Data Mining/ML

Validation

Jamal Atif
jamal.atif@dauphine.fr

M2 ID

Université Paris-Dauphine

2015-2016

Plan

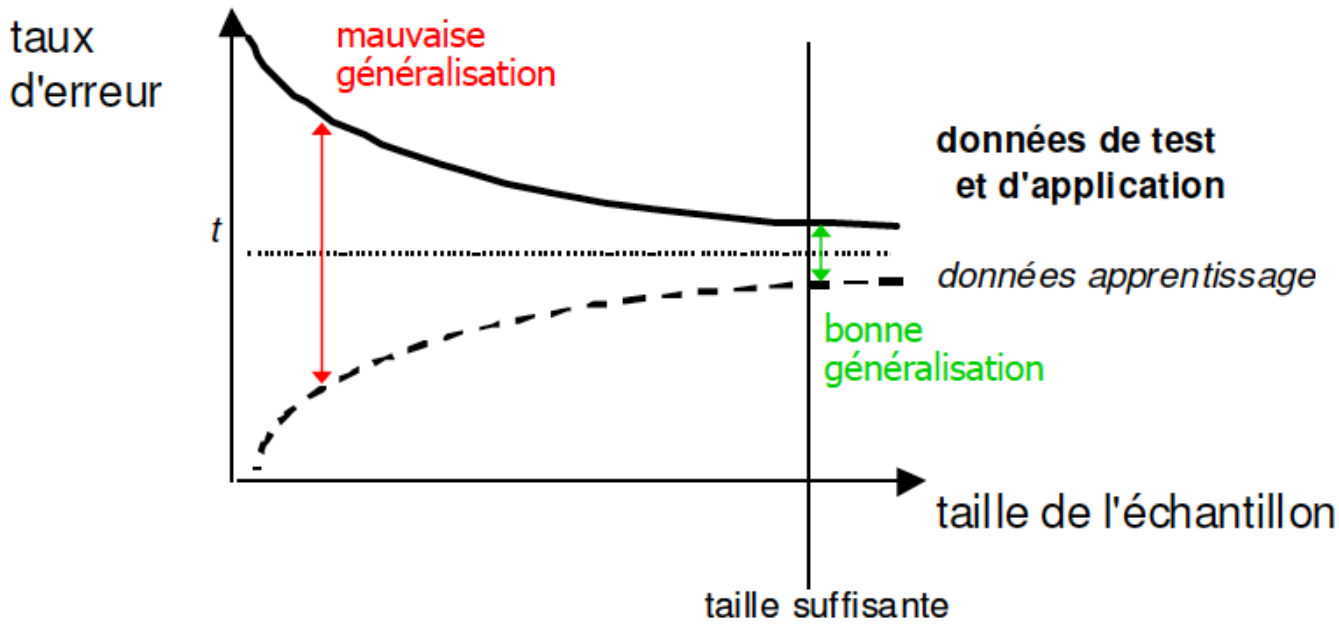
1. Quelques considérations générales
   - Evaluation
   - Ré-échantillonage

Qualités attendues d'une technique de classification

- **Précision** : le taux d'erreur, proportion d'individus mal classés doit être le plus bas possible.
- **Robustesse** : le modèle doit dépendre aussi peu que possible de l'échantillon d'apprentissage et se généraliser à d'autres échantillons.
- **Concision, parcimonie** : les règles du modèles doivent être aussi simples et aussi peu nombreuses que possible.
- **Diversité des types de données utilisées** : données qualitatives, discrètes, continues et manquantes.
- **Rapidité de calcul du modèle** : apprentissage rapide pour affinement du modèle.
- **Paramètrage** : pouvoir pondérer les erreurs de classement.

