Model generation and model selection in credit scoring

Vadim STRIJOV

Russian Academy of Sciences
Computing Center

EURO 2010 Lisbon
July 14th

The workflow

Client's application & history

↓

Client's score: probability of fraud / default

↓

Accept (refuse) the application

↓

Make the agreement

↓

Client's history

State of art

Classics

- N. Siddiqi: Credit Risk Scorecards developing, 2004
- D. Hosmer, S. Lemeshov: Logistic Regression, 2000

New strategy

- H. Madala: Group Method of Data Handling, 1995
- J. Koza, I. Zelinka: Genetic Programming, 2004
- Y. LeCun: Optimal Brain Surgery, 1985
- C. Bishop, J. Nabney: Model Selection and Coherent Bayesian Inference, 2004
- P. Grunwald: Minimum Description Length Principle, 2009

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

$$\overbrace{0 \longrightarrow \quad \underbrace{30+ \longrightarrow 60+ \longrightarrow 90+}_{\text{Default: deliquency 90+ on any, but } 1^{\text{st}}} \longrightarrow 120+}^{\text{Fraud: deliquency 90+ on } 3^{\text{rd}}} \quad \longrightarrow 150+$$

Scorecard developing

- 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

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

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

Scale conversion and grouping

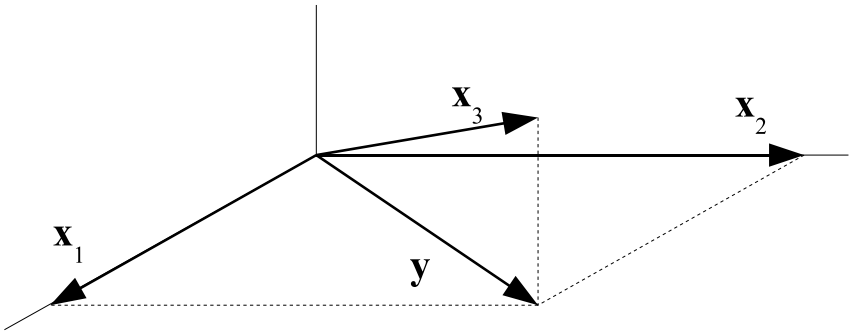
- Applicant's industry, nominal scale

Nominal	Tourism	Banking	Education
John	1	0	0
Thomas	0	1	0
Sara	0	0	1

- Applicant's education, ordinal scale

Ordinal	Primary	Secondary	Higher
John	1	0	0
Thomas	1	1	0
Sara	1	1	1

Univariate vs. multivariate



Problem statement, the data

1. The data set: $\mathbf{x} \in \mathbb{R}^n$, $y \in \mathbb{R}$,
$$D = \{(\mathbf{x}^1, y^1), \ldots, (\mathbf{x}^i, y^i), \ldots, (\mathbf{x}^m, y^m)\};$$
2. the design matrix $X \in \mathbb{R}^{m \times n}$,
$$X = (\mathbf{x}_1, \ldots, \mathbf{x}_j, \ldots, \mathbf{x}_n);$$
3. dependent variable $\mathbf{y} \sim \text{Bernoulli}(\boldsymbol{\sigma})$;
$$\mathbf{y} = (y^1, \ldots, y^m)^T,$$
4. the model
$$\mathbf{y} = \boldsymbol{\sigma}(\mathbf{w}) + \varepsilon, \qquad \boldsymbol{\sigma}(\mathbf{w}) = \frac{1}{1 + \exp(-X\mathbf{w})}.$$

Indexes of

- the objects, $\{1, \ldots, i, \ldots, m\} = \mathcal{I}$, split $\mathcal{I} = \mathcal{L} \sqcup \mathcal{T}$;
- the features $\{1, \ldots, j, \ldots, n\} = \mathcal{J}$, denote by $\mathcal{A}$ the active set.

