Strijov V.V. Credit Scorecard Development: Model Generation and Multimodel Selection
// 26th European Conference on Operational Research, 2013 : 220.

Credit Scorecard Development: Model Generation and Multimodel Selection

Vadim Strijov,
Alexander Aduenko and Kirill Pavlov

Russian Academy of Sciences
Computing Center

EURO|INFORMS MMXIII
Rome 1–4 July, 2013

The workflow

Client's application & history

$\downarrow$

Client's score: probability of fraud / default

$\downarrow$

Accept (refuse) the application

$\downarrow$

Make the agreement

$\downarrow$

Client's history

Types of scorecards

- Application
- Behavioral
- Collection

Number of the records:

- $\sim 10^4$ for long-term credits,
- $\sim 10^6$ point-of-sale credits,
- $\sim 10^7$ for churn analysis.

Type of detection

$$\overbrace{0 \longrightarrow \quad 30+ \longrightarrow 60+ \longrightarrow 90+}^{\text{Fraud: deliquency 90+ on 3}^{\text{rd}}} \longrightarrow 120+ \quad \longrightarrow 150+$$

$$0 \longrightarrow \quad \underbrace{30+ \longrightarrow 60+ \longrightarrow 90+ \longrightarrow 120+}_{\text{Default: deliquency 90+ on any, but 1}^{\text{st}}} \quad \longrightarrow 150+$$

List of variables

Variable	Type	Categories
Loan currency	Nominal	3
Applied amount	Linear	
Monthly payment	Linear	
Tetm of contract	Linear	
Region of the office	Nominal	7
Day of week of scoring	Linear	
Hour of scoring	Linear	
Age	Linear	
Gender	Nominal	2
Marital status	Nominal	4
Education	Ordinal	5
Number of children	Linear	
Industrial sector	Nominal	27
Salary	Linear	
Place of birth	Nominal	94
...	...	...
Car number shown	Nominal	2

The data, general statistics

- Loans of 90+ delinquency, default cases, applications
- The fraud cases are rejected
- Overall number of cases $\sim 10^4$–10^6
- Default rate $\sim$ 8–16%
- Period of observing: no less 91 days after approval
- Number of source variables $\sim$ 30–50
- Number records with missing data > 0, usually very small
- Number of cases with outliers > 0, $3\sigma^2$-cutoff

Scorecard developing, regular way

- Create the data set (the design matrix and the target vector)
- Map ordinal and nominal-scaled features to the binary ones
- Make the regression model
- Test it (multi-collinearity, stability, pooling, etc., see Basel-II)
- Determine the cut-off, according to the bank policy

There given

- the set $D = \{(\mathbf{x}_i, y_i)\}$,
  $\mathbf{x} = [x_{i1}, \dots, x_{ij}, \dots, x_{in}] \in \mathbb{R}^n, \quad y_i \in \{0, 1\}$;
  $i \in \mathcal{I} = \{1, \dots, m\}, \qquad j \in \mathcal{J} = \{1, \dots, n\}$;
- learning/control $i \in \mathcal{I} = \mathcal{L} \sqcup \mathcal{C}$;
- the error function S and the model $f(\mathbf{w}, \mathbf{x}) = \mu(\mathbf{w}^\mathsf{T}\mathbf{x})$, where μ is the link function.

Find

the subset $\mathcal{A} \subseteq \mathcal{J}$, which brings

$$\mathcal{A}^* = \arg\min_{\mathcal{A} \subseteq \mathcal{J}} S(f_{\mathcal{A}} | \mathbf{w}^*, D_{\mathcal{C}}) \tag{1}$$

while parameters $\mathbf{w}^*$ bring

$$\mathbf{w}^* = \arg\min_{\mathbf{w} \in \mathcal{W}} S(\mathbf{w} | D_{\mathcal{L}}, f_{\mathcal{A}}). \tag{2}$$

The dependent variable $\mathbf{y} \sim \text{Bernoulli}(\mathbf{f})$

$$\mathbf{y} = [y_1, \ldots, y_m]^\mathsf{T}$$

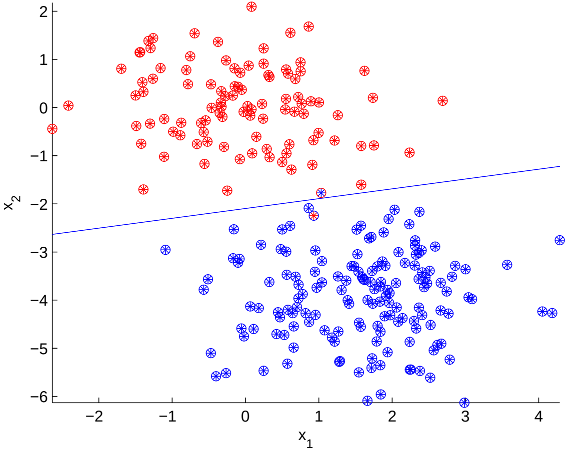
and the model

$$\mathbf{f} = \frac{1}{1 + \exp(-X\mathbf{w})}$$

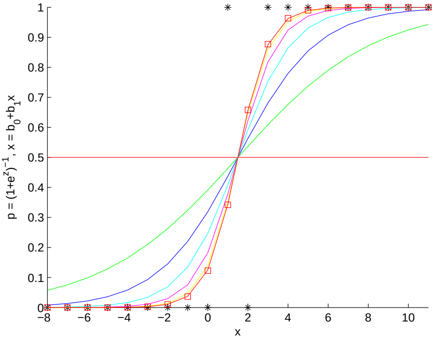
define the (error function) log likelihood function