Pouvoir de généralisation

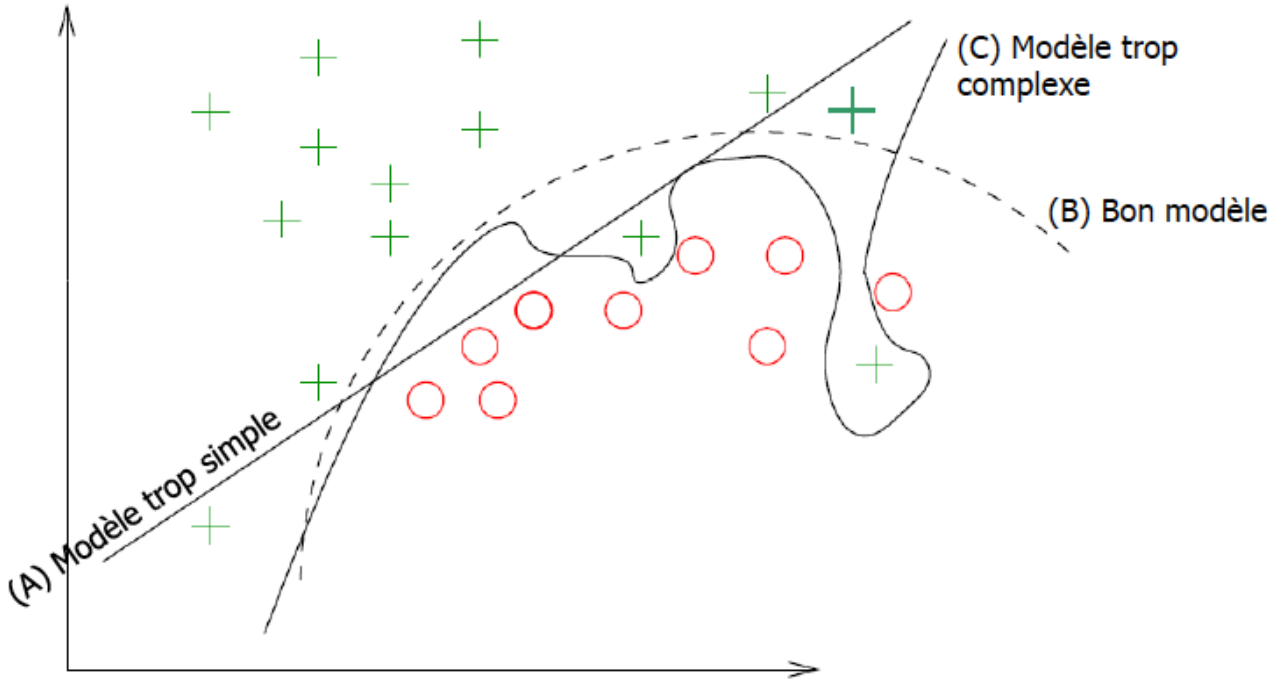
Courbes du taux d'erreur en apprentissage et en test.



Taille minimale de l'échantillon d'apprentissage :

- en deçà de laquelle le modèle obtenu en apprentissage se généralise mal en test en en application
- au dela de laquelle on n'observe plus de baisse sensible du taux d'erreur en test et en application.

