Data Mining & Machine Learning : Courbes ROC

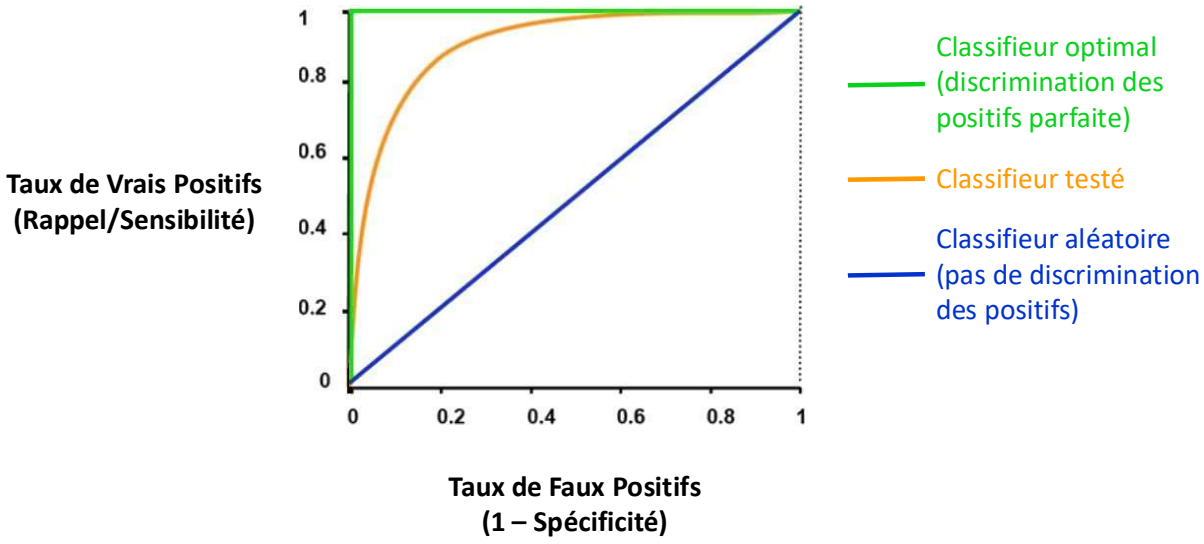
Nicolas PASQUIER
Laboratoire I3S (UMR-7271 UCA/CNRS)
Département Informatique
Université Côte d'Azur
http://www.i3s.unice.fr/~pasquier

Intérêt des Courbes ROC

- Courbe ROC : « Receiving Operating Characteristics »
- Outil de représentation graphique pour évaluer et comparer les performances de classifieurs pour la distinction d'une classe cible, appelée classe positive
- Avantages
  - Résultats valables même si l'échantillon de test n'est pas représentatif
  - Opérationnel même dans le cas de distributions très déséquilibrées
  - Indépendant des matrices de coûts : permet de savoir si un modèle sera toujours meilleur qu'un autre quelle que soit la matrice de coût
  - Les tracés des courbes ROC de différents classifieurs peuvent être affichées simultanément afin de comparer leurs performances visuellement
- Les indicateurs numériques AUC (*Area Under Curve*) et Coefficient de Gini calculés à partir de la courbe permettent d'en évaluer la qualité globale (i.e. capacité de discrimination de la classe positive)

Représentation Graphique des Courbes ROC

- La position de la courbe relativement aux axes gauche et supérieur permet d'appréhender la capacité du classifieur à discriminer la classe positive



Taux de Vrais Positifs et Taux de Faux Positifs

- Chaque point de coordonnées (TVP, TFP) de la courbe ROC est calculé à partir d'une matrice de confusion

Matrice de Confusion

		Prédiction		
		Positif	Négatif	Total
Classe réelle	Positif	VP	FN	VP + FN
	Négatif	FP	VN	FP + VN
	Total	VP + FP	VN + FN	∑

- Taux de Vrais Positifs (TVP) : VP / (VP + FN)
  - Interprétation : proportion d'exemples positifs correctement classés
  - Valeur : mesure du *Rappel*, ou *Sensibilité*
- Taux de Faux Positifs (TFP) : = FP / (FP + VN)
  - Interprétation : proportion d'exemples négatifs incorrectement classés
  - Valeur : 1 – mesure de *Spécificité*