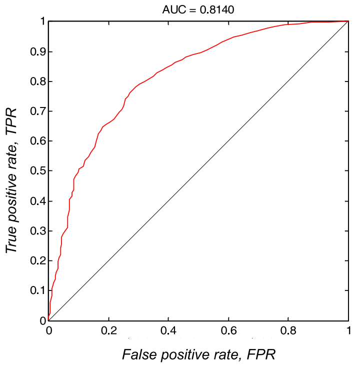
$$-\ln P(D|\mathbf{w}) = -\sum_{i \in \mathcal{L}} (y_i \ln \mathbf{w}^\mathsf{T}\mathbf{x}_i + (1 - y_i) \ln(1 - \mathbf{w}^\mathsf{T}\mathbf{x}_i)) = S(\mathbf{w}).$$

Create the one-level model



Use the ROC-curve as the quality criterion





	P	N
$P*$	TP	FP
$N*$	FN	TN

$$TPR = TP/P = TP/(TP + FN)$$
$$FPR = FP/N = FP/(FP + TN)$$

List of primitive functions

Description	In	N in	Out	N out	Comm	Param
Nominal to binary	nom	1	bin	1–4	-	Yes
Ordinal to binary	ord	1	bin	1–4	-	Yes
Linear to linear segments	lin	1	lin	1–4	-	Yes
Linear segments to binary	lin	1	bin	1–4	-	Yes
Get one column of n-matrix	bin	1–4	bin	1	-	Yes
Conjunction	bin	2–6	bin	1	Yes	-
Dijsunction	bin	2–6	bin	1	Yes	-
Negate binary	bin	1	bin	1	-	-
Logarithm	lin	1	lin	1	-	-
Hyperbolic tangent sigmiod	lin	1	lin	1	-	-
Logistic sigmoid	lin	1	lin	1	-	-
Sum	lin	2–3	lin	1	Yes	-
Divfference	lin	2	lin	1	No	-
Multiplication	lin,bin	2–3	lin	1	Yes	-
Division	lin	2	lin	1	No	-
Inverse	lin	1	lin	1	-	-
Polynomial transformation	lin	1	lin	1	-	Yes
Radial basis function	lin	1	lin	1	-	Yes
Monomials: $x\sqrt{x}$, etc.	lin	1	lin	1	-	-

Feature generation

There given

- the measured features $\Xi = \{\xi\}$,
- the expert-given primitive functions $G = \{g(\mathbf{b}, \xi)\}$,

$$g : \xi \mapsto x;$$

- the generation rules: $\mathcal{G} \supset G$, where the superposition $g_k \circ g_l \in \mathcal{G}$ w.r.t. numbers and types of the input and output arguments;
- the simplification rules: g_u is not in $\mathcal{G}$, if there exist a rule

$$r : g_u \mapsto g_v \in \mathcal{G}.$$

The result is

the set of the features $X = \{\mathbf{x}_1, \dots, \mathbf{x}_j, \dots, \mathbf{x}_n\}$.

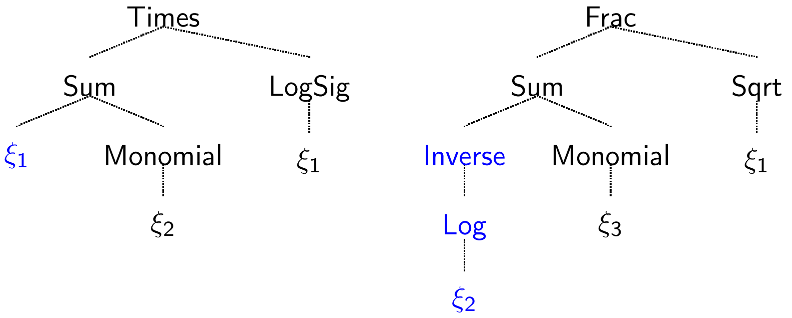
The number of features exceeds the number of clients!

Examples of generated features

- **Frac**(Period of residence, Undeclared income)
- **Frac**(**Seg**(Period of employment), Term of contract)
- **And**(Income confirmation, Bank account)
- **Times**(**Seg**(Score hour), **Frac**(**Seg**(Period of employment), Salary))

1. Select random nodes in two features,
2. exchange the corresponded subtrees,
3. modify the function at a random node for another one from the primitive set.

Any modification must result an admissible superposition.



Exhaustive search in the set of the generalized linear models

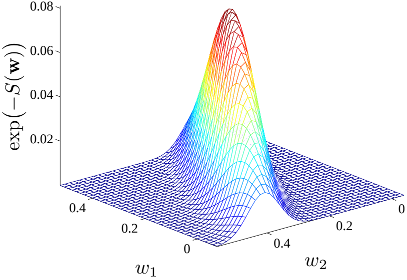
$$\mu(y) = w_0 + \alpha_1 w_1 x_1 + \alpha_2 w_2 x_2 + \ldots + \alpha_R w_R x_R.$$

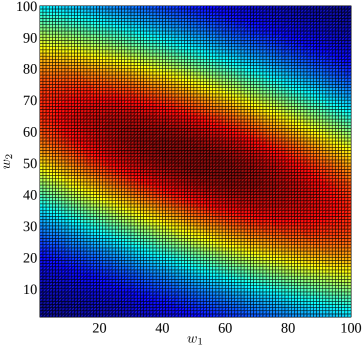
Here $\alpha \in \{0, 1\}$ is the structural parameter.