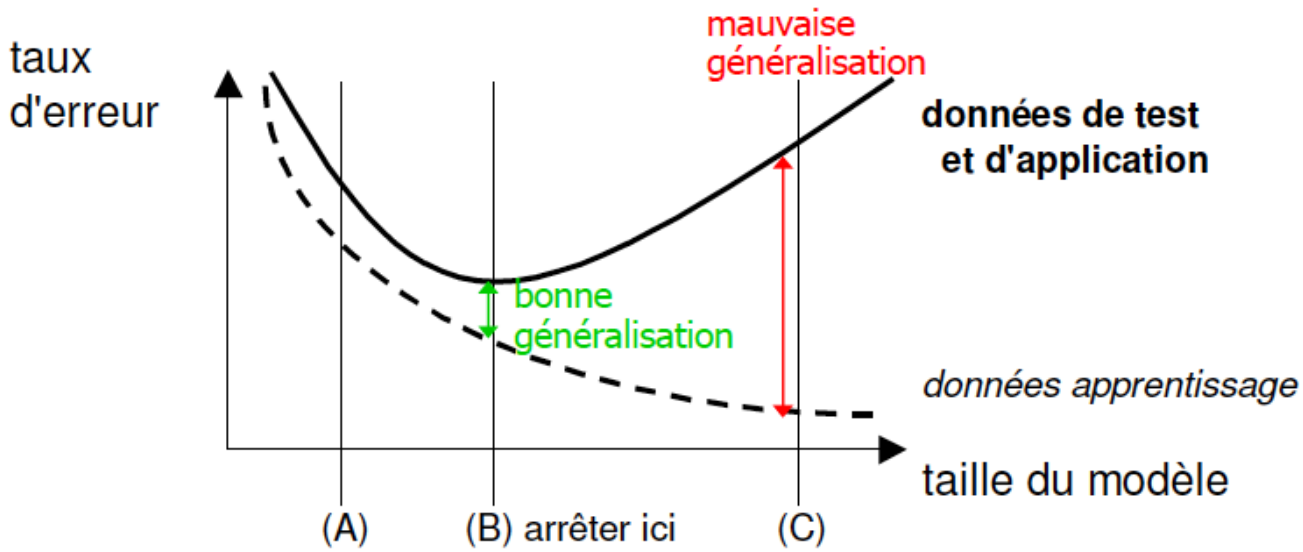
Sur apprentissage



Source : Olivier Bousquet

Sur apprentissage

taux d'erreur en fonction de la complexité du modèle.



Plan

1. Quelques considérations générales
   - Evaluation
   - Ré-échantillonage

Evaluation de la qualité d'un classifieur

- Panoplie de méthodes de classification.
- Laquelle choisir ? Y-a-t-il une méthode supérieure aux autres quelque soit le problème ?
- Y-a-t-il un ensemble de caractéristiques meilleur qu'un autre ?
- Comment évaluer une méthode de classification ? Quelles métriques ? Quelles méthodes ?
- Comment comparer les méthodes de classification entre elles ?

Métriques pour l'évaluation

Matrice de confusion

		PREDICTED CLASS	
		Class=Yes	Class=No
ACTUAL CLASS	Class=Yes	a	b
	Class=No	c	d

a: TP (true positive)

b: FN (false negative)

c: FP (false positive)

d: TN (true negative)

- True positive = correctly identified
- False positive = incorrectly identified
- True negative = correctly rejected
- False negative = incorrectly rejected

Métriques pour l'évaluation

Taux d'erreur : accuracy

		PREDICTED CLASS	
		Class=Yes	Class=No
ACTUAL CLASS	Class=Yes	a (TP)	b (FN)
	Class=No	c (FP)	d (TN)

$$\texttt{Accuracy} = \frac{a+d}{a+b+c+d} = \frac{TP+TN}{TP+TN+FP+FN}$$

Métriques pour l'évaluation

Taux d'erreur : accuracy

		PREDICTED CLASS	
		Class=Yes	Class=No
ACTUAL CLASS	Class=Yes	a (TP)	b (FN)
	Class=No	c (FP)	d (TN)

$$\texttt{Accuracy} = \frac{a+d}{a+b+c+d} = \frac{TP+TN}{TP+TN+FP+FN}$$

Métriques pour l'évaluation

Taux d'erreur : accuracy

Quelques limitations

- On considère un problème à 2 classes avec : 9990 instances de classe 0 et 10 instances de classe 1.
- Si le modèle prédit que tout instance est de classe 0, on a

$$\texttt{Accuracy} = \frac{9990}{10000} = 99,9$$

Métriques pour l'évaluation

Recall vs precision

		PREDICTED CLASS	
		Class=Yes	Class=No
ACTUAL CLASS	Class=Yes	a (TP)	b (FN)
	Class=No	c (FP)	d (TN)

