Choix de modèles

Gilbert Saporta
Chaire de Statistique Appliquée & CEDRIC
CNAM
292 rue Saint Martin, F-75003 Paris

gilbert.saporta@cnam.fr
http://cedric.cnam.fr/~saporta

Plan

1. Introduction
2. Modèles pour comprendre: quelques problèmes
3. Choix de modèles en régression
4. A la recherche du « vrai » modèle
5. Modèles pour prédire

1. Introduction

- Un modèle statistique cherche à:
  - Fournir une certaine compréhension des données et du mécanisme qui les a engendrées à travers une représentation parsimonieuse d'un phénomène aléatoire. Nécessite en général la collaboration d'un statisticien et d'un expert du domaine.
  - Prédire de nouvelles observations avec une bonne précision.

- Comprendre ou prévoir ne se réduit pas à l'opposition Non-supervisé vs Supervisé
  - « comprendre » peut signifier un modèle de distribution pour un vecteur aléatoire mais aussi un modèle de régression du type $y = f(x;\theta)+\varepsilon$
  - Dans une vision classique, un modèle doit être simple, et ses paramètres interprétables en termes du domaine d'application : elasticité, odds-ratio, etc.

- Paradoxe n° 1
  - Un « bon » modèle statistique ne donne pas nécessairement des prédictions précises au niveau individuel. Exemple facteurs de risque en épidémiologie
- Paradoxe n°2
  - On peut prévoir sans comprendre:
    - pas besoin d'une théorie du consommateur pour faire du ciblage
    - un modèle n'est qu'un algorithme

- En data mining, un bon modèle est celui qui donne de bonnes prévisions
  - capacité prédictive sur de nouvelles observations «généralisation »
  - différent de l'ajustement aux données (prédire le passé)
    - Un modèle trop précis sur les données se comporte de manière instable sur de nouvelles données : phénomène de surapprentissage
    - Un modèle trop robuste (rigide) ne donnera pas un bon ajustement sur les données
  - modéles issus des données

2. Modèles pour comprendre quelques problèmes

- Le statisticien et le savant...
  - Une vue naïve: le savant (economiste, biologiste, etc.) formule un modèle, le statisticien l'aide à estimer les paramétres et (ou) éventuellement réfute le modèle selon un test d'ajustement. Si le modèle est rejeté, le savant en cherche un autre.

- L’estimation n’est pas si facile:
  - Il faut une technique générale
    - Le maximum de vraisemblance ne l’a emporté que récemment sur la méthode des moments et le minimum du chi-carré (cf. Berkson, 1980 pour la régression logistique)
    - Les moindres carrés restent utilisés : souvent plus robustes, et nécessitant moins d’hypothèses (PLS)

- Avec peu de données:
  - Utiliser des estimateurs contraints: eg régularisation ridge
  - Ou devenir bayesien
    - Utiliser la régression ridge fait devenir inconsciemment bayesien
  - Dans les deux cas difficulté d'application des tests d'ajustement:
    - Quel degré de liberté?
    - Impossibilité de rejeter des modèles surparamétrés

3. Choix de modèles

- Quand l' "expert" hésite entre plusieurs formulations
  - Dans une famille paramétrée
  - Utilisation la plus fréquente: sélection de variables
- Parsimonie
  - Le rasoir d'Ockham* : un principe scientifique pour éviter les hypothèses inutiles

* Ou Occam

la recherche de la parsimonie: le rasoir d'Ockham

Guillaume d'Occam (1285? – 1349?), dit le « docteur invincible » franciscain philosophe logicien et théologien scolastique.
Etudes à Oxford, puis Paris. Enseigne quelques années à Oxford.
Accusé d'hérésie, convoqué pour s'expliquer à Avignon, excommunié pour avoir fui à Munich à la cour de Louis IV de Bavière. Meurt vraisemblablement de l'épidémie de peste noire.

Principe de raisonnement attribué à Occam : « Les multiples ne doivent pas être utilisés sans nécessité » (*pluralitas non est ponenda sine necessitate*).