Find a model defined by the set $\mathcal{A} \subseteq \mathcal{J}$:

α_1	α_2	$\ldots$	$\alpha_{\lvert\mathcal{J}\rvert}$
1	0	...	0
0	1	...	0
...	...	...	...
1	1	...	1

Let there given a sampled set $\{\mathbf{w}_1, \dots, \mathbf{w}_K\}$ realizations of the random variable $\mathbf{w}$ and the error function $S(\mathbf{w}|D, \mathbf{f})$. Consider the set $\{s_k = \exp\big(-S(\mathbf{w}_k|D, \mathbf{f})\big)|k = 1, \dots, K\}$.





Let $\mathbf{w} \sim \mathcal{N}(\mathbf{w}_0, A^{-1})$:

$$p(\mathbf{w}|A, f) = (2\pi)^{-\frac{n}{2}} \det{}^{-\frac{1}{2}}(A^{-1}) \exp\left(\frac{1}{2}(\mathbf{w} - \mathbf{w}_0)^T A (\mathbf{w} - \mathbf{w}_0)\right).$$

The posterior distribution of the model parameters, given A, B:

$$p(\mathbf{w}|D, A, B, f) = \frac{p(D|\mathbf{w}, B, f)p(\mathbf{w}|A, f)}{p(D|A, B, f)}.$$

Rewrite the error function $S = E_{\mathbf{w}} + E_D$ as...

The error function now includes the second part

The distribution $\mathbf{y} \sim \mathcal{N}(\mathbf{f}, A^{-1})$, LM

$$S(\mathbf{w}|D, f) = \frac{1}{2}(\mathbf{w} - \mathbf{w}_{\mathrm{MP}})^{\mathsf{T}} A(\mathbf{w} - \mathbf{w}_{\mathrm{MP}}) + \frac{1}{2}(\mathbf{f} - \mathbf{y})^{\mathsf{T}} B(\mathbf{f} - \mathbf{y}).$$

The distribution $\mathbf{y} \sim \mathcal{B}(f, 1 - f)$, GLM

The likelihood function is $p(D|w, B, f) = \prod_{i \in \mathcal{I}} f_i^{y_i}(1 - f_i)^{1-y_i}$, and the error function

$$S(\mathbf{w}) = \frac{1}{2}(\mathbf{w} - \mathbf{w}_{\mathrm{MP}})^{\mathsf{T}} A(\mathbf{w} - \mathbf{w}_{\mathrm{MP}}) + \sum_{i \in \mathcal{I}} y_i \ln f_i + (1 - y_i) \ln (1 - f_i).$$

The covariance matrix B^{-1} is estimated using Newton-Raphson method iteratively:

$$\mathbf{w}_{k+1} = \mathbf{w}_k - (X^{\mathsf{T}} B X)^{-1} X^{\mathsf{T}} (\mathbf{f} - \mathbf{y}) = (X^{\mathsf{T}} B X)^{-1} X^{\mathsf{T}} B \left(X \mathbf{w}_k - B^{-1}(\mathbf{f} - \mathbf{y})\right).$$

There are nine possible variants for data generation hypothesis

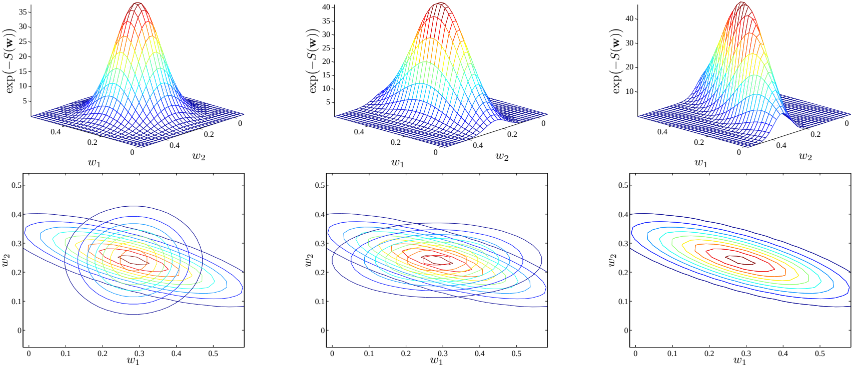
The (inverse) covariance matrix of

parameters	target variable
$A = \alpha I_n$	$B = \beta I_m$
$A = \text{diag}(\alpha_1, \dots, \alpha_n)$	$B = \text{diag}(\beta_1, \dots, \beta_m)$
A	B

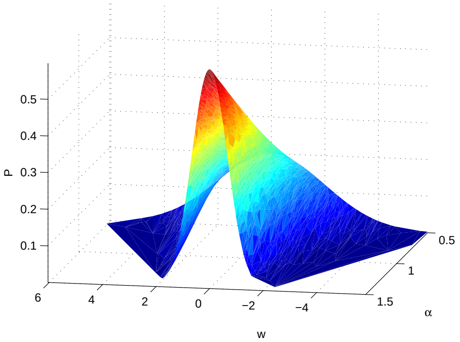
Empirical distribution: approximation and hypothesis

Approximate the set $\{s_k\}$ with the function $p(w|A)$ from $\mathcal{N}$ using the following hypothesis on the covariance matrix A^{-1}:

$A = \alpha I, \quad \alpha \geqslant 0;$ $\qquad A = \operatorname{diag}(\alpha_1, \dots, \alpha_n);$ $\qquad A, \quad \mathbf{w}^\mathsf{T} A \mathbf{w} \geqslant 0.$



How the distribution of parameters depends on $A = \alpha I_n$



- z-axis: $p(\mathbf{w}|D, f, A, B)$ the distribution of parameters,
- y-axis: α the inverted covariance,
- x-axis: w the model parameter.