Recall (True positive rate sensitivity)

De ceux qui existent, combien l'algorithme a pu trouver $TPR = \frac{TP}{TP+FN}$

Precision

De ceux que l'algorithme a pu classer, combien sont corrects. $PPV = \frac{TP}{TP+FP}$

Métriques pour l'évaluation

F-mesure

		PREDICTED CLASS	
		Class=Yes	Class=No
ACTUAL CLASS	Class=Yes	a (TP)	b (FN)
	Class=No	c (FP)	d (TN)

Moyenne harmonique entre la precision et le rappel :

$$F1 = 2.\frac{precision.recall}{precsion + recall} = \frac{2TP}{2TP + FP + FN}$$

Courbe ROC

Définition

La courbe ROC (Receiver Operating Characteristic) dessine l'évolution du taux du vrai positif (TPR) en fonction du taux du faux positif (FPR) en faisant varier un seuillage sur la confiance (probabilité) qu'un exemple soit dans la classe positif

Rappel :
$TPR = \frac{TP}{TP+FN}, FPR = \frac{FP}{TN+FP}$

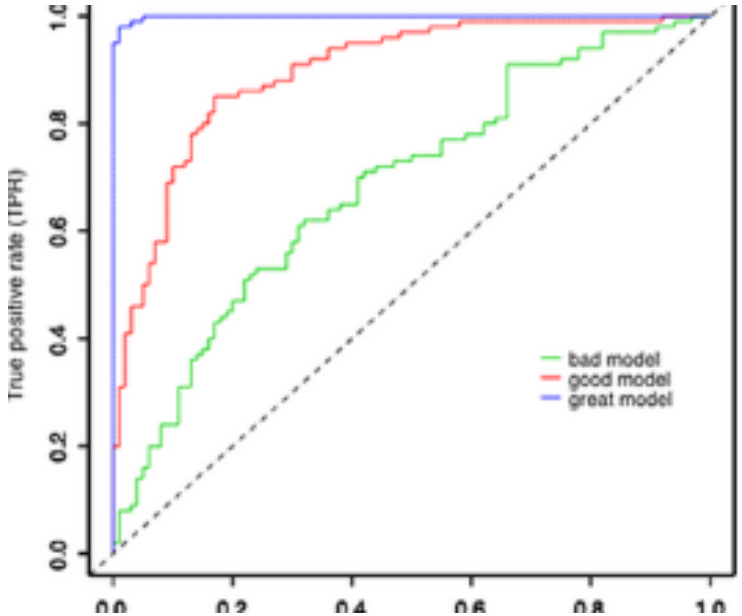
Different methods can work better in different parts of ROC space. This depends on cost of false + vs. false -

Courbe ROC

- Soit x_i positif (+) si $p(y = 1|x_i) > \theta$, sinon il est négatif (-) $(y = 0)$

$$\hat{y}_i = 1 \Leftrightarrow p(y = 1|x_i) > \theta$$

- Le nombre des TPs et FPs dépend du seuillage θ. Varier θ donne des points (TPR, FPR) différents.



ROC : exemple

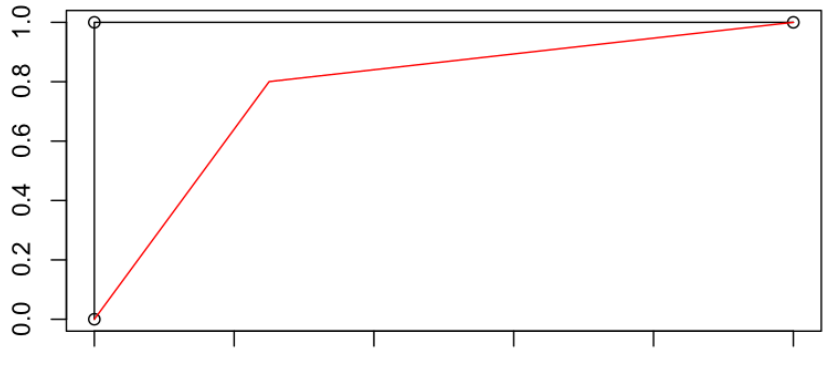
$TPR = p(\hat{y} = 1 \mid y = 1), FPR = p(\hat{y} = 1 \mid 0)$

