Sciences des données

Un voyage initiatique

Cécile Capponi, Rémi Eyraud, Hachem Kadri

LIS, Aix-Marseille Université, CNRS
Equipe QARMA

M1 Informatique – S5 2019-2020

Plan

Outline

Quelques mesures empiriques classiques d'une hypothèse h

Classification binaire

$$S = \{(x_i, y_i)\}, i = 1 \cdots n, y_i \in \{-1, +1\}$$

- taux d'erreur apparente $e_a(h) = \frac{1}{n} \sum_{i=1\cdots n} \mathbb{I}(h(x_i) \neq y_i)$
- taux de bonne classification $a_a(h) = \frac{1}{n} \sum_{i=1\cdots n} \mathbb{I}(h(x_i) = y_i)$
- précision(+1)

$$p_{+a}(h) = \frac{\sum_{i=1\cdots n, y_i=+1} \mathbb{I}(h(x_i) = +1)}{\sum_{i=1\cdots n} h(x_i) = +1}$$

- rappel(+1)

$$r_{+a}(h) = \frac{\sum_{i=1\cdots n, y_i=+1} \mathbb{I}(h(x_i) = +1)}{\sum_{i=1\cdots n, y_i=+1} y_i}$$

- compromis pondéré : le F_β-score (souvent $\beta = 1$)

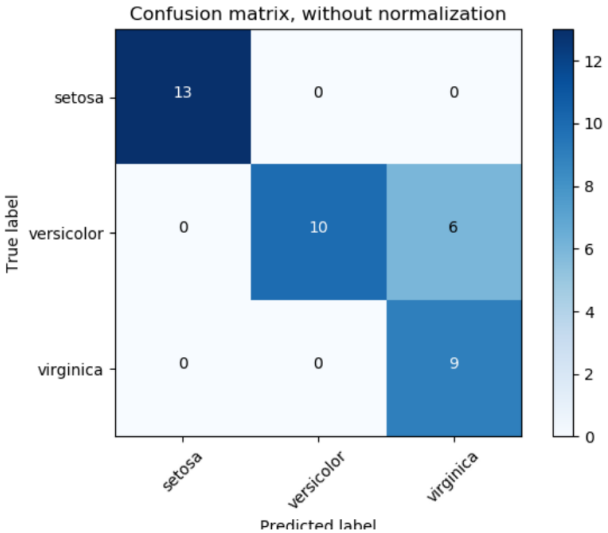
$$F_{\beta,+a}(h) = \frac{(1+\beta^2)(\text{precision.rappel})}{\beta^2.\text{precision} + \text{rappel}}$$

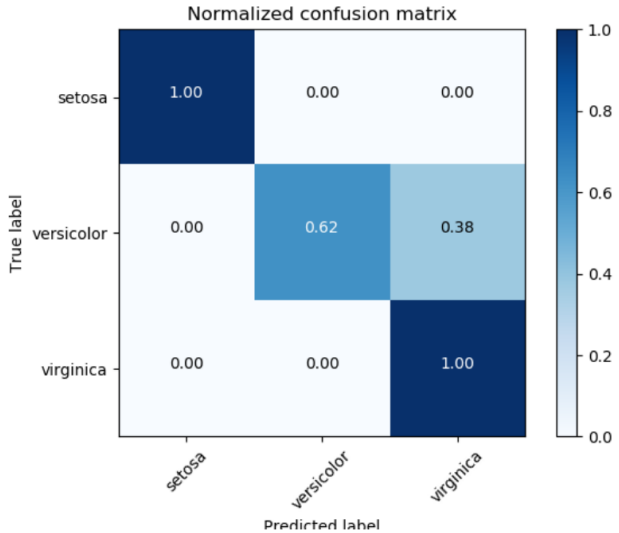
Matrice de confusion d'une hypothèse *h*

- Observer où se passent les confusions entre classes
- Diagonale = bonnes prédictions
- Hors-diagonale = les erreurs de prédictions
- Rappel+ = TP/(TP+FN)
- Précision+ = TP/(TP+FP)

		Classe réelle	
		-	+
Classe prédite	-	True Negatives *(vrais négatifs)*	False Negatives *(faux négatifs)*
	+	False Positives *(faux positifs)*	True Positives *(vrais positifs)*

(source : openclassroom)



