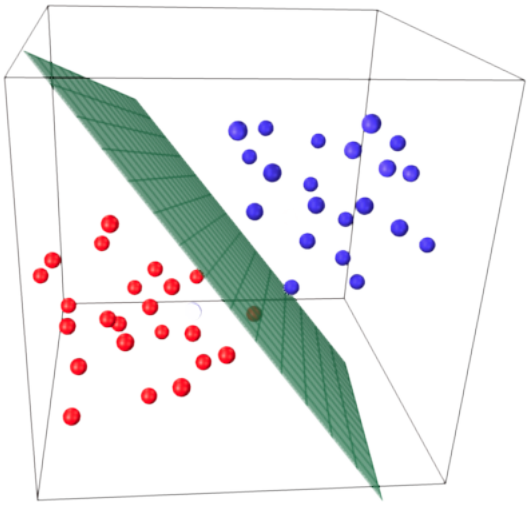
Rapport TPE - Régression Logistique

Lucien Mauffret et Yoann Pruvot

Mai 2019

1 Introduction

Dans le cadre du Master 1 ISIFAR, il est proposé un travail personnel encadré par deux enseignants, travail invitant à approfondir ou découvrir une méthode statistique.
Nous avons choisi de nous pencher sur la régression logistique, en adoptant une approche pédagogique, passant d'abord par une intuition du sujet traité, ensuite par une formalisation de ce dernier, et enfin appliquer la régression logistique à des cas spécifiques d'applications.



Les exemples ont été choisis avec en tête deux choses :

- Le premier, que l'on a appélé "Marketing Sport" nous met en situation dans la peau d'un data analyst qui est en support d'une campagne de promotion marketing afin de déterminer s'il faut cibler un type de client ou non. Cela constitue une situation classique ou la régression logistique trouve son sens.
- Le deuxième, que l'on a appelé "Banque de Dépôt à Terme" cherche à savoir comment mieux cibler, en tant que banque, dans le cadre de la vente de dépôt à terme, les clients à appeler pour optimiser ses recettes. C'est une situation que nous avons retenu pour sa proximité avec les sujets traités au cours du Master ISIFAR.

Nous espérons que la lecture de ce rapport soit tout aussi claire qu'agréable.

2 Intuition

La régression logistique est une méthode de classification d'une variable qualitative à modalité binaire. Plus précisement, en considérant $X = (X^1, ..., X^m)$ qualitatives ou quantitatives et Y qui est qualitative à modalité binaire, on souhaite connaître la modalité de Y_{n+1} .
On peut représenter sous forme de tableau de données :

X^1	...	X^m	Y
X_1^1	...	X_1^m	$Y_1 \in \{0,1\}$
...	...	...	...
X_n^1	...	X_n^m	$Y_n \in \{0,1\}$

On souhaite donc déterminer à partir de ces données Y_{n+1} . Cette modélisation est très utilisée par exemple en médecine pour essayer de prédire, en fonction de différents paramètres $(X_{n+1}^1, ..., X_{n+1}^m)$ d'un patient la présence ou non d'une maladie ou d'un virus, représenté par un codage binaire $\{0, 1\}$.
L'idée est de se dire que l'on peut peut-être faire une régression linéaire sur Y de la forme :

$$Y = a_0 + a_1 X^1 + ... + a_m X^m$$

Le problème est que Y étant à modalité binaire (0 ou 1), on ne peut trouver de droite de régression qui modélise correctement notre modèle. On va donc commencer par poser, en considérant $P(Y = 1|X) = p$, $\frac{p}{1-p}$ qui prend ses valeurs sur $]0; \infty[$. Toutefois, en regardant notre modélisation de Y, on remarque que la constante a_0 nous impose de prendre des valeurs sur $]-\infty; \infty[$. Pour cela, on va utiliser la fonction $x \mapsto ln(x)$.
On va donc travailler, dans le cadre de la régression logistique sur :