Use Bayesian inference to find the most probable parameters

The most probable parameters

$$\mathbf{w}_{\text{MP}} = \arg\max_{\mathbf{w}\in\mathcal{W}} p(\mathbf{w}|D, f, A, B),$$

of the model f are estimated using the Bayesian approach

$$p(\mathbf{w}|D, f, A, B) = \frac{p(D|\mathbf{w}, f, B)p(\mathbf{w}|f, A)}{\int p(D|\mathbf{w}', f, B)p(\mathbf{w}'|f, A)d\mathbf{w}'}.$$

The likelihood function $p(D|\mathbf{w}, f, B)$ is defined by the hypothesis of distribution of the dependent variable $\mathbf{y}$.
The model evidence

$$\mathcal{E}\left(f(\mathbf{w}, \mathbf{x})\right) = \int p(D|\mathbf{w}, f, B)p(\mathbf{w}|f, A)d\mathbf{w}.$$

The problem of the most evident model selection

There given:

- the sample set D,
- the finite set of models $\mathcal{F} = \{f_k | k \in \mathcal{K}\}$.

One must select the most evident model f_{k^*}, such that

$$k^* = \arg\max_{k \in \mathcal{K}} p(f_k|D) = \arg\max_{k \in \mathcal{K}} \int\limits_{\mathbf{w} \in \mathcal{W}} p(D|\mathbf{w}, B, f_k)p(\mathbf{w}|D, A, f_k)d\mathbf{w}.$$

If we assume the prior probabilities of models are equal,

$$p(f_1) = p(f_2) = \cdots = p(f_K),$$

then the most evident model selection problem is stated as the most probable model selection problem.

There given:

- the sample set D, the model $f = f(\mathbf{w}, \mathbf{x})$,
- the data generation hypothesis, it defines the error function

$$S(\mathbf{w}) = -\ln\bigl(p(D|\mathbf{w}, B, f)p(\mathbf{w}|A, f)\bigr).$$

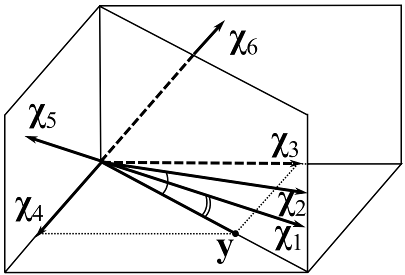
One must estimate the most probable parameters $\mathbf{w}_{\text{MP}}$

$$\mathbf{w}_{\text{MP}} = \arg\min_{\mathbf{w}\in\mathcal{W}} S(\mathbf{w}, D, \hat{A}, \hat{B}, f).$$

One must estimate corresponding hyperparameters A, B

$$\hat{A}, \hat{B} = \arg\min_{A,B} \Phi\bigl(S(\mathbf{w}_{\text{MP}}, D, A, B, f)\bigr).$$

What is the optimal feature set?



- Extract j-th column from the design matrix X,
- make regression $X_{\mathcal{J}\setminus\{j\}}$ on $\mathbf{y} \equiv X_{\{j\}}$,
- for the feature number j

$$\mathrm{VIF}_j = \frac{1}{1 - R_j^2},$$

where the determination coefficient

$$R_j^2 = 1 - \frac{\|\mathbf{x}_j - \mathbf{f}(\mathbf{x}_1, \dots, \mathbf{x}_{j-1}, \mathbf{x}_{j+1}, \dots, \mathbf{x}_n)\|^2}{\|\mathbf{x}_j - \tilde{\mathbf{x}}_j|^2};$$

here $\tilde{\mathbf{x}}_j$ is average vector for $\mathbf{x}_j$.

Decompose

$$X^{\mathsf{T}}XV = V\Lambda^2.$$

Find the conditional indexes

$$\eta_j = \frac{\lambda_{\max}}{\lambda_j}.$$

Obtain the variances of the parameters $\mathbf{w}$

$$Var(\mathbf{w}) = \sigma^2(X^T X)^{-1} = \sigma^2(V^T)^{-1}\Lambda^{-2}V^{-1} = \sigma^2 V\Lambda^{-2}V^T,$$

where σ^2 is the variance of the residuals.
The variance of w_j is j-th diagonal element of $Var(\mathbf{w})$.
Match the conditional index η_j and corresponding coefficients q_{ij}

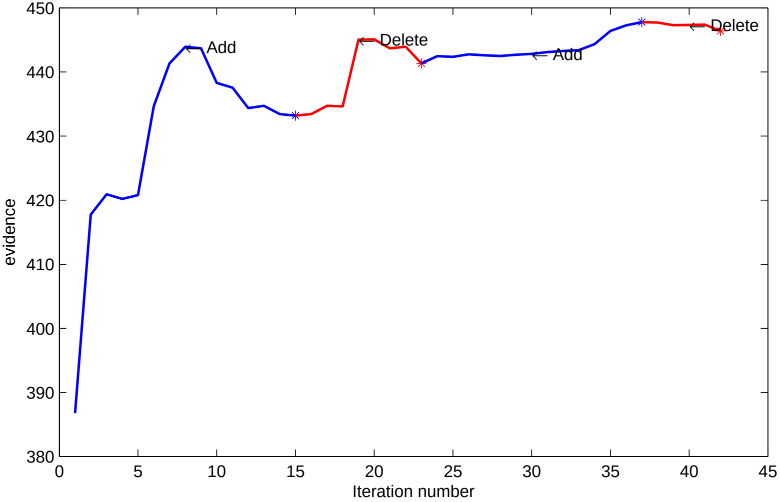
$$\sigma^{-2}\mathbf{var}(w_i) = \sum_{j=1}^{n} \frac{\upsilon_{ij}^2}{\lambda_j^2} = (q_{i1} + q_{i2} + \ldots + q_{in}).$$