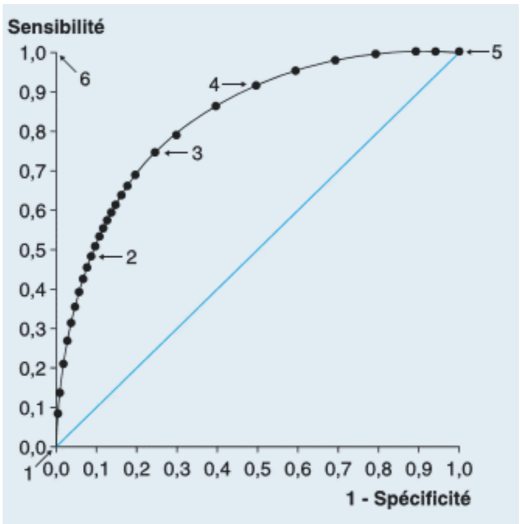


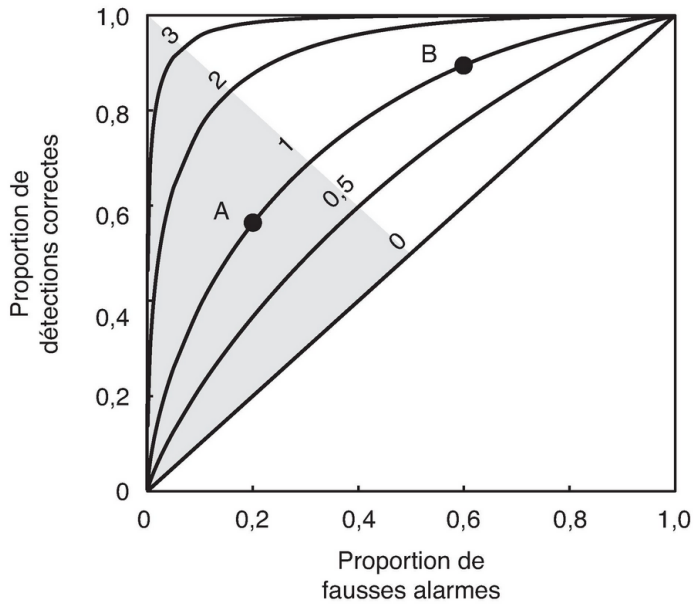
(source : scikit-learn)

La courbe ROC d'une hypothèse *h* (classification binaire)

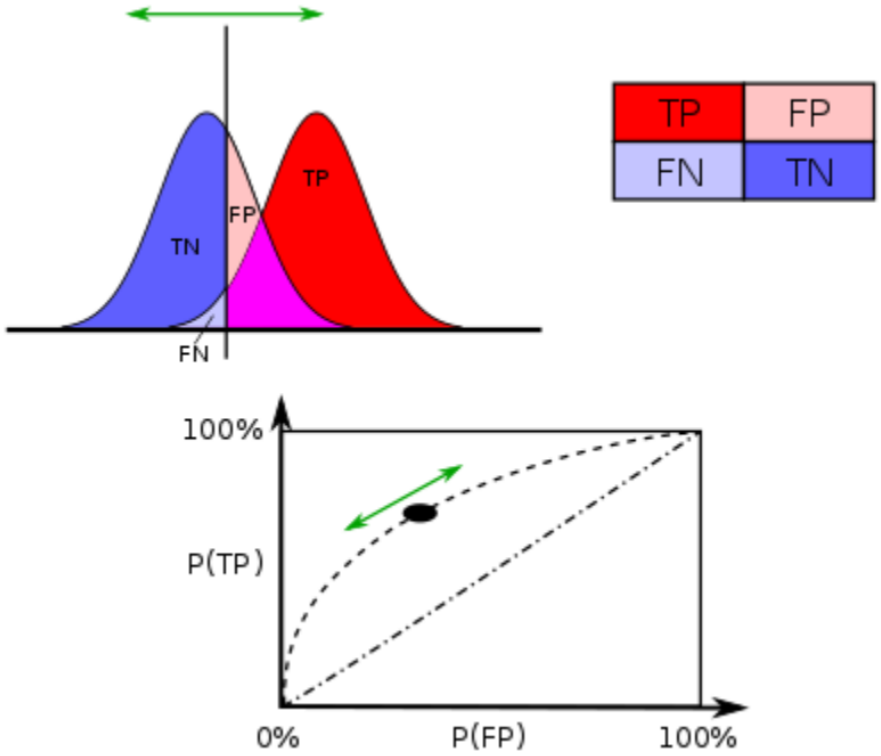
Principe : une courbe qui monte au plus tôt