$$ln(\tfrac{p}{1-p}) = a_0 + a_1 X^1 + ... + a_m X^m$$

L'idée est alors d'estimer tous ces paramètres. Une fois estimés, on peut alors prendre comme règle de décision :

$$\begin{cases} \text{p} > 0.5 \rightarrow \text{Y} = 1 \\ \text{p} \leq 0.5 \rightarrow \text{Y} = 0 \end{cases}$$

ou bien encore :

$$\begin{cases} \text{p}_{\frac{}{1-p}} > 1 \rightarrow \text{Y} = 1 \\ \text{p}_{\frac{}{1-p}} \leq 0.5 \rightarrow \text{Y} = 0 \end{cases}$$

Avant de commencer, remarquons que le rapport $\frac{p}{1-p}$ est ce que l'on appelle en anglais l' "odds", autrement dit la cote, qui est le rapport de probabilté que l'évènement Y arrive contre la probabilité qu'il n'arrive pas. Si il y a $\frac{3}{4}$ que Y arrive contre $\frac{1}{4}$ qu'il n'arrive pas, sa cote est de 3 contre 1. Tout au long de ce rapport, on s'efforcera de placer au centre de notre propos la pédagogie et la clarté.

3 Mise en place du modèle

3.1 Notation

On pose :

- $\theta = (\alpha, \beta)$
- $W = (1, X)$
- $\alpha + {}^t\beta X = {}^t\theta W$
- $L(\theta)$: la fonction de vraissemblance de θ
- $H(\theta)$: la hessienne de $L(\theta)$
- $\nabla(\theta) =$: le gradient de $L(\theta)$
- π : $P(Y = 1|X)$

3.2 Bayes et équivalence des approches

Rappelons que d'après la *formule de Bayes*:

$$\pi = \frac{P(Y=1)P(X|Y)}{P(X)}$$

En appliquant la même formule à $P(Y = 0|X)$, et en divisant les deux, on obtient :

$$\frac{\pi}{1-\pi} = \frac{P(Y=1)}{P(Y=0)} \frac{P(X|Y=1)}{P(X|Y=0)}$$

Enfin, on peut remarquer que :

$$\ln(\frac{\pi}{1-\pi}) = a_0 + \sum_{k=1}^{m} a_k X^k$$

a une solution qui est :

$$\pi = \frac{\exp^{a_0+\sum_{k=1}^{m} {}^{a_k} X^k}}{1+\exp^{a_0+\sum_{k=1}^{m} {}^{a_k} X^k}} \text{ et donc } 1 - \pi = \frac{1}{1+\exp^{a_0+\sum_{k=1}^{m} {}^{a_k} X^k}}$$

La transformation $p \mapsto \ln(\frac{p}{1-p})$ est appelée fonction $LOGIT$ de p, qui donnera le nom de régression *logistique*.
On remarque qu'il y a équivalence en terme de fonction LOGIT pour $\frac{\pi}{1-\pi}$ et $\frac{P(X|Y=1)}{P(X|Y=0)}$. En effet, s'il on pose:

- $\ln(\frac{\pi}{1-\pi}) = a_0 + \sum_{k=1}^{k=m} a_k X^k$
- $ln(\frac{P(X|Y=1)}{P(X|Y=0))} = b_0 + \sum_{k=1}^{k=m} b_k X^k$

On a alors :

$$\begin{cases} \mathrm{a}_0 & = & ln(\frac{P(Y=1)}{P(Y=0)}) + \mathrm{b}_0 \\ \mathrm{a}_k & = & \mathrm{b}_k, \; k \geq 1 \end{cases}$$

3.3 LOGIT d'un individu

Pour un individu ω, on défini son LOGIT :

$$\ln(\frac{\pi}{1-\pi}) = a_0 + \sum_{k=1}^{m} a_k X^k$$

C'est en fait l'hypothèse principale de la régression logistique. Beaucoup de distributions vérifient cette hypothèse:

- loi normale
- loi exponentielle
- lois discrètes
- loi gamma, beta, poisson

On se retrouve avec une fonction de densité nous permettant d'estimer la quantité $\frac{\pi}{1-\pi}$.