A inspiré le personnage du moine franciscain Guillaume de Baskerville dans le « Nom de la rose » d'Umberto Eco. *Premier jour, vêpres : « il ne faut pas multiplier les explications et les causes sans qu'on en ait une stricte nécessité. »*

Le rasoir d'Ockham ou principe de parcimonie

Principe de raisonnement attribué à Ockham : « Les multiples ne doivent pas être utilisés sans nécessité » (*pluralitas non est ponenda sine necessitate*).

Rasoir d'Ockham et science moderne

Le rasoir d'Ockham n'est malheureusement pas un outil très incisif, car il ne donne pas de principe opératoire clair pour distinguer entre les hypothèses en fonction de leur complexité : ce n'est que dans le cas où deux hypothèses ont la même vraisemblance qu'on favorisera l'hypothèse la plus simple (ou parcimonieuse). Il s'agit en fait d'une application directe du théorème de Bayes où l'hypothèse la plus simple a reçu la probabilité a priori la plus forte. Des avatars modernes du rasoir sont les mesures d'information du type AIC, BIC où des mesures de pénalité de la complexité sont introduites dans la log-vraisemblance.

wikipedia

Le principe de vraisemblance (Fisher, 1920)

- échantillon de n observations iid :

$$L\left(x_1,..,x_n;\theta\right)=\prod_{i=1}^{n} f\left(x_i;\theta\right)$$

- Pour une famille f, la meilleure estimation de θ est celle qui maximise la vraisemblance, ie la probabilité d'avoir obtenu les données observées. Le meilleur modèle devrait également avoir une vraisemblance maximale.
- Mais la vraisemblance croit avec le nombre de paramétres..

Choix de modèles par vraisemblance pénalisée

- Comparer des modèles ayant des nombres de paramètres différents: *K* nombre de paramètres à estimer.

Critère d'Akaike :

- $AIC = -2 \ln(L) + 2K$

Critère de Schwartz :

- $BIC = -2 \ln(L) + K \ln(n)$

- On préférera le modèle pour lequel ces critères ont la valeur la plus faible.

- AIC et BIC ne sont semblables qu’en apparence
- **Théories différentes**
  - AIC : approximation de la divergence de Kullback-Leibler entre la vraie distribution f et le meilleur choix dans une famille paramétrée

$$I(f;g) = \int f(t) \ln \frac{f(t)}{g(t)} dt = E_f(\ln(f(t)) - E_f(\ln(g(t))$$

Asymptotiquement:

$$E_{\hat{\theta}} E_f(\ln(g(t;\hat{\theta})) \sim \ln(L(\hat{\theta})) - k$$

- BIC : choix bayesien de modèles
  - m modèles M_i paramétrés par θ_i de probabilités *a priori* $P(M_i)$
  - Distribution *a priori* de θ_i pour chaque modèle $P(\theta_i / M_i)$.
  - Distribution *a posteriori* du modèle sachant les données ou vraisemblance intégrée $P(Mi/_x)$

$$P(M_i / \mathbf{x}) \propto P(M_i)P(\mathbf{x} / M_i) = P(M_i)\int P(\mathbf{x} / \theta_i, M_i)P(\theta_i / M_i)d\theta_i$$

Posterior odds entre deux modèles: $$\frac{P(M_i / \mathbf{x})}{P(M_j / \mathbf{x})} = \frac{P(M_i)}{P(M_j)}\frac{P(\mathbf{x} / M_i)}{P(\mathbf{x} / M_j)}$$

facteur de Bayes

Avec un prior constant sur l'ensemble des m modèles , le modèle le plus probable *a posteriori* est celui qui maximise

$$P(\mathbf{x} / M_i)$$

après quelques développements on trouve:

$$\ln(P(\mathbf{x} / M_i) \sim \ln(P(\mathbf{x} / \hat{\theta}_i, M_i)) - \frac{k}{2}\ln(n)$$