- *h* doit retourner un score *s*, typiquement une probabilité entre 0 et 1 (et non pas directement la classe)
- Faire varier le seuil *S* d'affectation de la classe $t = \{+1, -1\}$ visée : si $h(x) > S$ alors $h(x) = t$.
- Tracer sensibilité ($\frac{TP}{TP+FN}$ = rappel+) versus 1-spécificité ($1 - \frac{TN}{TN+FP} = \frac{FP}{TN+FP}$ = taux de faux positifs)





La courbe ROC et l'AUC



(source : Xavier Dupré)

AUC = Aire sous la courbe ROC

- Courbe ROC : lecture graphique de la performance des hypothèses
- AUC : traduction mathématique par intégration, représente le taux de bonne classification
- valeur de Wilcoxon

Dans d'autres cadres que la classification binaire

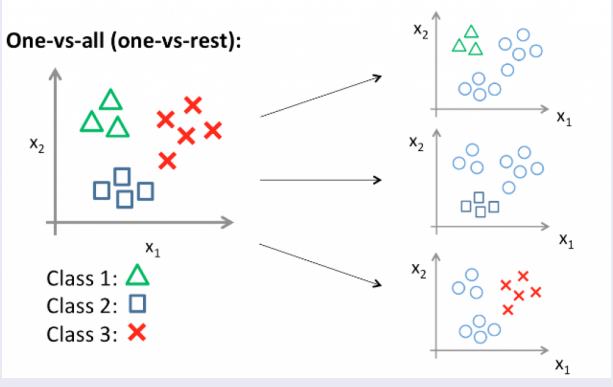
En régression, $S = \{(x_i, y_i)\}, y_i \in \mathbb{R}$: erreur quadratique moyenne

$$\text{MSE}(h) = \sum_i (h(x_i) - y_i)^2$$

En classification multi-classes

Erreur, précision, etc.

Calcul pour chaque classe et moyennes :
plusieurs schémas expérimentaux



(source : MrMint)

- Les approches OvA (one-vs-all), OvO (one-vs-one)
- Problème du déséquilibre de l'échantillon
  - Re-échantillonnage
  - MAUC (mean AUC)
  - Norme de la matrice de confusion (sans diagonale)

Outline

Train-Test split (hold-out)

Estimation : calculer la mesure choisie sur des exemples qui n'ont pas servi à apprendre. Soit $S = \{(x_i, y_i)\}, i = 1 \cdots n$.

Séparation de l'échantillon initial $S = A \cup T$

- Conserver la distribution $\rightarrow$ séparation au hasard
- k% pour l'apprentissage, $(100 - k)$% pour le test
- Apprendre h sur A (train), estimer la performance sur T (test)

Cas de l'erreur de classification : $\hat{e}(h) = \frac{\sum_{i=1..n} \mathbb{I}(h(x_i) \neq y_i)}{n}$

