D},f}(h)$ ne peut pas être fait par l'apprenant !

ERM : Principe de minimisation du risque empirique

- La chose qu'il peut calculer est le risque ou perte empirique

$$L_S(h) = \frac{|\{i \in [m] \mid h(x_i) \neq y_i\}|}{m},$$

pour $S = \{(x_1, y_1), \ldots, (x_m, y_m)\}$.

Limite du modèle

- Problème d'apprentissage par cœur et d'overfitting. La règle suivante est parfaite si on minimise le risque empirique :

$$h(x) = \begin{cases} y_i \text{ si il existe } i \text{ tq } x = x_i \\ 0 \text{ sinon} \end{cases}$$

- On peut arriver au même constat avec des classes de fonctions (comme des polynômes
- En généralisation, on ne fait pas mieux que random.

ERM avec un biais

- on fixe une classe d'hypothèses $\mathcal{H}$, a priori, avant de voir les données, c'est une connaissance antérieure

$$\mathrm{ERM}_{\mathcal{H}}(S) = h_S \in \operatorname*{argmin}_{h \in \mathcal{H}} L_S(h)$$

- Dans l'exemple du spam ou des papayes,
  - possibilité : des rectangles dans le plan fonctionne avec le principe ERM (limite overfit)
  - une classe plus large permet l'overfit
- Question fondamentale du ML : quelle classe choisir pour ne pas avoir de l'overfitting avec ce principe ?

ERM et calcul d'erreurs

- L'*Hypothèse de réalisabilité* dit que $\mathcal{H}$ contient une fonction $h^\star$ qui réalise une perte nulle selon la distribution $L_{\mathcal{D},f}(h^\star) = 0$.
- Cela entraîne que l'erreur empirique $L_S(h^\star) = 0$
- Et si on suit le principe ERM, alors $L_S(h_S) = 0$ avec probabilité 1 si S est tiré selon $\mathcal{D}$.
- Mais on veut calculer l'erreur réelle $L_{\mathcal{D},f}(h_S)$ et non pas l'erreur empirique,
- on testera sur de nouvelles données
- d'où l'importance de du tirage selon $\mathcal{D}$ selon l'hypothèse iid.

Big data, small data

- Rappel : des données, de bonnes données, etc...
- quid dans les cas très déséquilibrés, distributions difficiles,...

Confiance et approximation

- $L_S(h_S)$ est bien une variable aléatoire car le choix de h_S dépend de S.

Confiance

- On a un résultat en probabilité car on ne peut garantir d'éviter des tirages extrêmes (toujours la même valeur de y,...).
- On note souvent δ la probabilité d'avoir un échantillon non représentatif. Alors $(1-\delta)$ est la confiance.

Approximation

- Le calcul de h peut n'être qu'approximatif (on commet quelques erreurs).
- On peut les tolérer jusqu'à un certain seuil de précision, souvent noté ϵ : on cherche à avoir $L_{\mathcal{D},f}(h_S) \leq \epsilon$
- On se trompe pour tous les échantillons S tels que $L_{\mathcal{D},f}(h_S) > \epsilon$

Classe de taille finie

Theorem (Les classes finies sont apprenables sous l'hypothèse de réalisabilité)

Si $\mathcal{H}$ est de taille finie, pour tout $\epsilon > 0$ et $\delta \in [0,1]$, pour toute fonction cible f de l'apprentissage, pour toute distribution $\mathcal{D}$ telle que il existe $h \in \mathcal{H}, L_{\mathcal{D},f}(h) = 0$, alors avec probabilité $1-\delta$ sur le tirage iid d'un échantillon S de taille supérieure à $\frac{\log(|\mathcal{H}|/\delta)}{\epsilon}$, l'erreur réelle de la minimisation du risque empirique (ERM) est plus petite que ϵ

$$L_{\mathcal{D},f}(h_s) \leq \epsilon.$$

- On considère que $\mathcal{H}$ est de taille finie, on prend $m \geq \frac{\log(|\mathcal{H}|/\delta)}{\epsilon}$ exemples alors $\mathcal{D}^m(\{S \mid L_{\mathcal{D},f}(h_S) > \epsilon\}) \leq \delta.$