3.4 Modèle binomial et Vraissemblance

On peut modéliser, pour un individu ω, la probabilité $P(Y|X)$ par:

$$\pi(w)^{Y(w)}(1-\pi(w))^{1-Y(w)}$$

On peut donc écrire la vraissemblance associée (pour simplifier on n'écrit pas ω):

$$L = \prod_{i=1}^{n} \pi_i^{Y_i}(1-\pi_i)^{1-Y_i}$$

Ce qui nous donne une log-vraissemblance:

$$ln(L) = \sum_{i=1}^{n} \ln(\pi_i)Y_i + (1-Y_i)\ln(1-\pi_i)$$

On cherche alors à maximiser notre log-vraissemblance. Pour cela, on va s'appuyer sur la méthode de Newton-Raphson.

4 Résolution et algorithme de Newton-Raphson

4.1 Origines

La question que l'on se pose est la suivante : considérant une fonction $x \mapsto f(x)$, peut-on trouver un zéro de cette fonction ? L'idée est la suivante : on approxime $f(x)$ (on suppose f dérivable en x, on part de x_0) ainsi:

$$f(x) \approx f(x_0) + (x - x_0)f'(x_0)$$

Et en cherchant à annuler cette approximation, on veut donc:

$$0 = f(x_0) + (x - x_0)f'(x_0)$$

d'où le modèle numérique:

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}$$

On s'arrête dès que l'on trouve le zéro de notre fonction ou en fixant un seuil à ne pas dépasser.

4.2 Notre modèle

Dans notre cas, la méthode de Newton-Raphson permet de trouver le zéro de $\theta \mapsto L(\theta)$. le modèle se présente sous sa forme multidimensionnelle sous cette forme:

$$\begin{cases} \theta_0 &= 0 \\ \theta_{k+1} &= \theta_k - H(\theta)^{-1}\nabla(\theta) \end{cases}$$

On s'arrête dès que l'on trouve notre θ maximisant notre fonction de vraissemblance, ou au seuil que l'on fixe. Ainsi, on a le meilleur estimateur $\hat{\theta}$ de $\theta = (\alpha, \beta)$.
Dans cet méthode, on est amené à se poser la question de la forme de $\partial_j ln(L)$ et $\partial_j \partial_k ln(L)$, qui se calcule de manière directe:

$$\begin{cases} \partial_j \ln(L(\theta)) &= \sum_{i=1}^n W_i^j (Y_i - \pi_i) \\ \partial_j \partial_k \ln(L(\theta)) &= \sum_{i=1}^n W_i^j W_i^k \pi_i (1-\pi_i) \end{cases}$$

Le $\hat{\theta}$ est en fait une estimation de π, que l'on appelle communément "score", qui est une estimation de $P(Y=1|X)$.

$$\begin{aligned} \ln(L(\theta)) &= \sum_{i=1}^n Y_i \ln(\tfrac{e^{{}^t\theta W_i}}{1+e^{{}^t\theta W_i}}) + (1-Y_i)\ln(1 - \tfrac{e^{{}^t\theta W_i}}{1+e^{{}^t\theta W_i}}) \\ &= \sum_{i=1}^n Y_i({}^t\theta W_i - \ln(1+e^{{}^t\theta W_i})) + (1-Y_i)(-\ln(1+e^{{}^t\theta W_i})) \\ &= \sum_{i=1}^n Y_i {}^t\theta W_i - \ln(1+e^{{}^t\theta W_i}) \end{aligned}$$