Méthode 1

i	y_i	$p(y_i = 1 \mid x_i)$	$\hat{y}_i(\theta = 0)$	$\hat{y}_i(\theta = 0.5)$	$\hat{y}_i(\theta = 1)$
1	1	0.9	1	1	0
2	1	0.8	1	1	0
3	1	0.7	1	1	0
4	1	0.6	1	1	0
5	1	0.5	1	1	0
6	0	0.4	1	0	0
7	0	0.3	1	0	0
8	0	0.2	1	0	0
9	0	0.1	1	0	0
			TPR=5/5=1 FPR=4/4=1	TPR=5/5=1 FPR=0/4=1	TPR=0/5=0 FPR=0/4=0

Méthode 2

i	y_i	$p(y_i = 1 \mid x_i)$	$\hat{y}_i(\theta = 0)$	$\hat{y}_i(\theta = 0.5)$	$\hat{y}_i(\theta = 1)$
1	1	0.9	1	1	0
2	1	0.8	1	1	0
3	1	0.7	1	1	0
4	1	0.6	1	1	0
5	1	**0.2**	1	0	0
6	0	**0.6**	1	1	0
7	0	0.3	1	0	0
8	0	0.2	1	0	0
9	0	0.1	1	0	0
			TPR=5/5=1 FPR=4/4=1	TPR=4/5=0.8 FPR=1/4=.25	TPR=0/5=0 FPR=0/4=0

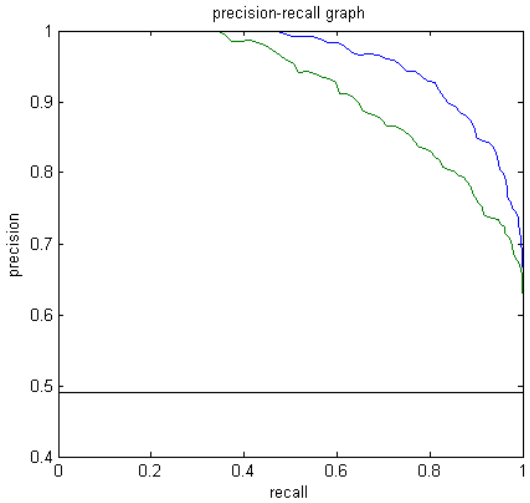


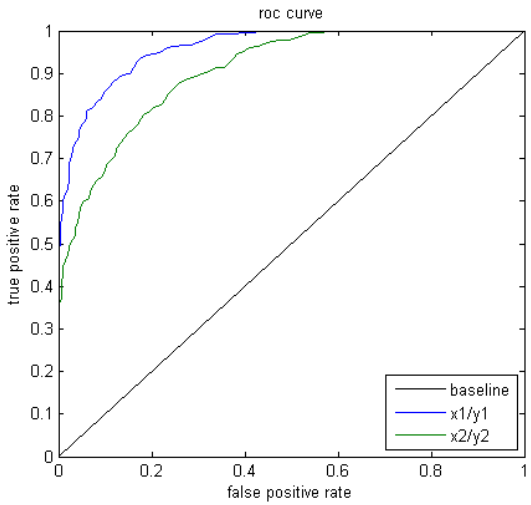
Courbe Precision/Recall

- Utile quand la notion de négatif (FPR donc) n'est pas définie ou il y a beaucoup de négatifs (détection d'évennements rares)