Calcul de la Courbe ROC

- Le calcul de la courbe ROC se fait à partir des résultats du test du classifieur
- Ce calcul nécessite pour chaque exemple de l'ensemble de test
  - La classe réelle de l'exemple
  - La probabilité de prédiction positive de l'exemple
- Exemple de calcul pour les résultats de test suivants :
  - Ensemble de test de 20 exemples
  - 6 exemples positifs (classe Buyer = True)
  - 14 exemples négatifs (classe Buyer = False)
- Les exemples de test sont ordonnés par probabilités décroissantes de prédiction positive : valeur de P(True)

Tri selon P(True)

Ensemble de Test

#	Age	...	Buyer	P(True)	P(False)
1	19	...	True	1.00	0.00
2	32	...	True	0.95	0.05
3	47	...	True	0.90	0.10
4	23	...	False	0.85	0.15
5	40	...	True	0.80	0.20
6	25	...	False	0.75	0.25
7	33	...	False	0.70	0.30
8	42	...	True	0.65	0.35
9	21	...	False	0.60	0.40
10	50	...	False	0.55	0.45
11	37	...	False	0.50	0.50
12	31	...	True	0.45	0.55
13	24	...	False	0.40	0.60
14	45	...	False	0.35	0.65
15	36	...	False	0.30	0.70
16	29	...	False	0.25	0.75
17	20	...	False	0.20	0.80
18	48	...	False	0.15	0.85
19	41	...	False	0.10	0.90
20	22	...	False	0.05	0.95

Calcul des Taux de Vrais et Faux Positifs

- Interprétation intuitive : si P(True) ≥ 0.50 alors Prédiction = True
- Selon cette règle d'affectation de seuil 50 %
  - Les exemples 1 à 11 sont prédits « True »
  - Les exemples 12 à 20 sont prédits « False »

Ensemble de Test

Exemples de test prédits positifs

Seuil P(True) ≥ 0.50

Exemples de test prédits négatifs

#	Age	...	Buyer	P(True)	P(False)
1	19	...	True	1.00	0.00
2	32	...	True	0.95	0.05
3	47	...	True	0.90	0.10
4	23	...	False	0.85	0.15
5	40	...	True	0.80	0.20
6	25	...	False	0.75	0.25
7	33	...	False	0.70	0.30
8	42	...	True	0.65	0.35
9	21	...	False	0.60	0.40
10	50	...	False	0.55	0.45
11	37	...	False	0.50	0.50
12	31	...	True	0.45	0.55
13	24	...	False	0.40	0.60
14	45	...	False	0.35	0.65
15	36	...	False	0.30	0.70
16	29	...	False	0.25	0.75
17	20	...	False	0.20	0.80
18	48	...	False	0.15	0.85
19	41	...	False	0.10	0.90
20	22	...	False	0.05	0.95

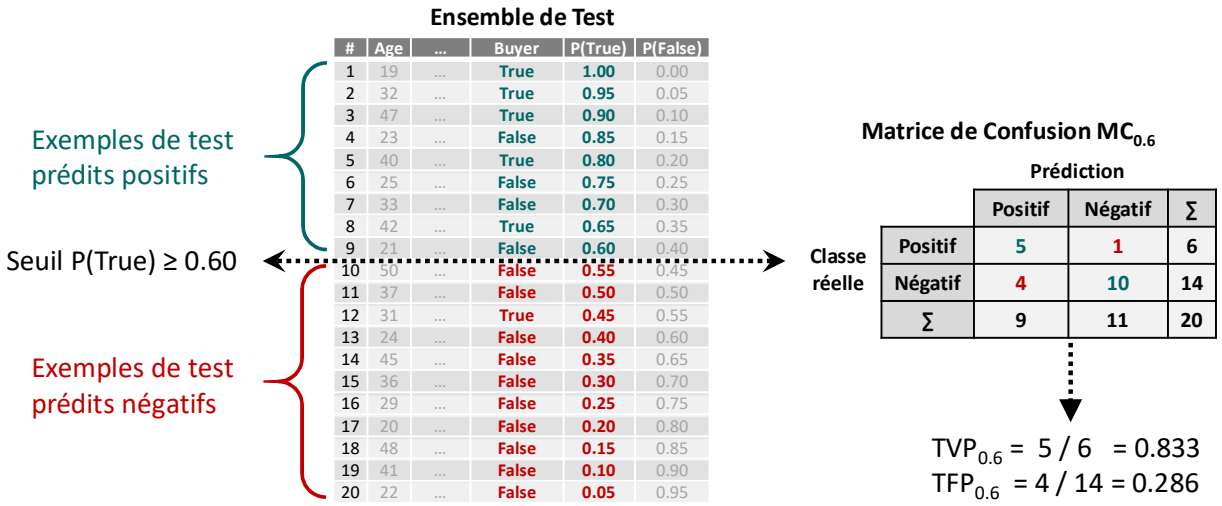
Matrice de Confusion

Classe réelle \ Prédiction	Positif	Négatif	Σ
Positif	5	1	6
Négatif	6	8	14
Σ	11	9	20