Le modèle le plus probable minimise :

$$-2\ln(P(\mathbf{x} / M_i) \sim -2\ln(P(\mathbf{x} / \hat{\theta}_i, M_i)) - k\ln(n) = BIC$$

$$P(M_i / \mathbf{x}) = \frac{e^{-0.5BIC_i}}{\sum_{j=1}^{m} e^{-0.5BIC_j}}$$

- Nombre effectif de paramétres
  - Comment calculer un degré de liberté pour un modèle linéaire avec contraintes (eg ridge)?
  - Si $\hat{\mathbf{y}} = \mathbf{S}\mathbf{y}$ où S est une matrice nxn ne dépendant que des prédicteurs et pas de y (généralisant $X(X'X)^{-1}X'$)

ddl= Trace(S)

3. Sélection de variables en régression linéaire classique

- Choix de k variables parmi p
  - Elimination de variables non pertinentes
  - Obtention de formules plus stables
- Critères:
  - A k fixé: équivalence de tous les critères, le même modèle est choisi

3.1 Quelques critères pour comparer des modèles de tailles différentes

- R^2 augmente avec le nombre de variables mais pas R^2 ajusté
- R^2 ajusté et $\hat{\sigma}$ sont équivalents
- AIC ou BIC

$$AIC = -2\ln(L) + 2(k+1) = n\ln(\frac{SSE}{n}) + 2(k+1) + n(\ln\pi + 1)$$

$$BIC = -2\ln(L) + \ln(n)(k+1) = n\ln(\frac{SSE}{n}) + \ln(n)(k+1) + n(\ln\pi + 1)$$

- *SSE*= somme des carrés des résidus du modèle à k variables

Le C_p de Mallows

- On cherche à estimer l'erreur quadratique de prédiction (MSPE) :

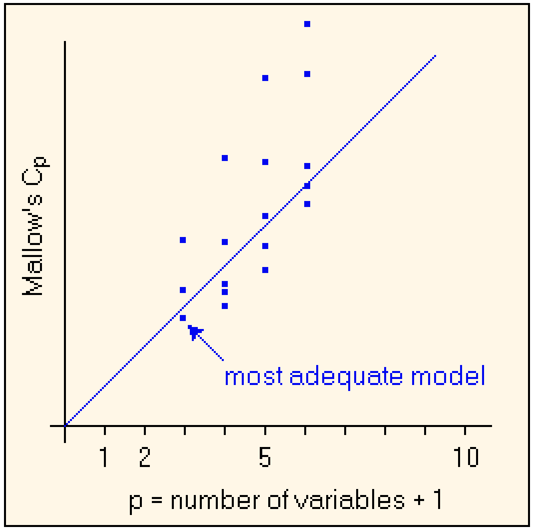
$$E \sum_j (\hat{Y}_j - E(Y_j|X_j))^2/\sigma^2,$$

- En sélectionnant P prédicteurs parmi K :

$$C_p = \frac{SSE_p}{S^2} - N + 2P,$$

- Si le modèle est le bon $E(C_p)=p$

```
The general procedure to find an
adequate model  by means of the Cp
statistic is to calculate Cp
 for all possible combinations of
variables and the Cp values against p.
The model with the lowest Cp value
approximately equal to p is the most
"adequate" model.
```



http://www.statistics4u.com/fundstat_eng/cc_varsel_mallowscp.html

```
proc reg;
title Regression OLS;
id nom;
model prix=cyl puis lon lar poids vitesse/AIC BIC ADJRSQ CP selection=ADJRSQUARE ;
run;
```