Problem statement, the target function

The quality criterion is the log likelihood function

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

We must find the active set $\mathcal{A} \subset \mathcal{J}$ and the model parameters $\mathbf{w}_\mathcal{A}$, such that

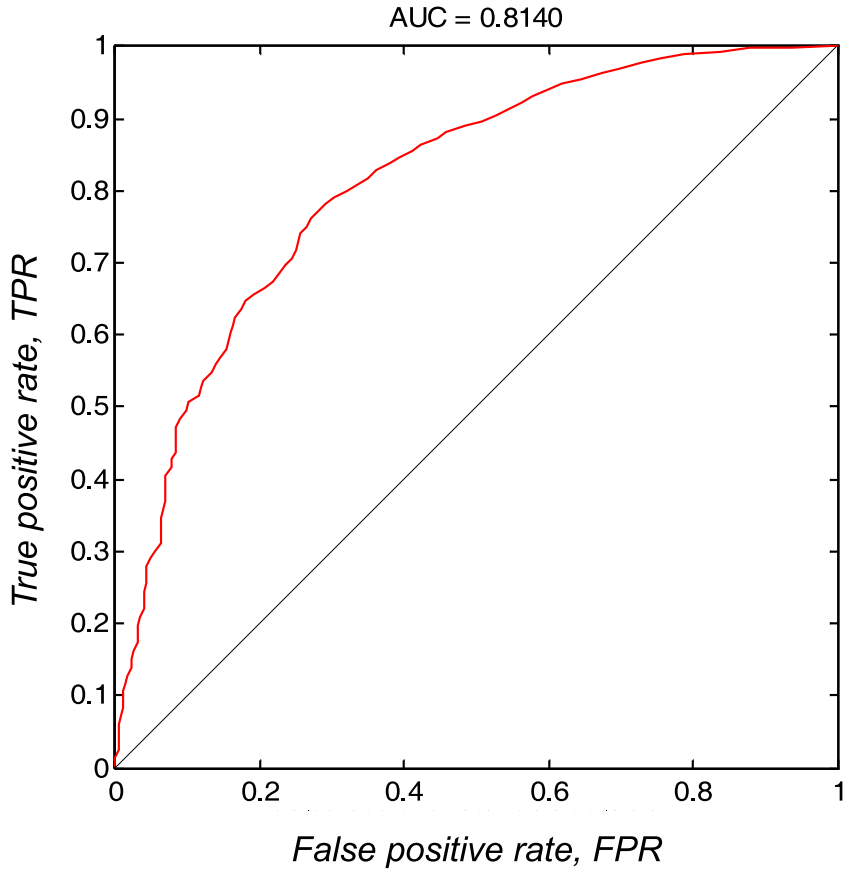
$$S(\mathbf{w})_\mathcal{A} \longrightarrow \min_{\mathcal{A}\subseteq\mathcal{J}, i\in\mathcal{T}},$$

where $\mathcal{I} = \mathcal{L} \sqcup \mathcal{T}$.

Indexes of

- the objects, $\{1, \ldots, i, \ldots, m\} = \mathcal{I}$, split $\mathcal{I} = \mathcal{L} \sqcup \mathcal{T}$;
- the features $\{1, \ldots, j, \ldots, n\} = \mathcal{J}$, denote by $\mathcal{A}$ the active set.

ROC-curve as the quality criterion



	P	N
P^*	TP	FP
N^*	FN	TN

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

Grouping, the optimization problem

We have an initial model defined by the set $\mathcal{A}$; append the generated set of the features and estimate their significance.

$$\begin{array}{lccccll} \xi = & 1 & 2 & 3 & \dots & c, & c \text{ is the number of categories}, \xi \in C; \\ & \downarrow & \downarrow & \downarrow & & \downarrow & \\ x_j = & \gamma_1 & \gamma_2 & \gamma_3 & \dots & \gamma_c, & |\Gamma| \text{ is the number of groups}, \gamma \in \Gamma. \end{array}$$

We must find the function

$$h : C \to \Gamma.$$

The optimization problem is

$$(h, |\Gamma|) = \arg \max_{h \in H} S(w)_{\mathcal{A} \cup j}.$$

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 $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_j, \ldots, \mathbf{x}_n\}$.

The number of features exceeds the number of clients!

