Evaluation des classifieurs

Méthodologie de validation

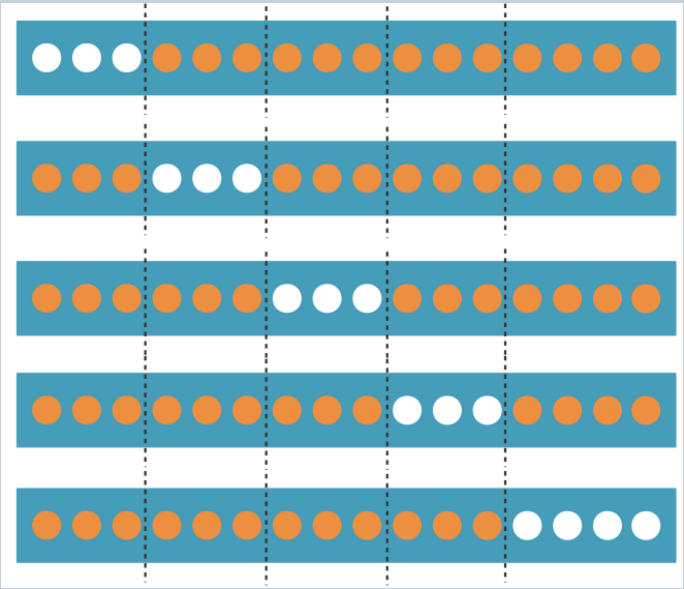
- Data = train data + test
  - Train data
    - = données pour entrainer le classifieur
  - Test data
    - = données pour évaluer le classifieur

Méthodologie de validation

- Validation simple
  - On divise les données
  - On mesure la performance sur les données test
  - Problème (← raisonnement inductif)
- Validation croisée
  - On divise n fois les données et on mesure les performances
  - Moyenne des performances
  - Techniques
    - K-fold
    - Leave one out (LOO)

Validation croisée

- Évaluer les performances sur des données non vues
  - Séparer les données en train/test set
  - Validation croisée
    - k-fold (ici k = 5)
    - Valeur habituelle : 10
    - leave one out (loo)
    - Perf finale = moyenne des 5 perf
    - Option
      - stratification



Evaluation des classifieurs

- Pourquoi ?
- Comment ?
- Matrice de confusion
  - Pour un problème à 2 classes
- Performance = proportion d'exemples bien classés

n=165	**Predicted: NO**	**Predicted: YES**	
Actual: NO	TN = 50	FP = 10	60
Actual: YES	FN = 5	TP = 100	105
	55	110	

Métriques

- Métriques
  - taux de bonnes prédictions (Accuracy)
    - Résultat partiel !
    - Exemple de classifieur
      - prénom = Lucie si et seulement si Leucémie

	Leucémie	Non Leucémie	Total
Lucie	70	4930	5000
Non Lucie	3 930	981 070	995 000
Total	4 000	996 000	1 000 000

Métriques

	Leucémie	Non Leucémie	Total
Lucie	70	4930	5000
Non Lucie	3 930	981 070	995 000
Total	4 000	996 000	1 000 000

- Accurracy
  - (70 + 981070)/100 000 = 9.82 !

Métriques

	Leucémie	Non Leucémie	Total
Lucie	70	4930	5000
Non Lucie	3 930	981 070	995 000
Total	4 000	996 000	1 000 000

- Rappel (recall)
  - Mesure à quel point les prédictions de chaque classe sont exactes (pures)
    - 70/4000 = 0.01 pour la classe Leucémie
    - 981 070/996 000 = 0.99 pour la classe Non Leucémie

Métriques

	Leucémie	Non Leucémie	Total
Lucie	70	4930	5000
Non Lucie	3 930	981 070	995 000
Total	4 000	996 000	1 000 000

- Précision (precision)
  - Mesure la proportion des exemples bien identifiés
    - 70/5000 = 0.01 pour la classe Leucémie
    - 981 070/995 000 = 0.99 pour la classe Non Leucémie

Métriques

- Combiner rappel et precision en 1 valeur
  - Score f1 : moyenne harmonique
  - f1Score = p*r/(p+r)
- Prendre en compte le coût des erreurs
  - Métrique d'erreur vs métrique d'utilité
  - Ex : filtre anti spam

Méthodes ensemblistes

- Principe
  - Produire un ensemble de n classifieurs
- Pour prédire la classe de newx :
  - Récupérer les n prédictions
  - la classe attribuée à newx est établie par un vote à la majorité des classifieurs.

Perspectives

- Courbe ROC
- Courbe d'apprentissage
- Sur-apprentissage (over fitting)

- Classifieurs usuels
  - Principe
  - forces et faiblesses
  - Paramètres et Grid Search