Number in Model	Adjusted R-Square	R-Square	C(p)	AIC	BIC	Variables in Model
2	0.6448	0.6866	-0.1501	$\boxed{300.5430}$	305.1183	PUIS POIDS
3	0.6342	0.6988	1.3900	301.8305	307.7996	CYL PUIS POIDS
3	0.6301	0.6954	1.5183	302.0321	307.9006	PUIS LAR VITESSE
3	0.6262	0.6922	1.6398	302.2209	307.9951	PUIS POIDS VITESSE
2	0.6235	0.6678	0.5608	301.5917	305.8148	PUIS VITESSE
2	0.6229	0.6673	0.5808	301.6203	305.8337	PUIS LON
3	0.6224	0.6890	1.7586	302.4037	308.0866	PUIS LON VITESSE
3	0.6206	0.6876	1.8137	302.4879	308.1287	PUIS LAR POIDS
3	0.6195	0.6866	1.8492	302.5419	308.1557	PUIS LON POIDS
2	0.6162	0.6613	0.8057	301.9393	306.0440	PUIS LAR
1	0.6153	0.6379	-0.3084	301.1433	$\boxed{304.2040}$	PUIS
4	0.6101	0.7018	3.2760	303.6495	310.9014	CYL PUIS LAR POIDS
4	0.6094	0.7013	3.2940	303.6782	310.9107	CYL PUIS LAR VITESSE
4	0.6091	0.7010	3.3048	303.6953	310.9162	CYL PUIS POIDS VITESSE
4	0.6064	0.6990	3.3807	303.8157	310.9554	CYL PUIS LON POIDS
4	0.6056	0.6984	3.4049	303.8540	310.9679	PUIS LAR POIDS VITESSE
3	0.6025	0.6726	2.3791	303.3294	308.5496	CYL PUIS LON
4	0.6021	0.6958	3.5049	304.0111	311.0192	PUIS LON LAR VITESSE
4	0.5991	0.6934	3.5925	304.1477	311.0640	CYL PUIS LON VITESSE
4	0.5991	0.6934	3.5926	304.1478	311.0640	PUIS LON POIDS VITESSE

3	0.5971	0.6682	2.5478	303.5731	308.6716	PUIS LON LAR
3	0.5966	0.6678	2.5608	303.5917	308.6809	CYL PUIS VITESSE
3	0.5941	0.6657	2.6416	303.7071	308.7387	CYL PUIS LAR
4	0.5920	0.6880	3.7970	304.4624	311.1676	PUIS LON LAR POIDS
2	0.5897	0.6379	1.6911	303.1426	306.8329	CYL PUIS
5	0.5874	0.7087	5.0134	305.2253	314.0329	CYL PUIS LAR POIDS VITESSE
2	0.5821	0.6313	1.9422	303.4699	307.0467	POIDS VITESSE
5	0.5785	0.7025	5.2499	305.6077	314.0878	CYL PUIS LON LAR POIDS
5	0.5783	0.7023	5.2565	305.6183	314.0894	CYL PUIS LON POIDS VITESSE
5	0.5781	0.7022	5.2608	305.6252	314.0904	CYL PUIS LON LAR VITESSE
4	0.5742	0.6744	4.3117	305.2312	311.4245	CYL PUIS LON LAR
5	0.5730	0.6986	5.3988	305.8444	314.1230	PUIS LON LAR POIDS VITESSE
3	0.5590	0.6368	3.7336	305.1985	309.4907	CYL POIDS VITESSE
3	0.5562	0.6345	3.8212	305.3129	309.5488	LON POIDS VITESSE
3	0.5548	0.6334	3.8643	305.3690	309.5774	LAR POIDS VITESSE
6	0.5504	0.7091	7.0000	307.2033	317.3025	CYL PUIS LON LAR POIDS VITESSE
1	0.5404	0.5675	2.3563	304.3443	306.7261	POIDS
4	0.5295	0.6402	5.6045	307.0284	312.0506	CYL LON POIDS VITESSE
4	0.5274	0.6386	5.6663	307.1100	312.0800	CYL LAR POIDS VITESSE
4	0.5222	0.6346	5.8162	307.3063	312.1512	LON LAR POIDS VITESSE
2	0.5157	0.5726	4.1601	306.1270	308.7842	CYL POIDS
2	0.5103	0.5679	4.3405	306.3268	308.9156	LON POIDS
2	0.5099	0.5675	4.3529	306.3406	308.9246	LAR POIDS
5	0.4904	0.6403	7.6033	309.0268	314.6964	CYL LON LAR POIDS VITESSE
3	0.4818	0.5732	6.1375	308.1018	310.9985	CYL LON POIDS
3	0.4811	0.5726	6.1600	308.1270	311.0119	CYL LAR POIDS
3	0.4764	0.5688	6.3055	308.2883	311.0981	LON LAR POIDS
2	0.4475	0.5125	6.4358	308.4985	310.3556	LON VITESSE
4	0.4424	0.5736	8.1231	310.0858	313.2228	CYL LON LAR POIDS
3	0.4238	0.5255	7.9422	310.0101	312.0358	CYL LON VITESSE