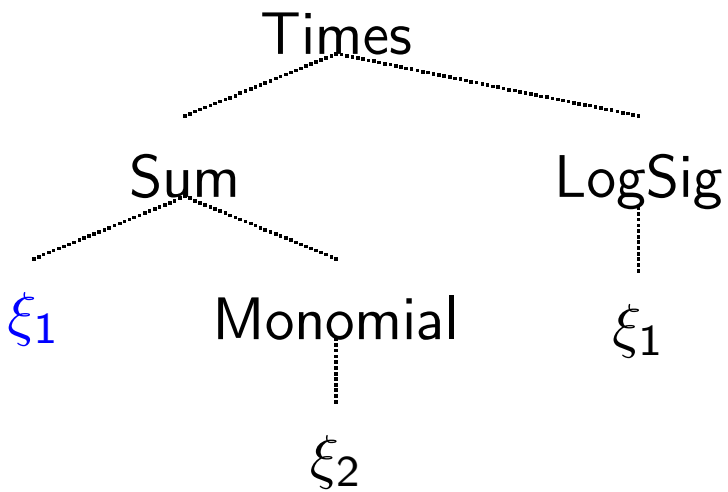
Examples of generated features

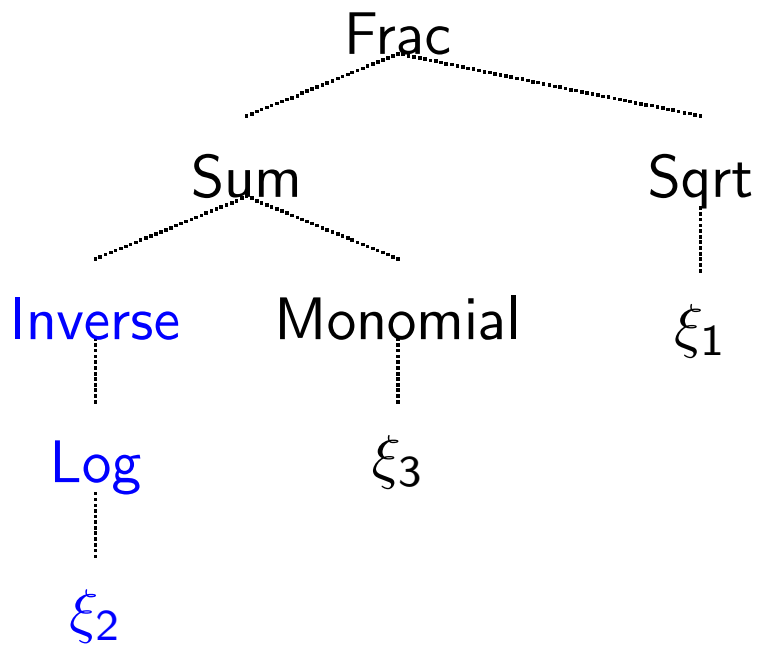
- **Frac**(Period of residence, Undeclared income)
- **Frac**(**Seg**(Period of employment), Term of contract)
- **And**(Income confirmation, Bank account)
- **Times**(**Seg**(Score hour), **Frac**(**Seg**(Period of employment), Salary))

Feature generation

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.





Structural parameters and model selection

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	$\ldots$	0
0	1	$\ldots$	0
$\ldots$	$\ldots$	$\ldots$	$\ldots$
1	1	$\ldots$	1

Coherent Bayesian inference

$f_1, ..., f_M$ are the competitive models,
$P(f_i|D)$ is the posterior probability, $P(D|f_i)$ is the evidence

$$P(f_i|D) = \frac{P(D|f_i)P(f_i)}{\sum_{j=1}^{M} P(D|f_j)P(f_j)}. \quad (1)$$

The models f_i and f_j could be compared as

$$\frac{P(f_i|D)}{P(f_j|D)} = \frac{P(D|f_i)P(f_i)}{P(D|f_j)P(f_j)}.$$

The posterior probability of the parameters $\mathbf{w}$ given D

$$P(\mathbf{w}|D, f_i) = \frac{P(D|\mathbf{w}, f_i)P(\mathbf{w}|f_i)}{P(D|f_i)}, \quad (2)$$

the model evidence in the parameter space is

$$P(D|f_i) = \int P(D|\mathbf{w}, f_i)P(\mathbf{w}|f_i)d\mathbf{w}.$$

Data generation hypothesis

$$y = f_i(\mathbf{w}, \mathbf{x}) + \nu,$$

the likelihood function is

$$P(y|\mathbf{x}, \mathbf{w}, \beta, f_i) \triangleq P(D|\mathbf{w}, \beta, f) = \exp(-\beta E_D(D|\mathbf{w}, f_i)) Z_D^{-1}(\beta),$$

the regularization function

$$P(\mathbf{w}|\alpha, f_i) = \exp(-\alpha E_W(\mathbf{w}|f_i)) Z_W^{-1}(\alpha),$$