$$prec = p(y = 1 \mid \hat{y} = 1)$$

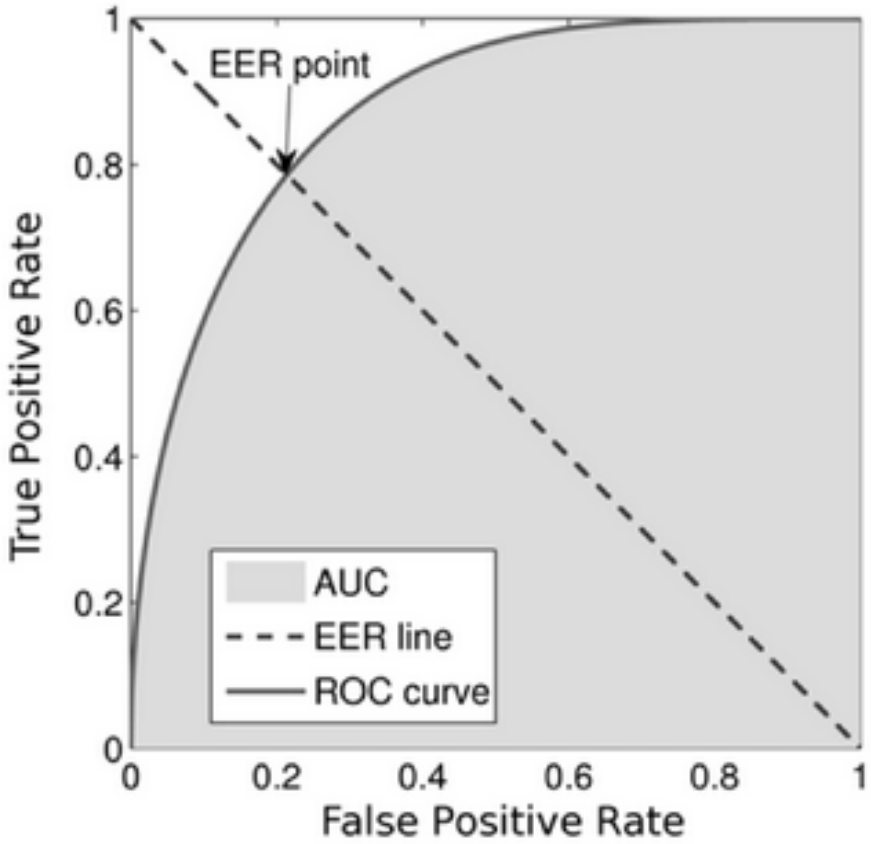
$$recall = p(\hat{y} = 1 \mid y = 1)$$





AUC et EER

- EER- Equal error rate/ cross over error rate (FPR = FNR), doit être petit
- AUC - Area under curve (aire sous la courbe), doit être large



Métriques pour l'évaluation

Matrice de cout

		PREDICTED CLASS	
ACTUAL CLASS	C(i|j)	**Class=Yes**	**Class=No**
	Class=Yes	C(Yes|Yes)	C(No|Yes)
	Class=No	C(Yes|No)	C(No|No)

$C(i|j)$: cout de mal classifier une instance de la classe j en i. La diagonale est prise comme nulle en gńérale, ou simplement $C(i \mid j) > C(i \mid i)$