2	0.4144	0.4833	7.5395	309.5451	311.0600	CYL LON
3	0.4107	0.5147	8.3522	310.4167	312.2624	LON LAR VITESSE
4	0.3826	0.5279	9.8526	311.9199	314.0023	CYL LON LAR VITESSE
2	0.3790	0.4521	8.7201	310.6011	311.7790	CYL VITESSE
1	0.3778	0.4144	8.1430	309.7966	310.9782	LON
3	0.3732	0.4838	9.5178	311.5251	312.8909	CYL LON LAR
1	0.3708	0.4078	8.3940	309.9995	311.1377	CYL
3	0.3685	0.4800	9.6643	311.6597	312.9683	CYL LAR VITESSE
2	0.3680	0.4424	9.0855	310.9158	311.9950	CYL LAR
2	0.3579	0.4334	9.4254	311.2036	312.1932	LAR VITESSE
2	0.3363	0.4144	10.1430	311.7966	312.6040	LON LAR
1	0.2971	0.3384	11.0161	311.9926	312.7134	VITESSE
1	0.2550	0.2988	12.5142	313.0395	313.5488	LAR

3.2 algorithmes de sélection

- Dénombrement exhaustif: 2^p-1 modèles
- Meilleurs sous-ensembles (algorithme de Furnival et Wilson jusqu'à quelques dizaines de variables)
- Méthodes pas à pas (stepwise selection)
  - Ascendant (forward)
  - Descendant (backward)
  - Ascendant avec élimination possible (stepwise)
  - ...

Les logiciels classiques utilisent des tests d'arret :
F pour entrer, pour rester

Summary of Forward Selection

Step	Variable Entered	Number Vars In	Partial R-Square	Model R-Square	C(p)	Valeur F	Pr > F
1	PUIS	1	0.6379	0.6379	-0.3084	28.19	<.0001
2	POIDS	2	0.0487	0.6866	-0.1501	2.33	0.1476

Summary of Backward Elimination

Step	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	Valeur F	Pr > F
1	LON	5	0.0004	0.7087	5.0134	0.01	0.9098
2	VITESSE	4	0.0069	0.7018	3.2760	0.29	0.6025
3	LAR	3	0.0030	0.6988	1.3900	0.13	0.7228
4	CYL	2	0.0122	0.6866	-0.1501	0.57	0.4646

- Les méthodes basées sur des tests F devraient être abandonnées:
  - utilisation incorrecte de tests multiples
  - erreurs standard ne tenant pas compte du processus de sélection

4. A la recherche du vrai modèle

- On suppose que le vrai modèle fait partie des m modèles en compétition

Comparaison AIC BIC

- Si n tend vers l'infini la probabilité que le *BIC* choisisse le vrai modèle tend vers 1, ce qui est faux pour l'*AIC*.
- *AIC* va choisir le modèle qui maximisera la vraisemblance de futures données et réalisera le meilleur compromis biais-variance
- L'*AIC* est un critère prédictif tandis que le *BIC* est un critère explicatif.
- Pour n fini: résultats contradictoires. *BIC* ne choisit pas toujours le vrai modèle: il a tendance à choisir des modèles trop simples en raison de sa plus forte pénalisation
- **Illogisme à utiliser les deux simultanément**

AIC BIC réalistes?

- Vraisemblance pas toujours calculable.
- Nombre de paramétres non plus (arbres, ..)

- « Vrai » modèle?