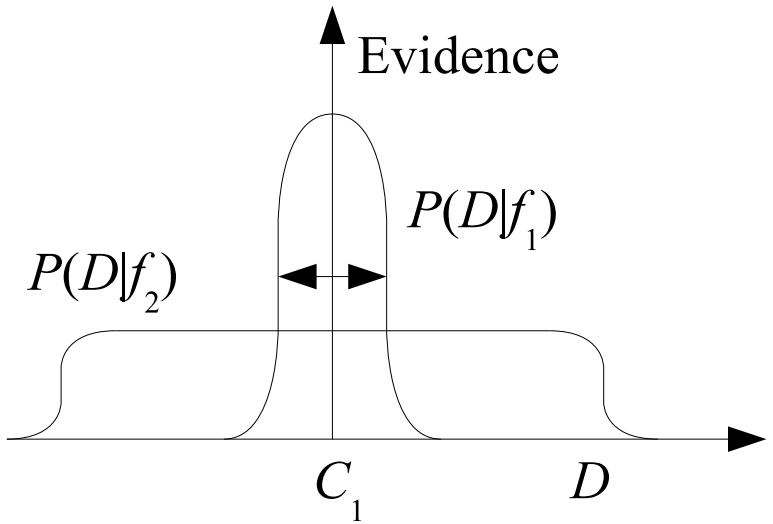
$\beta = \sigma_\nu^{-2}$ the variance of data noise, $\alpha = \sigma_{\mathbf{w}}^{-2}$ the variance of parameters.

The desired target function

$$P(\mathbf{w}|D, \alpha, \beta, f_i) = \frac{P(D|\mathbf{w}, \beta) P(\mathbf{w}|\alpha)}{P(D|\alpha, \beta)} = \frac{\exp(-S(\mathbf{w}|f_i))}{Z_S(\alpha, \beta)}$$

and the error function $S(\mathbf{w}) = \alpha E_W + \beta E_D$.

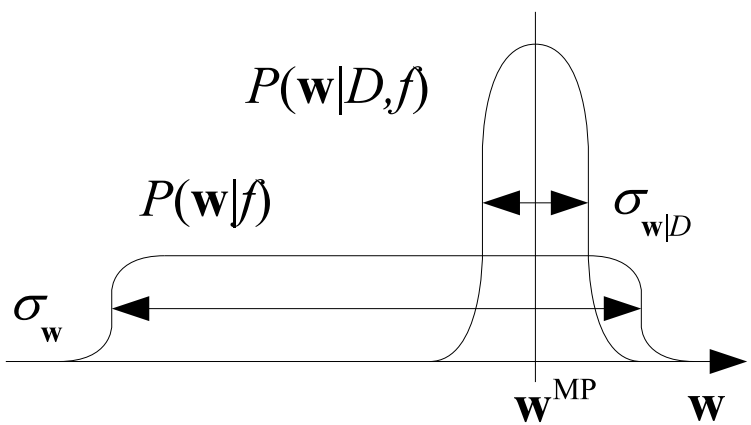
Model comparison

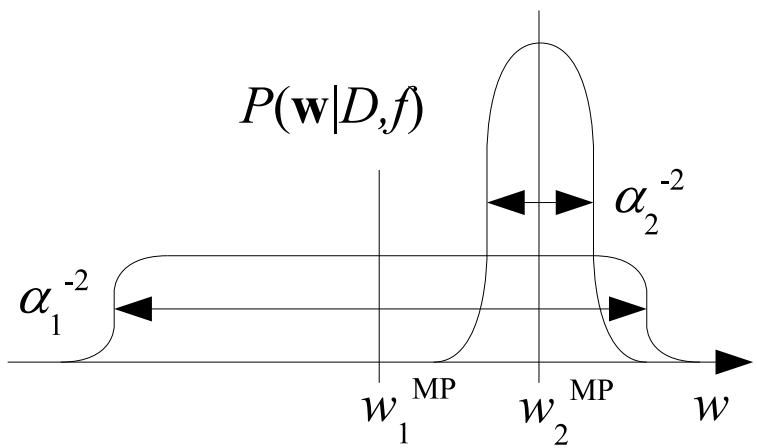


The posterior distribution and comparison of elements

The vector of the parameters $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, A)$ is a multivariate random variable. We could consider the covariance matrix A as

1. $A = \text{diag}(\alpha, \ldots, \alpha)$,
2. $A = \text{diag}(\alpha_1, \ldots, \alpha_W)$,
3. non-diagonal.