Belsley method, the decomposition of $\mathbf{var}(w_i)$

Conditional index	$\mathbf{var}(w_1)$	$\mathbf{var}(w_2)$	$\ldots$	$\mathbf{var}(w_n)$
η_1	q_{11}	q_{21}	$\cdots$	q_{n1}
η_2	q_{12}	q_{22}	$\cdots$	q_{n2}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
η_n	q_{1n}	q_{2n}	$\cdots$	q_{nn}

- the bigger q_{ij} the bigger impact of j-th parameter into the variance of i-th parameter;
- the bigger values of η_j mean there is a dependency between the features;
- the i-th feature in involved in the multicorrelation if η_j is larger and q_{ij} exceeds a given threshold.

Model selection by the evidence maximization



Add and **Delete** features until the evidence goes down.

Stepwise feature selection algorithm

Add stage:
Add the feature $\mathcal{E}(f_{\mathcal{A}_k})$, which brings minimum to the error function

$$j^* = \arg \min_{j \in \mathcal{J} \setminus \mathcal{A}_{k-1}} S(\mathbf{w}|D, f_{\mathcal{A}_{k-1} \cup \{j\}})$$

$\mathcal{A}_k = \mathcal{A}_{k-1} \cup \{j^*\}$ until exceeds its minimum value on this stage but no more than given $\Delta_{\mathcal{E}}$.

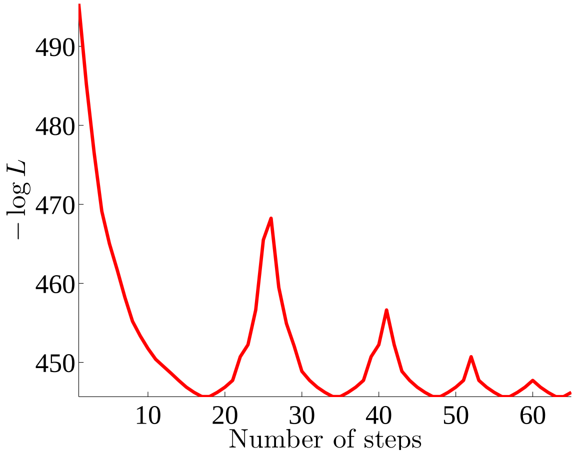
Del stage:
Delete the feature $\mathcal{A}_k = \mathcal{A}_{k-1} \setminus j^*$ according to the Belsley method:

$$i^* = \sum_{g=1}^{t} \left[\eta_g^2 > \eta_t\right], \qquad j^* = \arg \max_{j \in \mathcal{A}_{k-1}} \sum_{g=t-i^*+1}^{t} q_g^j$$

until $\mathcal{E}(f_{\mathcal{A}_k})$ exceeds its minimum value on this stage but no more than given $\Delta_{\mathcal{E}}$.

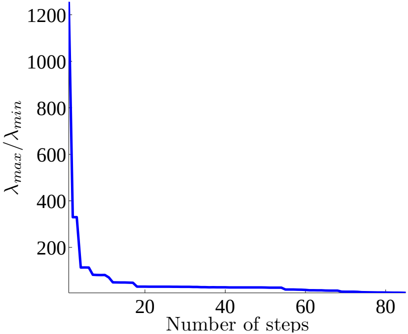
Repeat Add and Del stages until the evidence $\mathcal{E}(f_{\mathcal{A}})$ become stable.

Model selection by the error function minimization

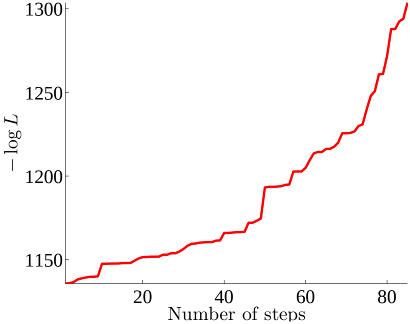


Add and **Delete** features until the the error function up.

Removing a feature during one stage

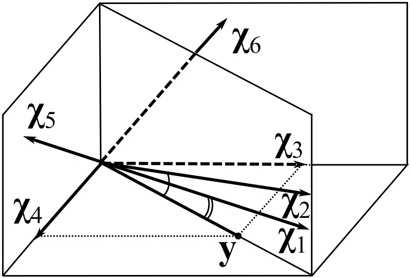


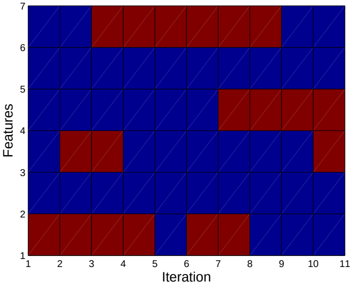
Condition number η



$-\ln S$

The condition number η and the likelihood $-\ln S$ depends on the number of the removed features.