"Essentially, all models are wrong, but some are useful " G.Box (1987)

- Aucun modèle parcimonieux ne s'ajuste à de grands ensembles de données et les tests sont alors inutiles
  - Avec un millions d'observations, une corrélation de 0.01 est significativement différente de zéro. A-t-elle un intérêt?
  - La plupart des modèles classiques sont rejetés car trop simples et le moindre écart devient significatif

"The Truth Is Out There" (X-Files, 1993)

5. Modèles pour prédire

- En data mining, un bon modèle est celui qui donne de bonnes prévisions
  - capacité prédictive sur de nouvelles observations («généralisation »)
  - différent de l’ajustement aux données (prédire le passé)
    - Un modèle trop précis sur les données se comporte de manière instable sur de nouvelles données : phénomène de surapprentissage
    - Un modèle trop robuste (rigide) ne donnera pas un bon ajustement sur les données
  - modéles issus des données

5.1 Ajuster ou prédire?

- Les critères de type AIC, BIC, Cp utilisent deux fois les données: une fois pour estimer, une autre pour mesurer la qualité
- Prédire les données futures et non le passé!
- Minimiser l’espérance de l’erreur quadratique de prédiction $E(y-\hat{y})^2$

- Solution pratique: la validation croisée
  - « Leave one out »: chaque observation est estimée à l'aide des n-1 autres
    - résidu prédit:

$$y_i - \hat{f}^{(-i)}(x_i) = y_i - \hat{y}_i^{(-i)} = \frac{y_i - \hat{y}_i}{1 - h_i}$$

h_i terme diagonal du projecteur $\mathbf{X(X'X)^{-1}X'}$

    - PRESS predicted error sum of squares

$$\sum_{i=1}^{n} \left(y_i - \hat{y}_i^{(-i)} \right)^2$$

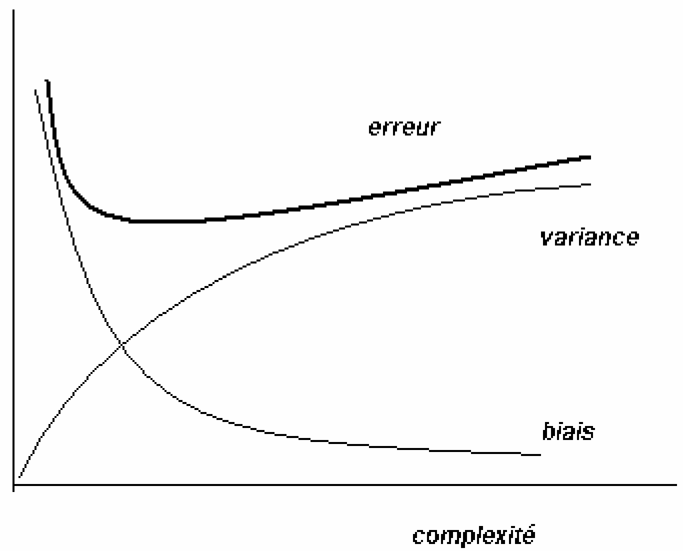
quelques press

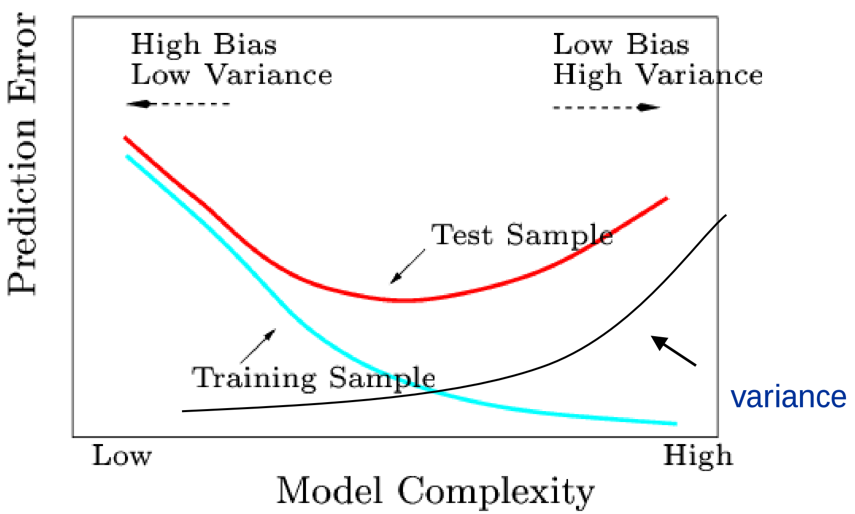
modéle complet:	732 726 946
puissance poids	308 496 438
cylindree puissance poids	369 112 558
puissance	327 142 373