Plan

1. Quelques considérations générales
   - Evaluation
   - Ré-échantillonage

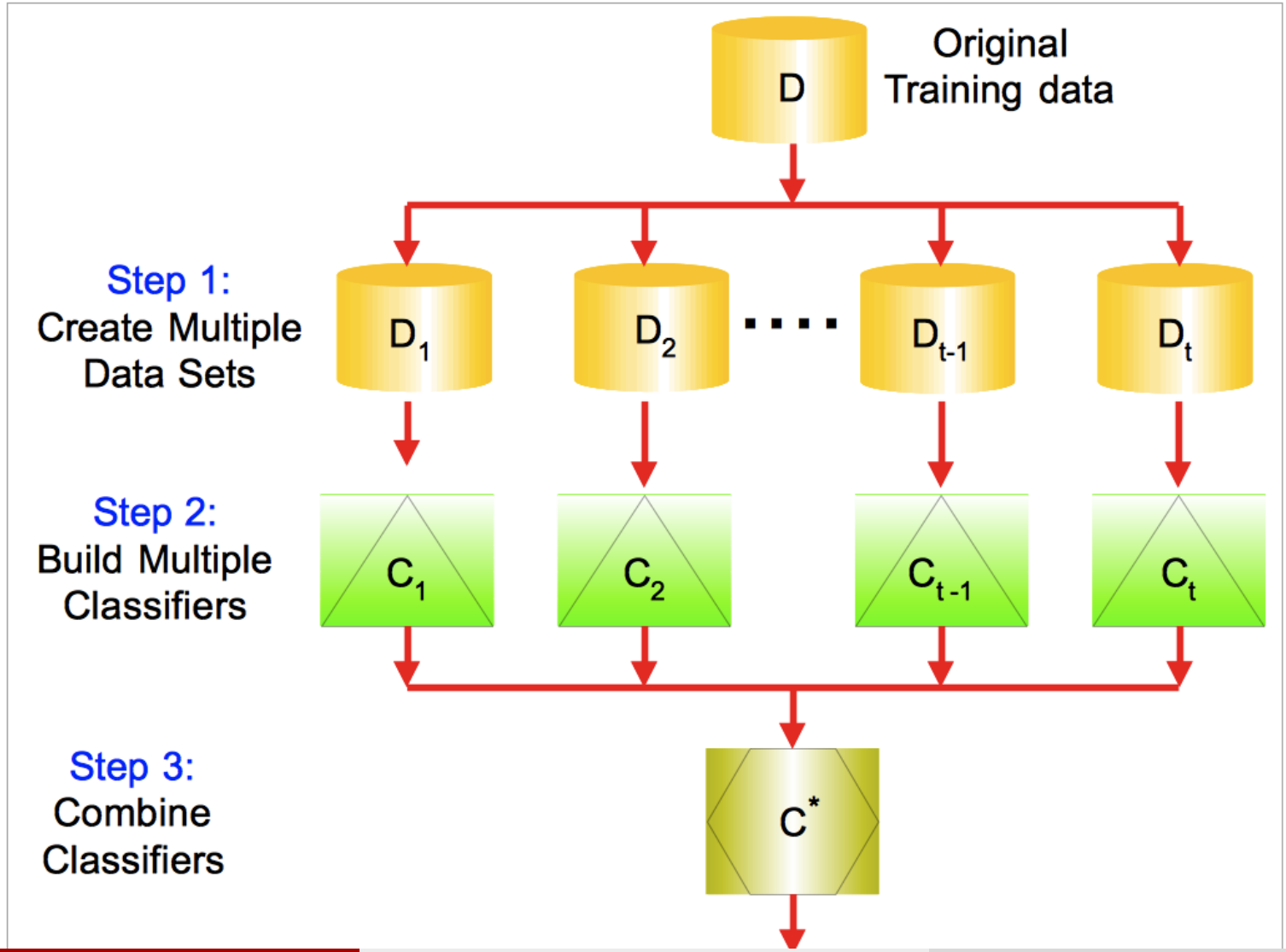
Techniques de ré-échantillonage

Le ré-échantillonage génère différents sous-ensembles de données à partir de l'ensemble initial $\mathcal{D}$.

- Pour améliorer un classifier :
  - Bagging
  - Boosting
- Pour comparer des classifieurs :
  - Validation croisée.
  - Boostrap

Amélioration des classifieurs

Idée générale



Amélioration des classifieurs par bagging

- On construit I sous-ensembles contenant m_i objets ($m_i < n$) de $\mathcal{D}$ (avec remise).
- Chaque sous-ensemble est utilisé pour entraîner un classifier.
- La décision finale est basée sur le vote des classifieurs.
- Généralement, les classifieurs sont de la même forme.
- **Avantage** : améliore la performance des classifieurs instables en moyennant leur réponse.
- Instable : petit changement dans les données $\Rightarrow$ gros changement dans le comportement du classifier.