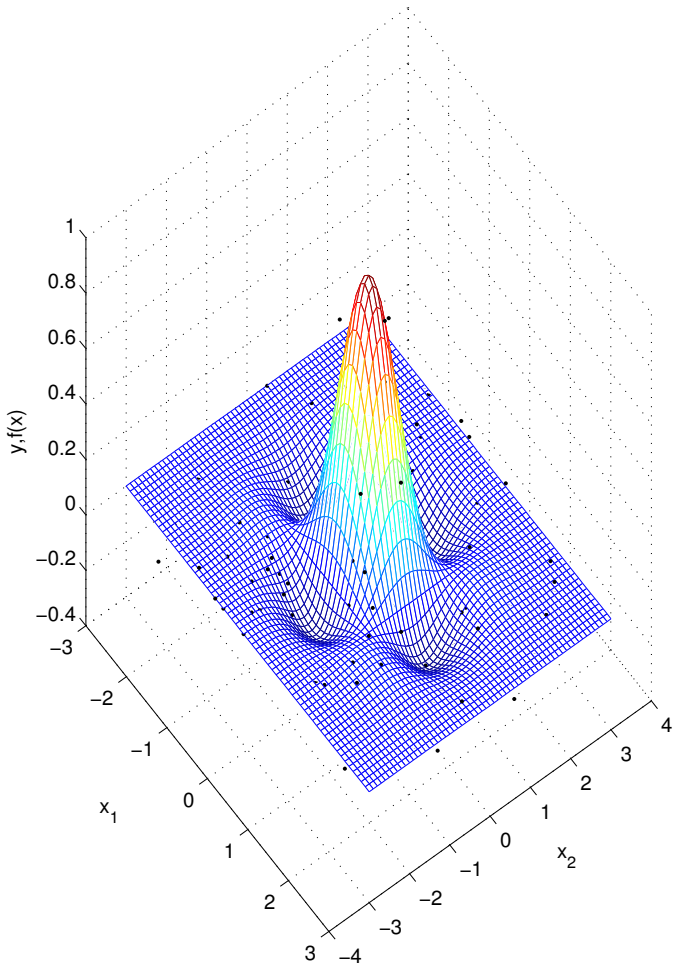


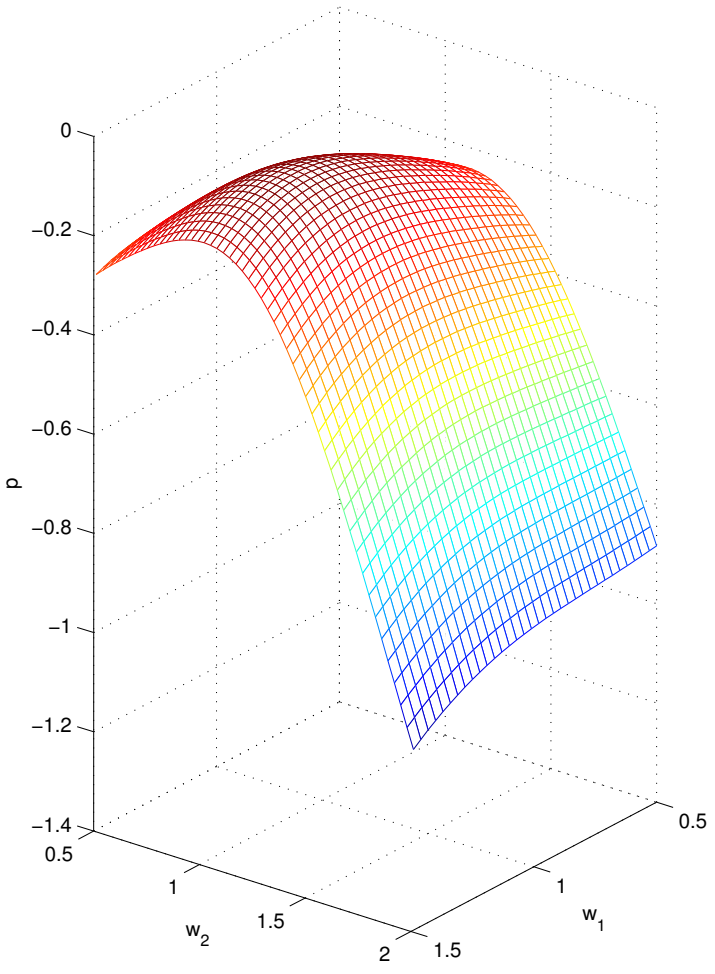
The evidence $P(D|f_i) = \int P(D|\mathbf{w}, f_i) P(\mathbf{w}|f_i) d\mathbf{w}.$