5.2 Le dilemme biais-variance

$$E\left(y_0-\hat{y}_0\right)^2=\sigma^2+E\left(f(x_0)-\hat{f}(x_0)\right)^2=$$

$$\sigma^2+\underbrace{\left(E\left(\hat{f}(x_0)\right)-f(x_0)\right)^2}_{\text{biais}}+\underbrace{V\left(\hat{f}(x_0)\right)}_{\text{variance}}$$





Adapted from Hastie et al.

complexité d’un modèle

- Plus un modèle est complexe, mieux il s’ajuste en apprentissage mais avec de grands risques en test.
- ∃ compromis optimal
- Comment mesurer la complexité d’un modèle?
  - V.Vapnik a montré que ce n’est pas le nombre de paramétres “Statistical Learning Theory”

5.3 Combinaison de modèles

- Bayesian Model Averaging

$$P(y/\mathbf{x}) = \sum_{i=1}^{m} P(y/M_i,\mathbf{x})P(M_i \; /\mathbf{x})$$

$$E(y/\mathbf{x}) = \sum_{i=1}^{m} E(y/M_i,\mathbf{x})P(M_i \; /\mathbf{x})$$

Moyenne des prévisions de chaque modèle, pondérées par les probabilités a posteriori

- Stacking
  - Combinaison non bayésienne de m prédictions obtenues par des modèles différents $\hat{f}_1(\mathbf{x}), \hat{f}_2(\mathbf{x}), \ldots, \hat{f}_m(\mathbf{x})$
  - Premiere idée : régression linéaire

$$\min \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{m} w_j \hat{f}_j(\mathbf{x}) \right)^2$$