TVP = 5 / 6 = 0.833
TFP = 6 / 14 = 0.429

Calcul des Taux de Vrais et Faux Positifs

- En faisant varier le Seuil, par exemple si P(True) ≥ 0.6 alors Prédiction = True, nous obtenons une autre matrice de confusion $MC_{0.6}$ et les deux indicateurs associés $TVP_{0.6}$ et $TFP_{0.6}$



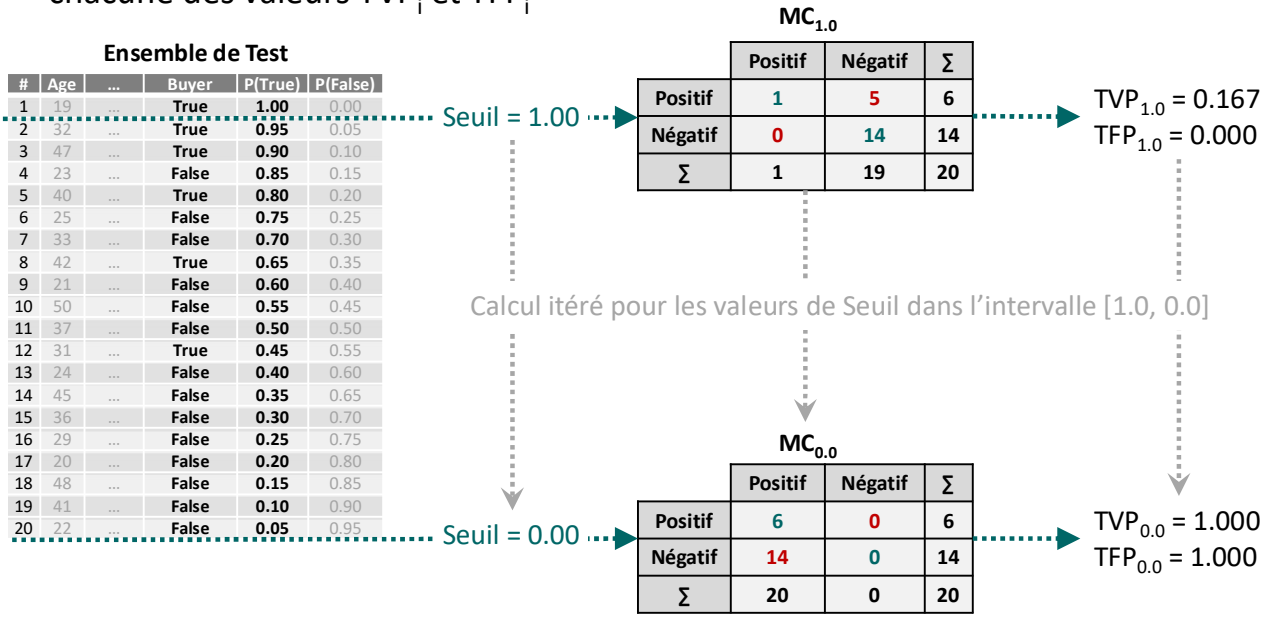
#	Age	...	Buyer	P(True)	P(False)
1	19	...	True	1.00	0.00
2	32	...	True	0.95	0.05
3	47	...	True	0.90	0.10
4	23	...	False	0.85	0.15
5	40	...	True	0.80	0.20
6	25	...	False	0.75	0.25
7	33	...	False	0.70	0.30
8	42	...	True	0.65	0.35
9	21	...	False	0.60	0.40
10	50	...	False	0.55	0.45
11	37	...	False	0.50	0.50
12	31	...	True	0.45	0.55
13	24	...	False	0.40	0.60
14	45	...	False	0.35	0.65
15	36	...	False	0.30	0.70
16	29	...	False	0.25	0.75
17	20	...	False	0.20	0.80
18	48	...	False	0.15	0.85
19	41	...	False	0.10	0.90
20	22	...	False	0.05	0.95