The red color means the feature is included into the active set $\mathcal{A}$.

Algorithms	$S_{\mathcal{L}}$	$S_{\mathcal{C}}$	C_p	lg κ	k
Genetic	0.073	0.107	337	13	26
GMDH	0.146	0.194	745	6	10
Stepwise	0.128	0.154	644	7	12
Ridge	0.111	0.146	832	33	160
Lasso	0.121	0.147	611	5	18
Stage	0.071	0.096	324	9	26
FOS	0.106	0.135	527	7	20
LARS	0.098	0.095	492	7	28
Evidence	0.097	0.123	469	5	21

The active variables, indexed by the set $\mathcal{A} \subseteq \mathcal{J}$

are fixed to define the model $f(\mathbf{w}_\mathcal{A}, \mathbf{x}_\mathcal{A})$.

The mixture model $\mathfrak{h}$

is the set of models $\mathfrak{h} = \{f_k | k = 1, \dots, K\}$, such that

$$\mathfrak{h} = \sum_{k=1}^{K} \pi_k f_k(\mathbf{w}_k), \text{ where } \sum_{k=1}^{K} \pi_k = 1, \quad \pi_k = 1 \geq 0.$$

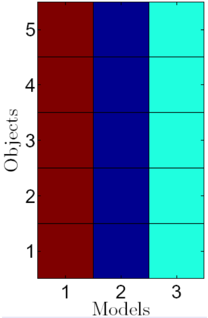
The multilevel model $\mathfrak{f}$, defined by indexed

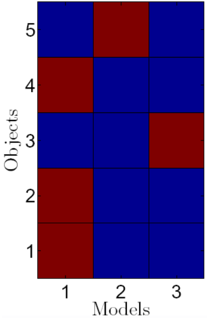
is the set of models $\mathfrak{f} = \{f_k | k = 1, \dots, K\}$, such that

$$\mathsf{E}(y_{i \in \mathcal{B}_k} | \mathbf{x}) = f(\mathbf{w}_k, \mathbf{x}_{i \in \mathcal{B}_k})$$

on the split

$$\mathcal{I} = \sqcup_{k=1}^{K} \mathcal{B}_k \ni i.$$





The evidence of the model

$$p(f_k \mid \mathbf{x}_i, y_i) = \frac{p(f_k, \mathbf{x}_i, y_i)}{p(\mathbf{x}_i, y_i)} = \frac{p(y_i \mid f_k, \mathbf{x}_i) p(f_k, \mathbf{x}_i)}{p(\mathbf{x}_i, y_i)}.$$

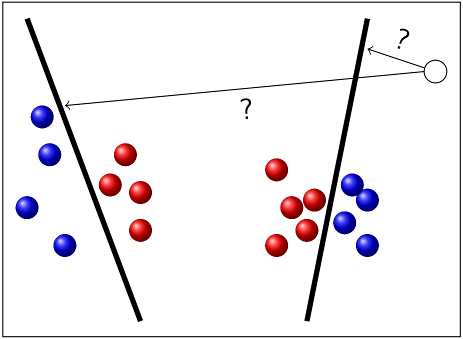
The evidence of two models

$$\frac{p(f_1 \mid \mathbf{x}_i, y_i)}{p(f_2 \mid \mathbf{x}_i, y_i)} = \frac{p(y_i \mid f_1, \mathbf{x}_i)}{p(y_i \mid f_2, \mathbf{x}_i)} \frac{p(f_1)}{p(f_2)}.$$

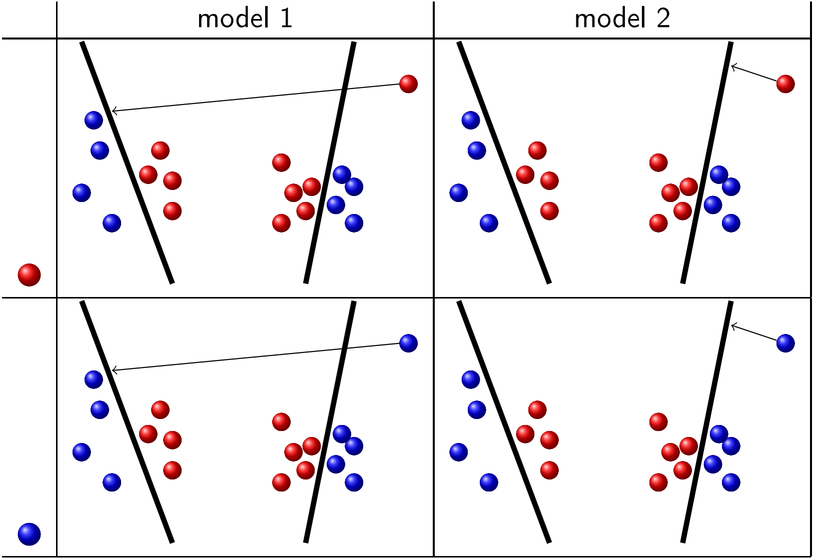
The decision rule: a sample corresponds to which model?

$$k_i^* = \arg \max_{k \in \{1, \ldots, K\}} p(y_i \mid f_k, \mathbf{x}_i).$$

Model detection



Model detection



Safe strategy of model selection

$$k_i^* = \arg \max_{k \in \{1,\dots,K\}} \min_{u \in \{0,1\}} p(u \,|\, f_k, \mathbf{x}_i).$$