On a $\partial_j({}^t\theta W_i) = \partial_j(\theta_0 W_i^0 + ... + \theta_m W_i^m) = W_i^j$
$\partial_u(ln(u)) = \frac{u'}{u}$ *et aussi* $\partial_j(1+e^{{}^t\theta W_i}) = \partial_j(1+e^{\theta_0 W_i^0 + ... + \theta_m W_i^m}) = W_i^j e^{{}^t\theta W_i}$

D'où,

$$\begin{aligned} \partial_j \ln(L(\theta)) &= \sum_{i=1}^n Y_i W_i^j - W_i^j \tfrac{e^{{}^t\theta W_i}}{1+e^{{}^t\theta W_i}} \\ &= \sum_{i=1}^n W_i^j (Y_i - \tfrac{1}{1+e^{-{}^t\theta W_i}}) \\ &= \sum_{i=1}^n W_i^j (Y_i - \pi_i) \end{aligned}$$

$$\begin{aligned} \partial_j \partial_k \ln(L(\theta)) &= \sum_{i=1}^n -W_i^j W_i^k \tfrac{e^{-{}^t\theta W_i}}{(1+e^{-{}^t\theta W_i})^2} \\ &= \sum_{i=1}^n W_i^j W_i^k \pi_i (1-\pi_i) \end{aligned}$$

5 Evaluation du modèle

5.1 Matrice de confusion

Pour évaluer la qualité du modèle, on se propose de mettre en place la matrice de confusion de ce dernier :

	Faux	Vrai
Faux	n_{11}	n_{12}
Vrai	n_{21}	n_{22}

On a alors :

- n_{11} : vrais négatifs
- n_{12} : faux négatifs
- n_{21} : faux positifs
- n_{22} : vrais positifs

On nomme également les coefficients suivants :

- sensibilité = $\frac{n_{22}}{n_{21}+n_{22}}$
- spécificité = $\frac{n_{11}}{n_{11}+n_{12}}$
- taux d'erreur = $\frac{n_{12}+n_{21}}{n}$

On va alors chercher à :

- maximiser la sensibilité et la spécificité
- minimiser le taux d'erreur

6 Optimisation du modèle et test de Wald

La question qui se pose est de savoir si notre modèle est idéal ou non. Intuitivement, on va se positionner de deux manières différentes : ascendante ou descendante. L'idée est alors de tester dans le cadre ascendant (respectectivement descendante) la significativité de chaque variable au modèle et de l'ajouter (respectivement l'enlever) si celle ci l'est (respectectivement si elle ne l'est pas).
D'un point de vu mathématiques, on va chercher à tester $\forall \hat{\theta}_i$, $i \in [1; m+1]$, (on note $\hat{\theta}_i = \theta_i$):

$$\begin{cases} \mathrm{H}_0 & : \quad \theta_i = 0 \\ \mathrm{H}_1 & : \quad \theta_i \neq 0 \end{cases}$$

On considère la statistique de Wald pour notre décision :

$$W = \frac{(\theta_i)^2}{Var(\theta_i)}$$

Enfin, on gardera la variable X^i si on a $W > t_{1-\alpha}$, $t_{1-\alpha}$ étant le quantile de niveau α, α étant l'erreur de première espèce pour une loi χ^2_1 (loi du chi deux à un degré de liberté). Pour calculer $Var(\theta)$, on peut passer (c'est ce qui est fait lors des études ci dessous) par une méthode Bootstrap.

7 Point de méthodologie

Dans les faits, à partir d'une base de données, on séparera systématiquement ce que l'on appelle plus communément notre "dataset" en deux : le "train set" et le "test set" qui correspondent dans notre cas à 70% du dataset contre 30%, et ceci appliqué donc à X et y. Cette méthodologie nous permet donc de tester notre modèle sur des données "vraies" et le vérifier !