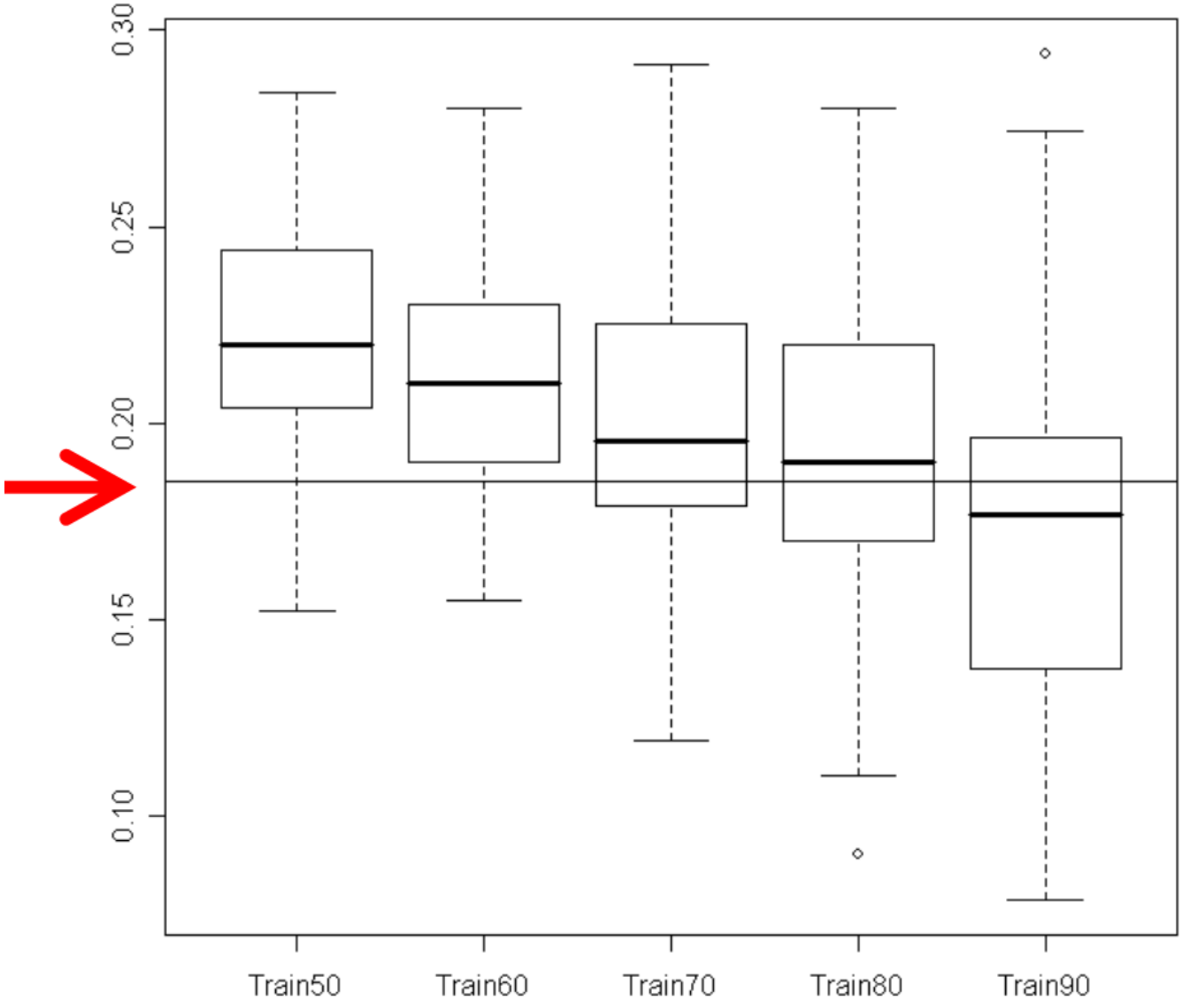
Répétition du hold-out et moyenne : stabilité de l'estimateur

Valeurs de k et qualité de l'estimateur de la performance

- Plus k est grand, plus le biais de l'estimateur est faible, plus sa variance est forte
- Plus k est petit, plus le biais de l'estimateur est fort, plus sa variance est faible
- OK si grand nombre de données, à éviter sinon

Les dangers du hold-out

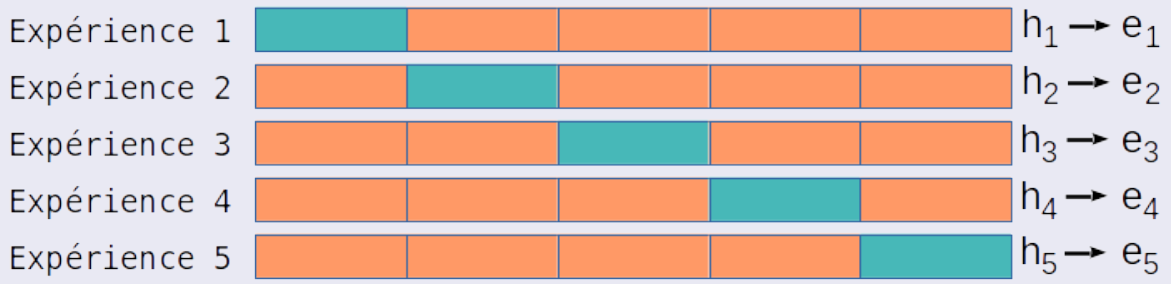
Difficulté du compris biais/variance *de l'estimateur*



Validation croisée et leave-one-out

Validation croisée *k* folds

- *k* Itération Train-Test sur *k* partitions différentes de $S = \{(x_i, y_i)\}, i = 1 \cdots n$
- Estimation de la mesure = moyenne des erreurs mesurées à chaque itération
- Sous-estimation de l'erreur en cas de sur-apprentissage
- Exemple avec $k = 5$ et estimation de l'erreur réelle e