Logistic regression case

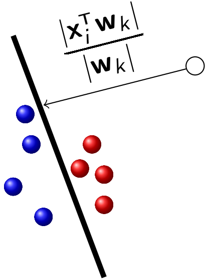
$$k_i^* = \arg \max_{k \in \{1,\dots,K\}} \{\min(\sigma(\mathbf{x}_i^{\mathsf{T}} \mathbf{w}_k), \sigma(-\mathbf{x}_i^{\mathsf{T}} \mathbf{w}_k))\}.$$

Transform the rule

$$k_i^* = \arg \max_{k \in \{1,\dots,K\}} \sigma(-|\mathbf{x}_i^{\mathsf{T}} \mathbf{w}_k|) = \arg \min_{k \in \{1,\dots,K\}} \sigma(|\mathbf{x}_i^{\mathsf{T}} \mathbf{w}_k|).$$

$$k_i^* = \arg \min_{k \in \{1,\ldots,K\}} \sigma(|\mathbf{x}_i^\mathsf{T} \mathbf{w}_k|),$$

$$k_i^* = \arg \min_{k \in \{1,\ldots,K\}} |\mathbf{x}_i^\mathsf{T} \mathbf{w}_k|.$$



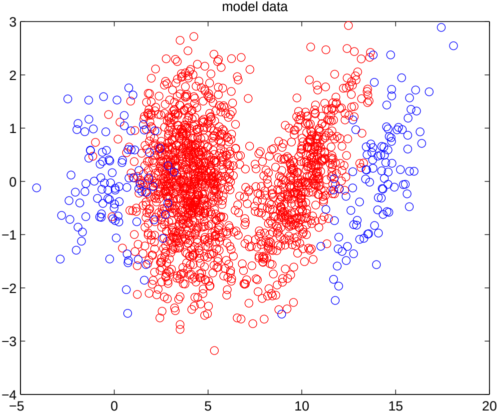
The object corresponds to the nearest separation hyperplane about accuracy up to $|\mathbf{w}_k|$.

M-step Estimate the model parameters $\mathbf{w}_k$ for each model $f_k, k = 1, \ldots, K$ using Newton-Raphson method (IRLS).

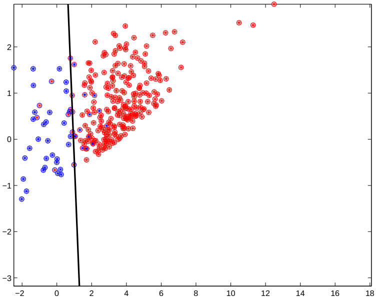
E-step Detect a corresponding model using the decision rule (the model evidence).

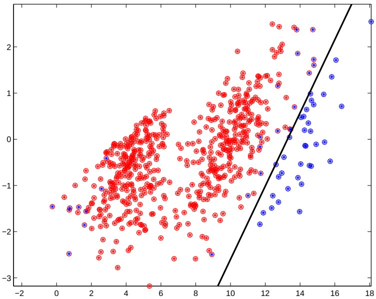
$$k_i^* = \arg \min_{k \in \{1,\ldots,K\}} |\mathbf{x}_i^{\mathsf{T}} \mathbf{w}_k|.$$