8 Cas d'étude 1 : Marketing Sport

Cet algorithme de classification est beaucoup utilisé dans la cadre du marketing. Cela permet, par exemple à la suite d'un sondage, de se positionner quant à la réalisation d'une offre spécifique. On peut alors mieux cibler le marché viser et rendre plus éfficace l'opération marketing.
Ici, on se mettra dans la peau d'une entreprise vendant des accessoires de sport.

8.1 Présentation des données

Suite au sondage, les données récupérées sont (age, sexe, nivtud, qualif, freres.soeurs, relig, trav.imp, trav.statis, hard.rock, lecture.bd, peche.chasse, cuisine, bricol, cinema, mins.tv, sport). Ces variables correspondent à des caractérisques de la vie courante des sondés au nombre de 934.
Ci dessous est présenté les 5 premières lignes de notre tableau de données.

	age	sexe	nivetud	qualif	freres.soeurs	relig	trav.imp	trav.satisf	hard.rock	lecture.bd	peche.chasse	cuisine	bricol	cinema	sport	heures.tv
0	28	Femme	superieur	Employe	8	non croyant	peu	Insatisfaction	0	0	0	1	0	0	0	0
1	59	Homme	primaire	Profession intermediaire	2	non croyant	moyen	Equilibre	0	0	0	0	0	0	1	0
2	34	Homme	superieur	Profession intermediaire	1	pas pratiquant	peu	Satisfaction	0	0	0	1	1	1	1	120
3	35	Femme	technique/professionnel	Employe	5	non croyant	tres	Equilibre	0	0	0	0	0	1	1	120
4	47	Homme	technique/professionnel	Ouvrier	5	non croyant	peu	Insatisfaction	0	0	1	1	1	0	0	60

Figure 1: 5 premières lignes de Marketing Sport

La question est de pouvoir prédire, en tant que spécialiste de la donnée, s'il est intéressant pour notre entreprise de faire des promotions sur nos produits de sport à de potentiels clients compte tenu de ces différentes caractéristiques. On se propose alors de mettre en place un modèle de régression logistique, que l'on a codé auparavant avec y = "sport", X = le reste de la matrice. Pour ce faire, on transforme nos données qualitatives en $0, 1$ correspondant à la présence ou non d'une caractéristique (procédé fait en deux temps). On applique alors notre modèle (sur donc 28 variables désormais).

8.2 Etude des résultats

L'estimateur du maximum de vraissemblance est alors calculé et vaut :

```
Les paramètres trouvés sont : [ 0.27654353 -0.03479754 -0.03211977 -1.12758306  0.18333281  0.09822498
  0.54873848  0.13428818  0.35739893 -0.00241792 -0.23196776  0.54217944
 -0.55283866  0.32128661  0.61761443 -0.07343462  0.44361962  0.0046725
 -0.41690132  0.29305276  0.01423566  0.11021352  0.19300573  0.10498926
  0.30473404 -0.1398927   0.05876085  0.39587757 -0.09231756]
```

Figure 2: Estimateur du Maximum de Vraisemblance - lecture de gauche à droite et de haut en bas

On obtient de plus une précision de : 68.7%.
La matrice de confusion associé est alors :

	Faux	Vrai
Faux	134	29
Vrai	47	71

ce qui nous permet de calculer :

- la spécificité : 82.2%
- la sensibilité : 60.2%

8.3 Optimisation du modèle

Afin d'essayer d'optimiser notre modèle, on se permet d'appliquer le test de Wald sur nos résultats précédents. A l'aide de la méthode Bootstrap, on obtien une variance de θ qui est :

```
[[0.013500, 0.000068, 0.001199, 17.720425, 0.195664,
0.049004, 0.035188, 0.025052, 0.021599, 0.000001,
0.017189, 0.018291, 0.043804, 0.033054, 0.028519,
0.016686, 0.034182, 0.020675, 0.031785, 0.026594,
0.014044, 0.014391, 0.022275, 0.044788, 0.034244,
0.245092, 0.014847, 0.033056, 0.016762]]
```