	Positif	Négatif	∑
Positif	5	1	6
Négatif	4	10	14
∑	9	11	20

- Cela équivaut à augmenter le poids des faux positifs d'une matrice de confusion

Calcul des Taux de Vrais et Faux Positifs

- Les points de la courbe (TFP_i, TVP_i) sont calculés en en faisant varier le Seuil entre 1.0 et 0.0 afin d'obtenir une suite de matrices de confusion MC_i et pour chacune des valeurs TVP_i et TFP_i



#	Age	...	Buyer	P(True)	P(False)
1	19	...	True	1.00	0.00
2	32	...	True	0.95	0.05
3	47	...	True	0.90	0.10
4	23	...	False	0.85	0.15
5	40	...	True	0.80	0.20
6	25	...	False	0.75	0.25
7	33	...	False	0.70	0.30
8	42	...	True	0.65	0.35
9	21	...	False	0.60	0.40
10	50	...	False	0.55	0.45
11	37	...	False	0.50	0.50
12	31	...	True	0.45	0.55
13	24	...	False	0.40	0.60
14	45	...	False	0.35	0.65
15	36	...	False	0.30	0.70
16	29	...	False	0.25	0.75
17	20	...	False	0.20	0.80
18	48	...	False	0.15	0.85
19	41	...	False	0.10	0.90
20	22	...	False	0.05	0.95

	Positif	Négatif	∑
Positif	1	5	6
Négatif	0	14	14
∑	1	19	20

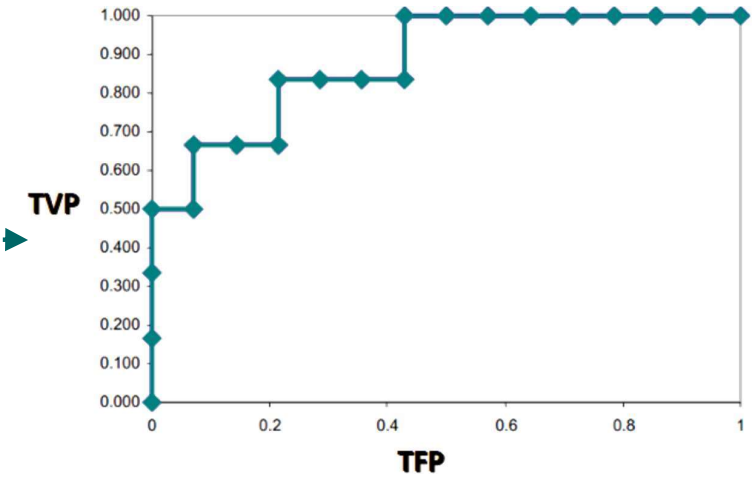
	Positif	Négatif	∑
Positif	6	0	6
Négatif	14	0	14
∑	20	0	20

Représentation de la Courbe ROC

- Les points de la courbe seront tracés avec
  - TVP_i pour ordonnée
  - TFP_i pour abscisse