Original Data	1	2	3	4	5	6	7	8	9	10
Bagging (Round 1)	7	8	10	8	2	5	10	10	5	9
Bagging (Round 2)	1	4	9	1	2	3	2	7	3	2
Bagging (Round 3)	1	8	5	10	5	5	9	6	3	7

Amélioration des classifieurs par boosting

- Approche collaboratrice : procédure itérative que change la distribution des données d'apprentissage en se focalisant plus sur les données mal classées à une étape précédente.
- Les classifieurs sont introduits un à la fois et travaillent sur des sous-ensembles différents.
- Chaque nouveau classifier s'occupe des cas mal compris par les autres.. les cas difficiles.
- Les classifieurs sont médiocres.
- Les classifieurs peuvent être de types différents

Amélioration des classifieurs par boosting

Procédure de boosting

A chaque objet on associe un poids. Au début, tous les objets ont le même poids.

- On construit un classifier à partir de l'ensemble pondéré.
- Les poids des objets sont modifiés en fonction du modèle construit :
  - diminution des poids des objets bien classés.
  - augmentation du poids des objets mal classés.
- On réitère le processus jusqu'à ce que le taux d'erreur soit acceptable.

Classification : vote

Original Data	1	2	3	4	5	6	7	8	9	10
Boosting (Round 1)	7	3	2	8	7	9	4	10	6	3
Boosting (Round 2)	5	4	9	4	2	5	1	7	4	2
Boosting (Round 3)	4	4	8	10	4	5	4	6	3	4

- Example 4 is hard to classify
- Its weight is increased, therefore it is more likely to be chosen again in subsequent rounds

Evaluation et comparaison des classifeurs

- Comment estimer le taux d'erreur ?
- Méthode naïve : utiliser tous les échantillons pour entraîner et calculer le taux d'erreur sur l'ensemble d'apprentissage.
- Un classifier tend à s'ajuster aux données d'apprentissage
- Un taux d'erreur généralement trop optimiste : pas rare d'avoir un taux de 0 à l'entrainement.
- Nécessité d'un ensemble de test indépendant de l'ensemble d'entrainement.
- Typiquement 10% de l'ensemble $\mathcal{D}$ pour tester.

Evaluation et comparaison des classifeurs

- Qu'arrive-t-il si on dispose de très peu d'échantillons ?
- Comment savoir si le taux d'erreur est précis ou si on est pas tombé par hasard sur un situation particulière en coupant l'ensemble $\mathcal{D}$?
- Si pour un ensemble de données $\mathcal{D}$, 2 classifieurs C_1 et C_2 ont 80% et 85% de précision, est-ce que $C_2 > C_1$?
- Solution : **Validation croisée** :
  - aléatoire
  - k- blocs
  - n-blocs (leave one out)

Evaluation et comparaison des classifeurs

Validation croisée aléatoire

- On prend aléatoirement k échantillons dans l'ensemble $\mathcal{D}$ (sans remise) pour chaque expérience.
- Le taux d'erreur est la moyenne des taux de chacune des expériences.
- La variance peut être calculée

Evaluation et comparaison des classifeurs

Validation croisée K-blocs

- On prend K ensemble disjoints de $\frac{n}{K}$ échantillons chacun
- On teste avec l'un d'entre eux.
- Taux d'erreur = moyenne des K expériences.
- La variance peut être calculée
- Avantage : tous les échantillons de $\mathcal{D}$ seront utilisés.

Evaluation et comparaison des classifeurs

Validation croisée n-blocs

- On prend un seul échantillon pour tester
- On teste avec l'un d'entre eux.
- Taux d'erreur = moyenne des n expériences.
- Avantage : utile quand $\mathcal{D}$ petit