Figure 3: Variance de θ

et ainsi, en effectuant le test de Wald avec un α fixé à 5%, on ne doit plus garder que les variables (age, cuisine, cinema, heures.tv, sexe_Homme, nivetud_primaire, nivetud_superieur, qualif_Cadre, qualif_Ouvrier, trav.satisf_Insatisfaction) nous donnant finalement un modèle ayant comme nouveau θ, pour une précision de 69.4%:

```
Les paramètres trouvés sont : [ 0.32576932 -0.03078941  0.5675411   0.37894597 -0.00253558  0.81819537
 -0.66543561  0.61407921  0.29798258 -0.57175715  0.41106582]
```

Figure 4: Nouveau θ post Wald

On a donc ainsi un modèle prédictif de classification relativement performant, nous permettant de pouvoir optimiser la prochaine campagne marketing et ainsi faire faire du bénéfice à notre entreprise.

9 Cas d'étude 2 : Banque de Dépot à Terme

Dans ce deuxième cas d'étude, on se positionne en tant que banque qui voudrait optimiser le ciblage de leur proposition d'offre de dépôt à terme. Les données sont issues de plusieurs campagnes d'appels d'une banque portugaise à des particuliers.

9.1 Présentation des données

Suite au sondage, les données récupérées sont (age, job, marital, education, default, balance, housing, loan, contact, month, duration, campaign, pdays, previous, poutcome, subscribed). Ces variables correspondent à des caractérisques des appels et informations personnelles effectuées auprès des sondés au nombre de 45211. Ci dessous est présenté les 5 premières lignes de notre tableau de données.

	age	job	marital	education	default	balance	housing	loan	contact	month	duration	campaign	pdays	previous	poutcome	y
0	58	management	married	tertiary	0	2143	1	0	unknown	may	261	1	0	0	unknown	0
1	44	technician	single	secondary	0	29	1	0	unknown	may	151	1	0	0	unknown	0
2	33	entrepreneur	married	secondary	0	2	1	1	unknown	may	76	1	0	0	unknown	0
3	47	blue-collar	married	unknown	0	1506	1	0	unknown	may	92	1	0	0	unknown	0
4	33	unknown	single	unknown	0	1	0	0	unknown	may	198	1	0	0	unknown	0

Figure 5: 5 premières lignes de Banque de Dépôt à Terme

La question est, encore une fois, de pouvoir prédire, pour quels clients potentiels il est intéressant d'appeler afin d'optimiser les ventes de dépôt à terme. On se propose toujours de mettre en place un modèle de régression logistique, avec y = "subscribed", X = le reste de la matrice. Pour ce faire, on transforme encore nos données qualitatives en $0, 1$ correspondant à la présence ou non d'une caractéristique. On applique alors notre modèle.

9.2 Etude des résultats

L'estimateur du maximum de vraissemblance est alors calculé et vaut :

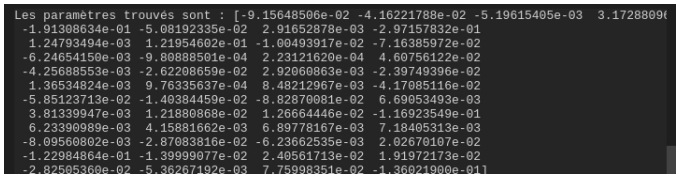


Figure 6: Estimateur du Maximum de Vraisemblance - lecture de gauche à droite et de haut en bas

On obtient de plus une précision de : 88.77%. La matrice de confusion associé est alors :

	Faux	Vrai
Faux	11694	250
Vrai	1322	298

ce qui nous permet de calculer :

- la spécificité : 97.8%
- la sensibilité : 18.4%

9.3 Optimisation du modèle

Afin d'essayer notre modèle, on se permet d'appliquer le test de Wald sur nos résultats précédents. On obtient donc une variance de θ qui est :