Optimal brain surgery

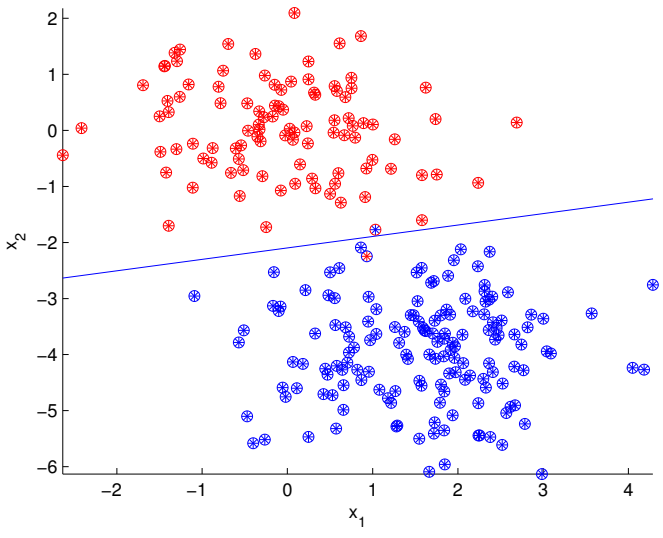
- Approximate $S(\mathbf{w})$:
  $S(\mathbf{w}+\Delta\mathbf{w}) = S(\mathbf{w}) + \mathbf{g}^T(\mathbf{w})\Delta\mathbf{w} + \frac{1}{2}\Delta\mathbf{w}^T H \Delta\mathbf{w} + o(\|\mathbf{w}\|^3)$.
- Elimination a feature is equivalent to $\mathbf{e}_i^T \Delta\mathbf{w} + w_i = 0$.
- Minimize the quadratic form $\Delta\mathbf{w}^T H \Delta\mathbf{w}$ subject to $\mathbf{e}_i^T + w_i = 0$, for all i.
- The index of the eliminated feature is $i = \arg\min_i(\min_{\Delta\mathbf{w}}(\Delta\mathbf{w}^T H \Delta\mathbf{w} | \mathbf{e}_i^T + w_i = 0))$.
- Introduce Lagrange function $S = \Delta\mathbf{w}^T H \Delta\mathbf{w} - \lambda(\mathbf{e}_i^T + w_i)$.
- For all i $\Delta\mathbf{w} = -\frac{w_i}{[H^{-1}]_{ii}} H^{-1}\mathbf{e}_i$.
- The salience of the target function is $L_i = \frac{w_i^2}{2[H^{-1}]_{ii}}$.