Synthetic data

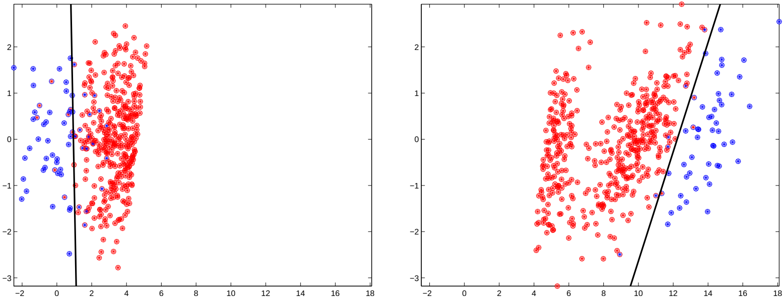


Step 1

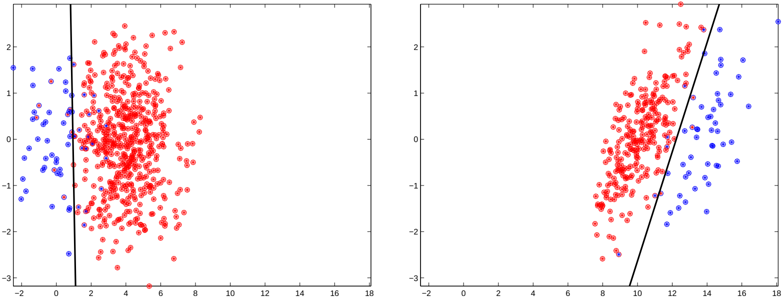




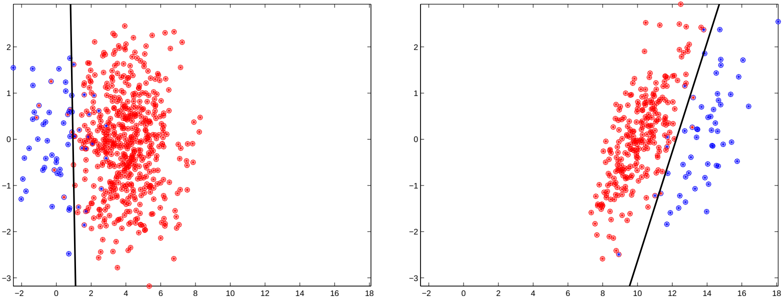
Step 2



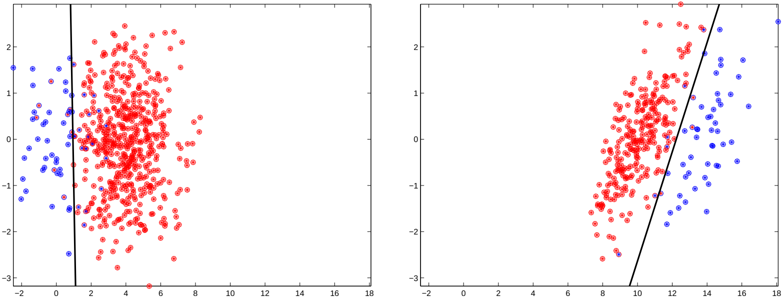
Step 3



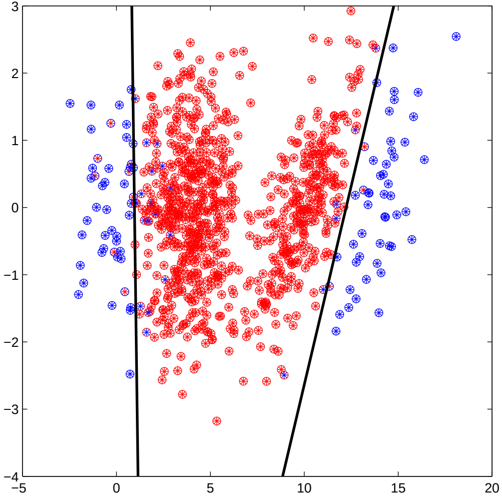
Step 4



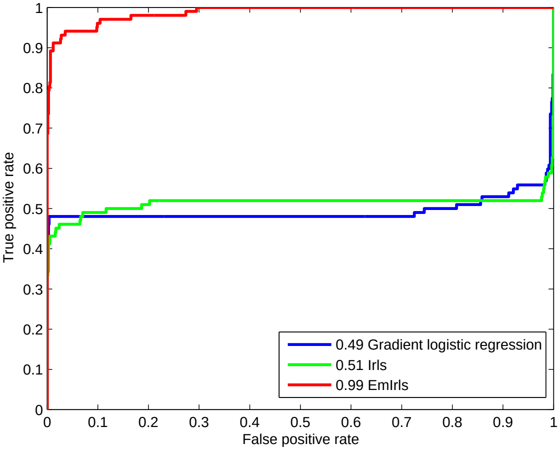
Step 5



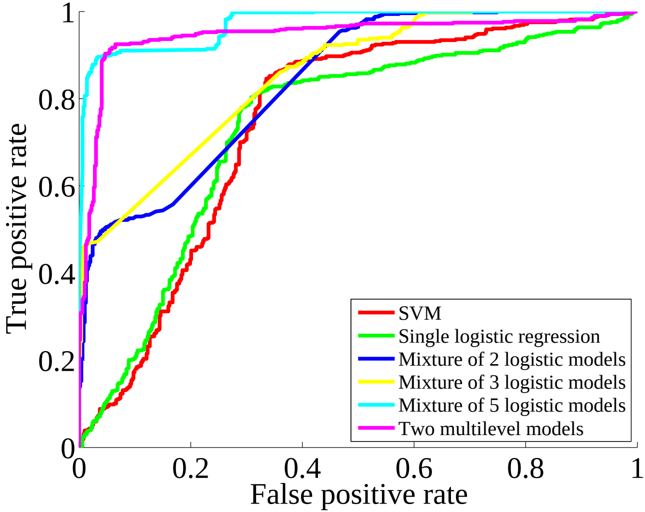
The sample set classification

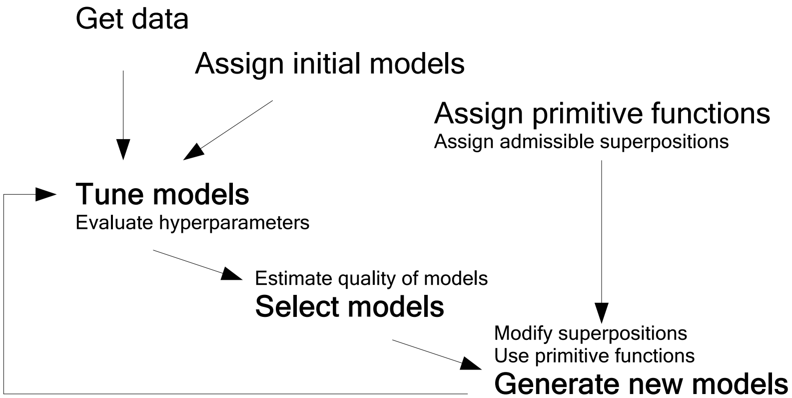


Computational experiment: the synthetic data-1, AUC



Computational experiment: the synthetic data-2, AUC




Get data
Assign initial models
Assign primitive functions
Assign admissible superpositions
Tune models
Evaluate hyperparameters
Estimate quality of models
Select models
Modify superpositions
Use primitive functions
Generate new models

The principle

- Hyperparameters are defined by the variance of model parameters,

  they could be used to select the stable and precise set of features.

Outline

- The strategy «generate various — select the best» is appeared to be successful for the credit scoring.
- One shall use primitive functions to generate non-linear features...

  ... and evaluate hyperparameters to select the best features for the generalized linear model.