Calcul de l'erreur réelle

- Comment borner la probabilité de tirer un échantillon qui entraîne un échec ? Quel est le poids selon $\mathcal{D}^m$ de cet échantillon de taille m qui calcule une hypothèse erronée ?

$$\mathcal{D}^m(\{S \mid L_{\mathcal{D},f}(h_S) > \epsilon\})$$

Des exemples non suffisamment informatifs

- On va construire une mauvaise hypothèse h_S si les exemples font croire que h_S est correcte : $L_{\mathcal{D},f}(h_S) > \epsilon$ et $L_S(h_S) = 0$ car on suppose l'*Hypothèse de réalisabilité*.
- On note M cet ensemble d'échantillons pour lesquels on peut construire une mauvaise hypothèse :

$$M = \bigcup_{h \text{ s.t. } L_{\mathcal{D},f}(h)>\epsilon} \{S' \mid L_{S'}(h) = 0\},$$

et S appartient donc à cet ensemble M donc

$$\mathcal{D}^m(\{S \mid L_{\mathcal{D},f}(h_S) > \epsilon\}) \leq \mathcal{D}^m(M).$$

Union Bound

- On utilise souvent cette inégalité en probabilités, appelée inégalité de Boole ou union bound en anglais : $\mathcal{D}(\bigcup_i A_i) \leq \sum_i \mathcal{D}(A_i)$.
- Appliqué ici, on trouve

$$\mathcal{D}^m(\{S \mid L_{\mathcal{D},f}(h_S) > \epsilon\}) \leq \sum_{h \text{ s.t. } L_{\mathcal{D},f}(h)>\epsilon} \mathcal{D}^m(\{S' \mid L_{S'}(h) = 0\}).$$

Borner le membre droit

- $L_{S'}(h) = 0$ si pour tout exemple x_i de S' on a $h(x_i) = f(x_i)$. Les exemples sont tirés de façon iid, donc

$$\mathcal{D}^m(\{S' \mid L_{S'}(h) = 0\}) = \prod_i^m \mathcal{D}(\{x_i \mid h(x_i) = f(x_i)\}).$$

- Comme $L_{\mathcal{D},f}(h) = \mathcal{D}(\{x \mid f(x) \neq h(x)\})$ on a

$$\mathcal{D}(\{x_i \mid h(x_i) = f(x_i)\}) = 1 - L_{\mathcal{D},f}(h).$$

- Pour un h pour lequel l'erreur est plus grande que ϵ on a

$$\mathcal{D}(\{x_i \mid h(x_i) = f(x_i)\}) \leq 1 - \epsilon,$$

et

$$\mathcal{D}^m(\{S' \mid L_{S'}(h) = 0\}) \leq (1 - \epsilon)^m.$$

Résultat

- En combinant

$$\mathcal{D}^m(\{S \mid L_{\mathcal{D},f}(h_S) > \epsilon\}) \leq \sum_{h \text{ s.t. } L_{\mathcal{D},f}(h)>\epsilon} \mathcal{D}^m(\{S' \mid L_{S'}(h) = 0\})$$

et

$$\mathcal{D}^m(\{S' \mid L_{S'}(h) = 0\}) \leq (1-\epsilon)^m.$$

on obtient en posant $\mathcal{H}_B = \{h \text{ s.t. } L_{\mathcal{D},f}(h) > \epsilon\}$ l'ensemble des mauvaises hypothèses

$$\mathcal{D}^m(\{S \mid L_{\mathcal{D},f}(h_S) > \epsilon\}) \leq |\mathcal{H}_B|(1-\epsilon)^m \leq |\mathcal{H}_B|e^{-\epsilon m} \leq |\mathcal{H}|e^{-\epsilon m}.$$