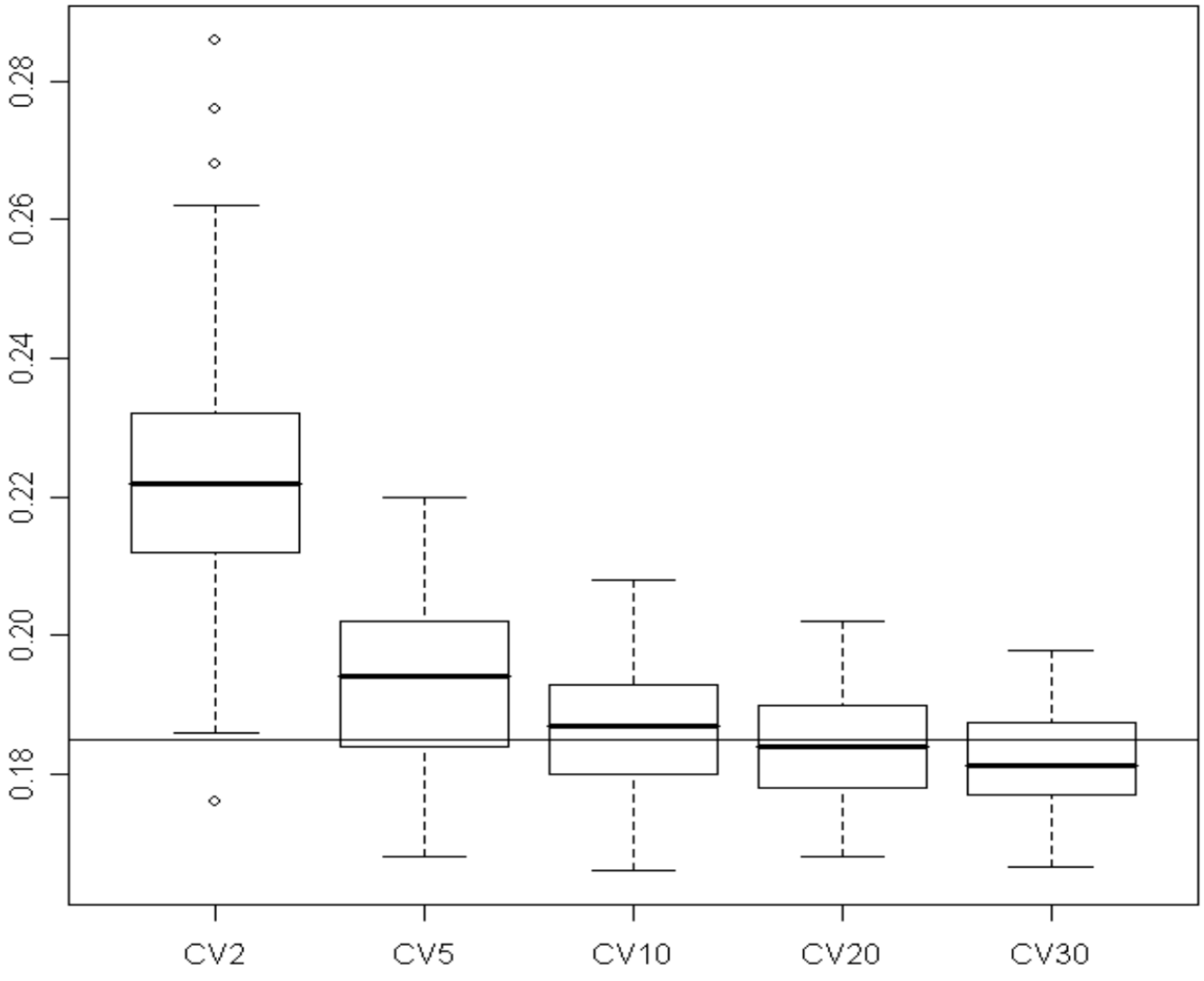
Apprendre sur
Tester sur

$$\hat{e} = \frac{e_1+e_2+e_3+e_4+e_5}{5}$$

Leave-one-out

Validation croisée avec $k = n$ folds

Qualité de l'estimateur par validation croisée



Estimation des hyper-paramètres (ex. le meilleur k des k-NN)

Via un ensemble de validation (si possible)

Une variante du hold-out

- Pour régler les hyper-paramètres, par exemple le meilleur k des k-ppv
- troisième extrait de l'échantillon initial (ou extrait de l'ensemble de train)
- Apprendre sur train, tester les hyper-paramètres sur val, estimer la mesure finale sur test.

Grid search et validation croisée

- Pour une sélection des valeurs possibles d'un hyper-paramètre : estimation de la performance via cross-validation
- Considérer des couples, trios, etc., d'hyper-paramètres
- Retenir *in fine* l'ensemble des valeurs des hyper-paramètres qui optimisent ensemble la mesure de performance