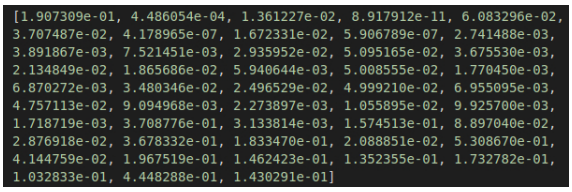


Figure 7: Variance de θ

et ainsi, avec un α fixé à 5%, on ne doit plus garder que les variables (age, duration, campaign, previous), nous donnant finalement un modèle ayant comme nouveau θ, pour une précision de 88.78%:

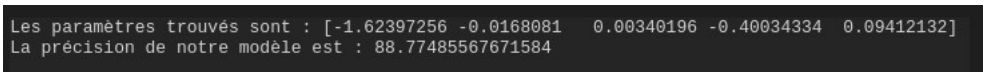


Figure 8: Nouveau θ post Wald

On a donc ainsi un modèle prédictif de classification relativement performant, nous permettant de pouvoir optimiser le ciblage de notre offre de dépôt à terme et rendre plus performante notre banque. Toutefois, on note deux choses :

- le test de Wald ne semble pas augmenter significativement la précision de notre modèle mais sert à enlever des variables
- on a une très faible sensibilité compensé par une très grande spécificité

On pourrait penser que c'est en parti dû à la faible proportion de $Y_i = 1$, environ 10%.

10 Autres tests et indicateurs

Afin d'ouvrir le sujet, on se propose de présenter quelques autres méthodes utilisées tout au long de la régression logistique. On trouve dans la littérature d'autres critères et indicateurs nous permettant de comprendre et d'optimiser notre modèle au mieux.
En premier lieu vient la courbe ROC, qui consiste en la chose suivante : à la suite de la matrice de confusion, une courbe ROC (Receiver Operating Characteristic) afin de mieux visualiser la qualité de notre test. On cherchera à maximiser l'aire entre la courbe ROC (x : spécificité, y : sensibilité) et la courbe représentative de $x \mapsto f(x) = x$. Une courbe ROC ressemble grossièrement à ça :

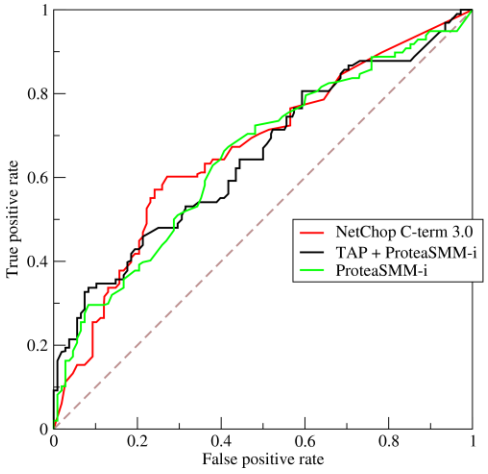


Figure 9: Exemple de courbe ROC - source : wikipédia

On pourrait également penser à mettre en place des indicateurs qui ressemblerait à un R^2 dans le cadre d'une régression linéaire. On appelle ces indicateurs les pseudo-R^2 :

- $R^2_{MF} = 1 - \frac{ln(L(\theta}{ln(L(0)}$
- $R^2_{CS} = 1 - (\frac{L(0)}{L(\theta)})^{\frac{2}{n}}$
- $R^2_N = \frac{R^2_{CS}}{max[R^2_{CS}]}$

De plus, à la place de Wald, on peut utiliser le critère d'information d'Akaike, dit AIC, communément utilisé, consistant en la chose suivante : en considérant un ensemble de modèle et le critère $AIC = 2k - 2ln(L)$ (k est nombre de paramètre, L la vraissemblance du modèle), on choisira le modèle qui aura le plus petit AIC. Il existe plusieurs version de l'AIC (AIC corrigé par exemple).