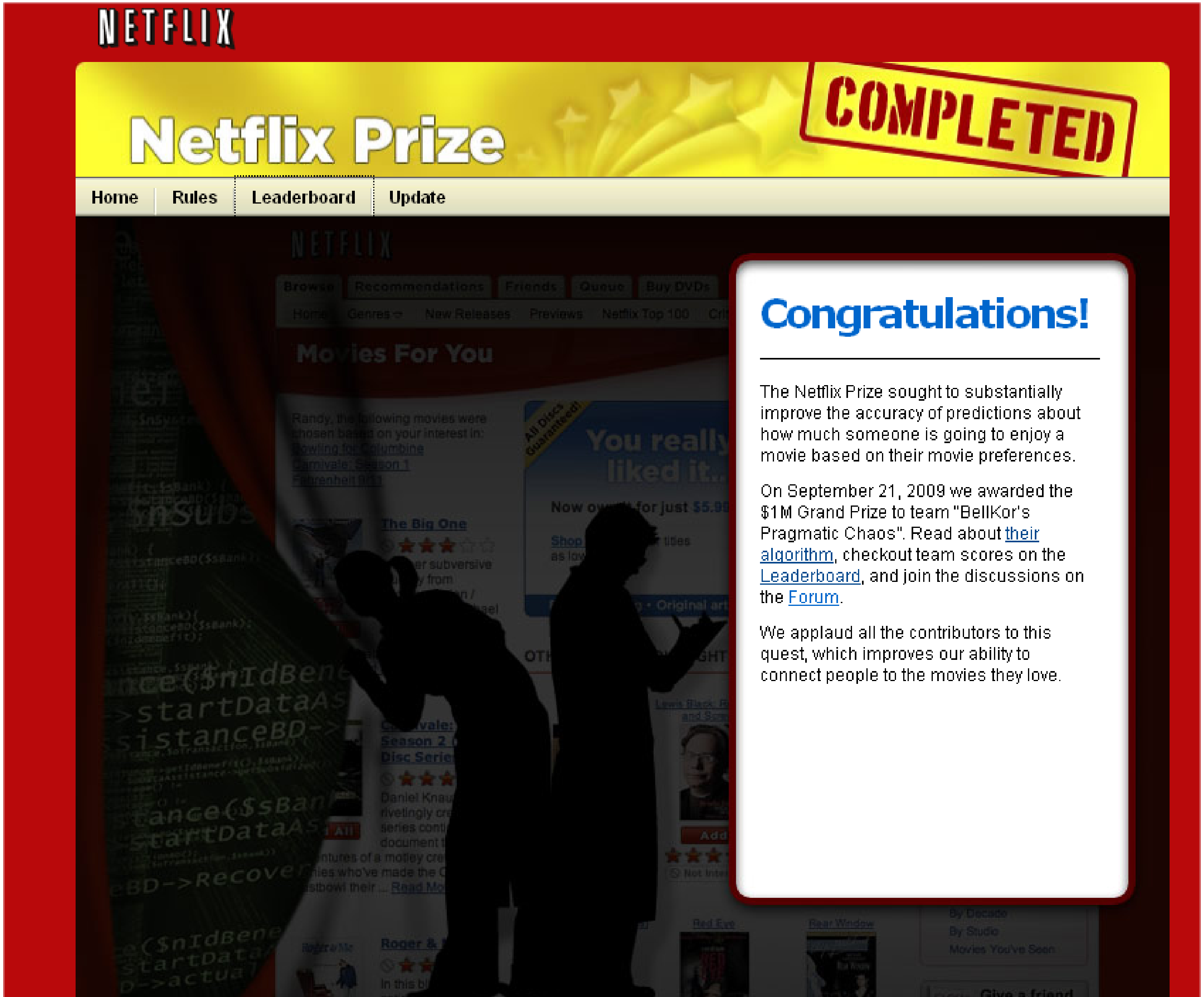
    - Favorise les modèles les plus complexes: surapprentissage

- Solution: utiliser les valeurs prédites en otant à chaque fois l'unité i

$$\min \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{m} w_j \hat{f}_j^{-i}(\mathbf{x}) \right)^2$$

- Améliorations:
  - Combinaisons linéaires à coefficients positifs (et de somme 1)
  - Régression PLS ou autre méthode régularisée car les m prévisions sont très corrélées

- Avantages
  - Prévision meilleures qu'avec le meilleur modèle
  - Possibilité de mélanger des modéles de toutes natures: arbres , ppv, réseaux de neurones etc. alors que le BMA utilise des modèles paramétrés de la même famille


NETFLIX
Netflix Prize
COMPLETED
Home
Rules
Leaderboard
Update
Congratulations!
The Netflix Prize sought to substantially improve the accuracy of predictions about how much someone is going to enjoy a movie based on their movie preferences.
On September 21, 2009 we awarded the $1M Grand Prize to team "BellKor's Pragmatic Chaos". Read about their algorithm, checkout team scores on the Leaderboard, and join the discussions on the Forum.
We applaud all the contributors to this quest, which improves our ability to connect people to the movies they love.

- BellKor's Pragmatic Chaos team bested Netflix's own algorithm for predicting ratings by 10.06%.
- “The Netflix dataset contains more than 100 million datestamped movie ratings performed by anonymous Netflix customers between Dec 31, 1999 and Dec 31, 2005 This dataset gives ratings about m = 480,189 users and n = 17,770 movies The contest was designed in a training-test set format. A Hold-out set of about 4.2 million ratings was created consisting of the last nine movies rated by each user (or fewer if a user had not rated at least 18 movies over the entire period).The remaining data made up the training set.”
- Blend of 24 predictions