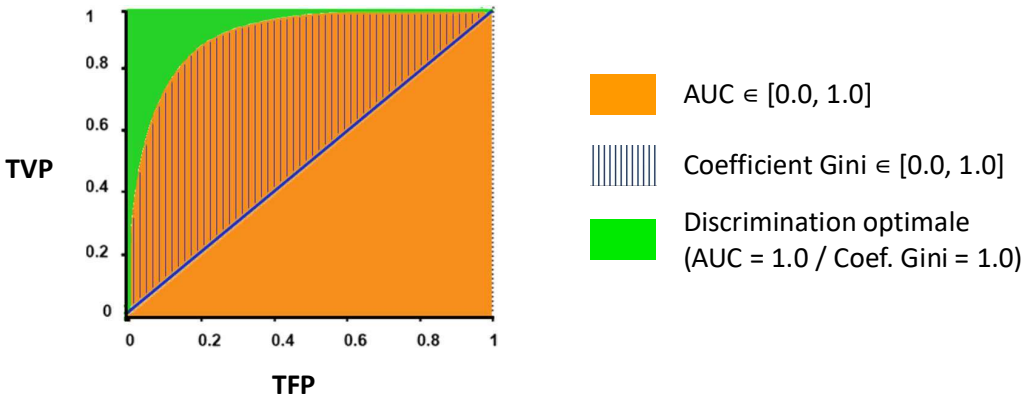
Ensemble de Test

#	Age	...	Buyer	P(True)	P(False)	TVP	TFP
1	19	...	True	1.00	0.00	0.167	0.000
2	32	...	True	0.95	0.05	0.333	0.000
3	47	...	True	0.90	0.10	0.500	0.000
4	23	...	False	0.85	0.15	0.500	0.071
5	40	...	True	0.80	0.20	0.667	0.071
6	25	...	False	0.75	0.25	0.667	0.143
7	33	...	False	0.70	0.30	0.667	0.214
8	42	...	True	0.65	0.35	0.833	0.214
9	21	...	False	0.60	0.40	0.833	0.286
10	50	...	False	0.55	0.45	0.833	0.357
11	37	...	False	0.50	0.50	0.833	0.429
12	31	...	True	0.45	0.55	1.000	0.429
13	24	...	False	0.40	0.60	1.000	0.500
14	45	...	False	0.35	0.65	1.000	0.571
15	36	...	False	0.30	0.70	1.000	0.643
16	29	...	False	0.25	0.75	1.000	0.714
17	20	...	False	0.20	0.80	1.000	0.786
18	48	...	False	0.15	0.85	1.000	0.857
19	41	...	False	0.10	0.90	1.000	0.929
20	22	...	False	0.05	0.95	1.000	1.000



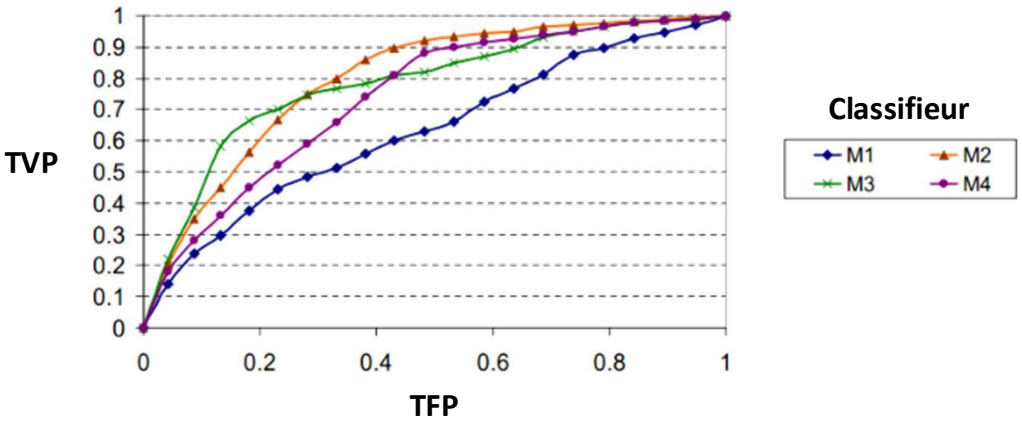
Indicateurs AUC et Coefficient de Gini

- AUC (*Area Under the Curve*) : aire sous la courbe ROC
  - Proportion représentée par l'aire sous la courbe ROC relativement à la surface totale de l'espace du graphique
- Indicateur équivalent : Coefficient de Gini
  - Proportion représentée par l'aire entre la courbe ROC et la diagonale de l'espace (modèle classant les exemples au hasard)



Interprétation des Courbes ROC

- Enveloppe convexe : formée par les courbes dont aucune autre courbe n'est au-dessus sur une portion donnée
- Les modèles correspondant aux courbes situées sur cette enveloppe convexe sont ceux potentiellement les plus performants
- Exemple :



- Enveloppe convexe : M2 et M3 sont les classifieurs les plus performants

Références et Bibliographie

- Principales Librairies R
  - ROCR : visualisation et évaluation des performances de classifieurs
  - pROC : analyse et affichage de courbes ROC
- Bibliographie
  - Data Classification: Algorithms and Applications. Chapter 24 (Evaluation of Classification Methods). Charu C. Aggarwal. Chapman and Hall/CRC, 2014. ISBN 978-1-466-58674-1
  - Data Mining Applications with R. Chapter 6 (Response Modeling in Direct Marketing) and Chapter 7 (Caravan Insurance Customer Profile Modeling with R). Yanchang Zhao & Yonghua Cen. Academic Press, 2014. ISBN 978-0-124-11511-8
  - ROC graphs: Notes and practical considerations for researchers. Tom Fawcett. Machine Learning, vol. 31, num. 1, pages 1-38, 2004.