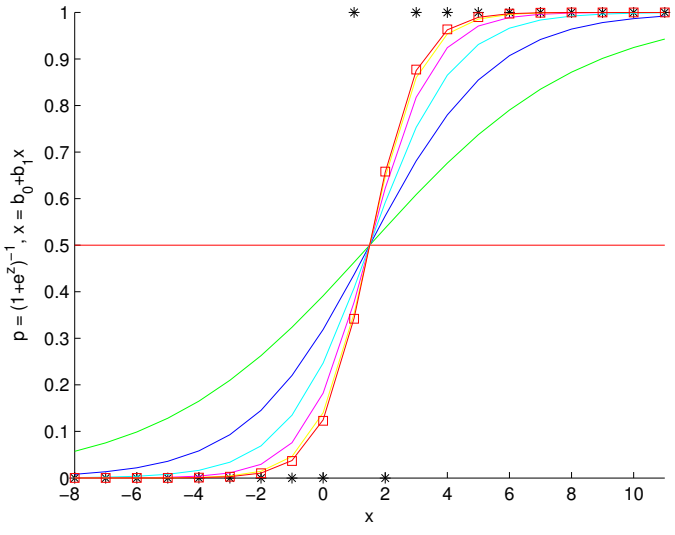
Compare elements of a model



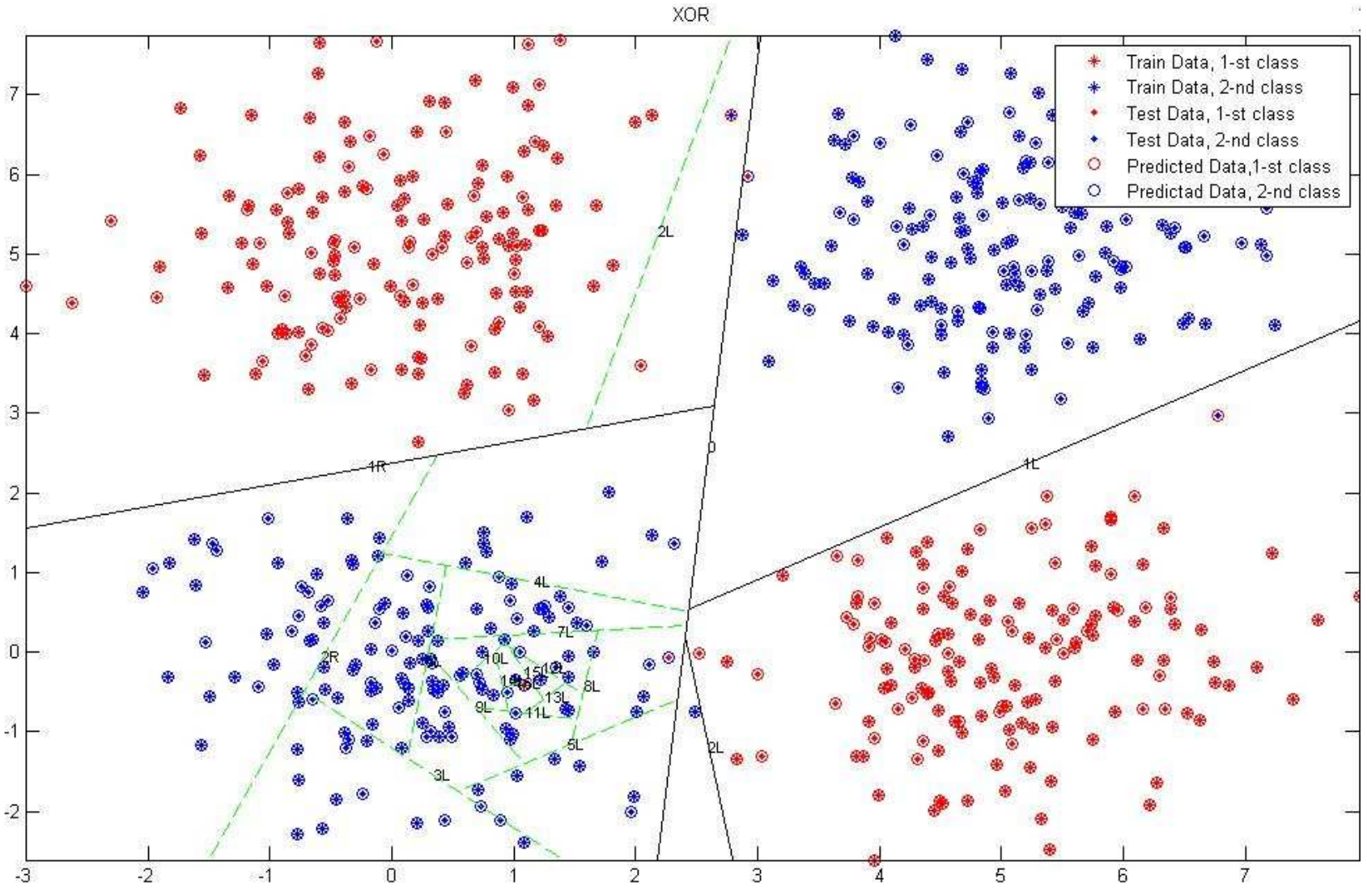


One-level modelling

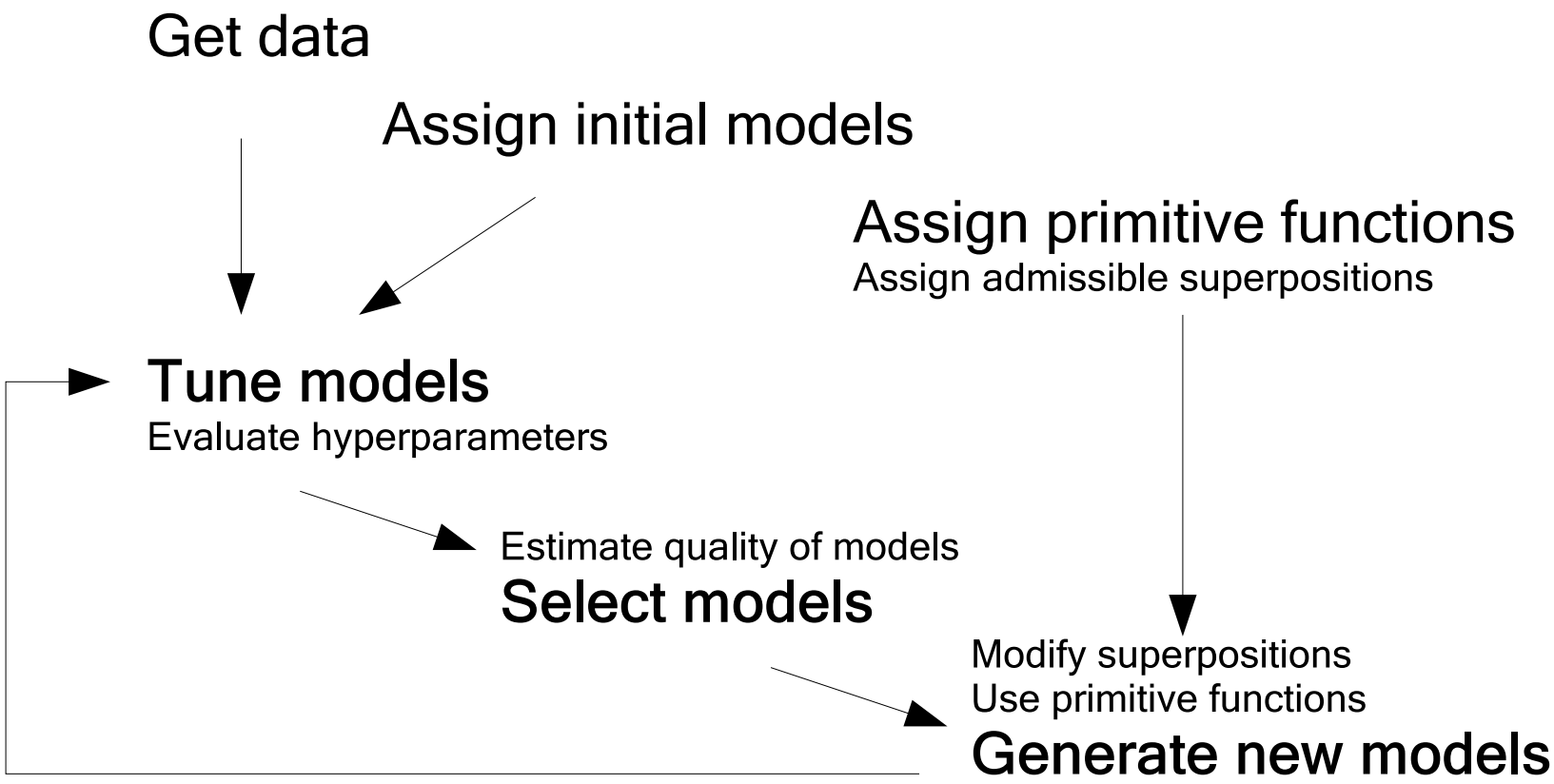




Multi-level modelling



The process of the model construction



Conclusion

Principle

- Hyperparameters are defined by the variance of model parameters,

  if the variance is large the model parameter and corresponded element could be eliminated.

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.