

Cours IFT6266, Estimation de l'erreur de généralisation et Bootstrap

- Jusqu'à présent nous avons présenté l'objectif de l'apprentissage comme la minimisation de l'erreur de généralisation espérée. Mais en pratique, nous avons aussi un objectif secondaire, qui est d'évaluer ET DE COMPARER l'erreur de généralisation espérée. Considérons d'abord le problème de l'évaluation de l'erreur et ensuite de la fiabilité de cette estimation.
- **Estimation de l'erreur de généralisation espérée par validation croisée.**
Il existe plusieurs variantes du principe de validation croisée, qui permet d'estimer l'erreur de généralisation d'un algorithme d'apprentissage pour un certain ensemble de données. Dans tous les cas, on fait la moyenne de plusieurs erreurs de test, mesurées sur des exemples qui n'ont pas été utilisés pour l'apprentissage. La validation croisée peut par exemple être utilisée pour essayer de choisir la capacité "optimale", ou pour choisir un algorithme d'apprentissage plutôt qu'un autre.
 - *train/test*: une certaine fraction des données est utilisée pour entraîner et le reste pour tester. C'est la méthode la plus simple et la plus rapide, mais elle ne fonctionne pas bien si on a peu de données (et c'est là que le problème de la précision de l'estimation de l'erreur par validation croisée est plus criant).
 - *K-fold XV*: on répète la procédure *train/test* pour K différents ensembles de test disjoints dont l'union est l'ensemble original de donnée. L'erreur estimée est la moyenne de toutes les erreurs de tests sur toutes les partitions.
 - *validation séquentielle*: les deux méthodes ci-haut ne s'appliquent pas si les données sont séquentielles (et donc probablement pas i.i.d.). La validation séquentielle procède avec plusieurs phases de *train/test* à chaque fois en entraînant avec les données jusqu'au temps t et en testant avec les données de $t + 1$ à $t + s$, et en recommençant avec l'entraînement jusqu'à $t + s$ et test de $t + s + 1$ à $t + 2s$, etc... L'erreur estimée est encore la moyenne de toutes les erreurs de tests.

Notez que la méthode *K-fold XV* et la méthode de *validation séquentielle* ont une complexité de calcul qui rajoute un facteur proportionnel au nombre d'exemples dans le temps de calcul total...

- **Bootstrap**

Une statistique est une mesure quelconque obtenue à partir d'un ensemble de données, comme par exemple (simple) sa moyenne, ou bien, (compliqué) un estimé de l'erreur de généralisation obtenue par validation croisée. Une manière générale d'obtenir un estimé des variations qu'une statistique peut subir à cause de notre ensemble d'apprentissage particulier est la méthode du Bootstrap. L'idée de base de la méthode du Bootstrap est la suivante: on va supposer vraie une hypothèse sur nos données (par exemple, elles sont i.i.d.), et on va s'en servir pour générer des ensembles de données alternatifs mais qui sont tout aussi plausibles que l'ensemble original (selon notre hypothèse). Avec le bootstrap non-paramétrique, on va supposer que les données sont i.i.d., et on va tirer un nouvel ensemble d'exemples de la même taille l que l'original en faisant l tirages avec remplacement

à partir des exemples originaux. Notez que cela est équivalent à substituer la véritable distribution des données par la distribution empirique correspondant à notre ensemble de données. On va tirer B tels ensembles de données sur chacun desquels on va mesurer notre statistique (e.g., l'erreur de généralisation estimée par validation croisée), et on va regarder par exemple l'écart-type de ces valeurs (à travers les B différentes expériences) pour mesurer la précision de la valeur de notre statistique.

- **Estimation de la robustesse de l'erreur estimée**

Le processus d'apprentissage et d'évaluation de l'erreur de généralisation dans son ensemble est le suivant. À partir de la distribution P , un ensemble d'apprentissage D_1 est tiré. Étant donné D_1 , on obtient une fonction f avec un algorithme d'apprentissage. On tire un ensemble de test D_2 de P , et on mesure l'erreur de f sur les exemples de D_2 (en faisant la moyenne des erreurs sur ces exemples). On peut donc considérer deux sources de variation qui font que notre estimé de l'erreur de généralisation est "bruité":

1. le choix particulier des exemples d'apprentissage D_1 (tirés d'une distribution de probabilité inconnue P),
2. le choix particulier des exemples de test D_2 .

La méthode du Bootstrap décrite plus haut permet de simuler tout le processus et donc d'obtenir un estimé qui tient compte des deux sources de variation. Cela revient à substituer un estimé $\hat{P}$ de P dans le processus ci-haut, par exemple $\hat{P}$ est la distribution empirique i.i.d. (bootstrap non-paramétrique), ou bien $\hat{P}$ est un modèle probabiliste des données (bootstrap paramétrique). Cependant cette méthode est très coûteuse en temps de calcul. Dans le cas où les erreurs sur les exemples de test sont indépendantes (étant donné la fonction f), ce qui sera toujours le cas si les données sont i.i.d., on peut facilement obtenir un estimé de la variance de notre estimé d'erreur, qui ne tient compte que de la deuxième source de variation (les exemples de test):

- **Variance due aux exemples de test**

Soit un ensemble i.i.d. d'erreurs e_i (i de 1 à l), par exemple dans le cas de la régression $e_i = (y_i - f(x_i))^2$ hors-échantillon. Considérons la moyenne des erreurs

$$\bar{e} = \frac{1}{l} \sum_i e_i.$$

et l'écart-type de l'échantillon

$$\hat{\sigma}_e = \sqrt{\frac{1}{l-1} \sum_i (e_i - \bar{e})^2}.$$

On sait que $E[\bar{e}] = E[e]$, i.e., $\bar{e}$ est un estimé non-biaisé de la véritable erreur de généralisation espérée $E[e]$. Par la loi des grands nombres, on sait aussi que asymptotiquement ($l \rightarrow \infty$),

$$Var[\bar{e}] \rightarrow \frac{Var[e]}{l}$$

ce qui nous donne immédiatement un “écart-type” $\hat{\sigma}_{\bar{e}} = \frac{\hat{\sigma}_e}{\sqrt{l}}$ autour de $\bar{e}$. De plus, par le théorème centrale limite, $\bar{e}$ converge vers une Normale asymptotiquement, donc on peut faire des test d'hypothèses (par exemple pour vérifier si on peut rejeter l'hypothèse nulle $E[e] = \theta$: il n'y a qu'à voir la masse $P(\bar{e} \leq \theta)$ (selon une distribution normale centrée à θ et d'écart-type $\hat{\sigma}_{\bar{e}}$) en deçà de ce seuil pour obtenir la p-value de cette hypothèse nulle. On peut raffiner un peu en observant qu'on a en fait affaire à une distribution t de Student (qui est une meilleure approximation que la Normale pour l petit) avec $l - 1$ degrés de libertés.

- Comparaison de deux modèles

En utilisant le modèle ci-haut, on peut comparer les erreurs de test $\bar{e}_A$ et $\bar{e}_B$ obtenus sur deux modèles A et B , en considérant que sous l'hypothèse nulle selon laquelle les modèles ont en fait la même erreur de généralisation, la variable

$$\sqrt{l} \frac{\bar{e}_A - \bar{e}_B}{\sqrt{\bar{\sigma}_A^2 + \bar{\sigma}_B^2}}$$

est asymptotiquement normale centrée réduite (donc si on observe une valeur beaucoup plus grande que 0, cela rejetera l'hypothèse nulle). Ici j'ai noté $\bar{\sigma}_A$ pour l'écart-type des erreurs du modèle A , etc...