Model Genetation and Model Selection

Vadim Strijov,
Visiting Professor at IAM METU

Computing Center of the Russian Academy of Sciences

Institute of Applied Mathematics,
Middle East Technical University
June 6^{th}, 2012

The plan and the goal

1. Bank credit scoring as the example of the problem that requests feature generation.
2. Table of primitive functions.
3. Genetic and exhaustive generation algorithms.
4. Robust feature selection.

The goal is to show how the dimensions reduction algorithms could be robust.

Client's application & history

↓

Client's score: probability of fraud / default

↓

Accept (refuse) the application

↓

Make the agreement

↓

Client's history

- 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

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

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

- 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

- 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	
Term 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

- 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

Problem statement, the data

1. The data set: $\mathbf{x} \in \mathbb{R}^n$, $y \in \mathbb{R}$,
$$D = \{(\mathbf{x}_i, y_i)\};$$
2. the design matrix $X \in \mathbb{R}^{m \times n}$,
$$X = [\mathbf{x}_1^\mathsf{T}, \dots, \mathbf{x}_m^\mathsf{T}]^\mathsf{T};$$
3. dependent variable $\mathbf{y} \sim \text{Bernoulli}(\boldsymbol{\sigma})$;
$$\mathbf{y} = [y_1, \dots, y_m]^\mathsf{T},$$
4. the model
$$\mathbf{y} = \boldsymbol{\sigma}(\mathbf{w}) + \varepsilon, \qquad \boldsymbol{\sigma}(\mathbf{w}) = \frac{1}{1 + \exp(-X\mathbf{w})}.$$

Problem statement, the target function

Since

$$p(\mathbf{y}|\mathbf{w}) = \prod_{i=1}^{m} p_i^{y_i}(1-p_i)^{1-y_i}.$$

the quality criterion is the log likelihood function

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

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

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}\in\mathbb{R}^{|\mathcal{A}|}} S(\mathbf{w}|\mathcal{A}, D_\mathcal{L}) \quad \text{and}$$

$$\hat{\mathcal{A}} = \arg\min_{\mathcal{A}\subseteq\mathcal{J}} S(\mathcal{A}|\hat{\mathbf{w}}, D_\mathcal{T}),$$

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

Indexes of

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

Let the dependent variable $\mathbf{y}$ is distributed binomially:

$$\mathbf{y} \sim \mathcal{B}(f, 1-f).$$

The likelihood function

$$p(D|w, B, f) = \prod_{i \in \mathcal{I}} f_i^{y_i}(1-f_i)^{1-y_i},$$

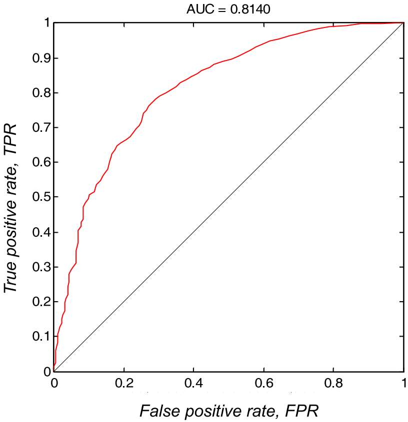
and the error function

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

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

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

ROC-curve as the quality criterion



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

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

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

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

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	-
Disjunction	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	-
Difference	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

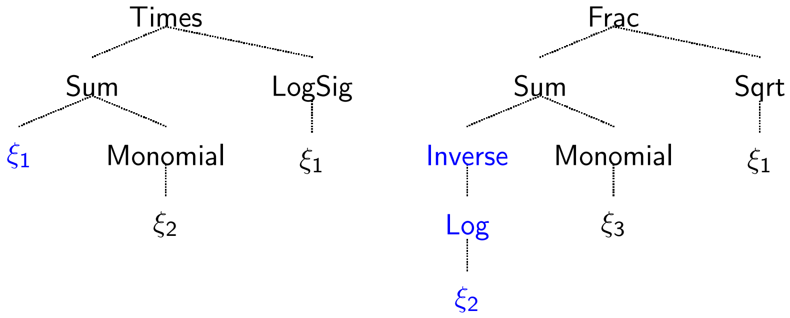
the set of the features $X = \{\mathbf{x}_1, \dots, \mathbf{x}_j, \dots, \mathbf{x}_n\}$.

The number of features exceeds the number of clients!

- **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.



1. Consider cartesian product $G \times \Xi$ of the set of non-generated variables Ξ the primitives G. Denote by a_ι the superpositions $g_v(\xi_u)$
2. Product superpositions a_ι no more than P times

$$a_\iota = g_v(\xi_u), \quad \text{where the index} \quad \iota = (v-1)U + u$$

and

$$x_j = \prod \underbrace{a_{\iota_1} \dots a_{\iota_p}}_{p \text{ times}}, \quad \text{where} \quad \iota \in \{1, \dots, UV\}, \quad p \in \{1, \dots, P\}.$$

In the other words

$$\xi_u \xrightarrow{g_v} g_v(\xi_u) \equiv a_\iota \xrightarrow{\prod^p} x_j, \qquad j \in \mathcal{J}.$$

Consider the linear models as the polynomial with a monomial $a_\iota = g_v(\xi_u)$

$$f(\mathbf{w}, \mathbf{x}) = \sum_{\iota=1}^{UV} w_\iota a_\iota + \sum_{\iota=1}^{UV} \sum_{\kappa=1}^{UV} w_{\iota\kappa} a_\iota a_\kappa + \sum_{\iota=1}^{UV} \sum_{\kappa=1}^{UV} \sum_{\tau=1}^{UV} w_{\iota\kappa\tau} a_\iota a_\kappa a_\tau + \cdots .$$

Let $G = \{g_1, \dots, g_l | g = g(\mathbf{b}, \cdot, \dots, \cdot)\}$ such that there are given

- the function $g : (\mathbf{b}, x) \mapsto x'$,
- its parameters $\mathbf{b}$ (the empty set is allowed),
- number of arguments $v(g)$ of the function g and the order of the arguments (zero arguments is allowed),
- domain $\text{dom}(g)$ and codomain $\text{cod}(g)$.

Consider the model $f(\mathbf{w}, \mathbf{x})$ as a superposition

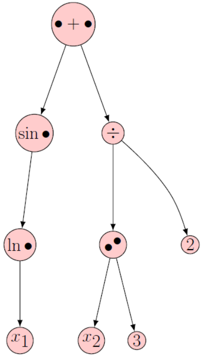
$$f(\mathbf{w}, \mathbf{x}) = (g_{i(1)} \circ \dots \circ g_{i(K)})(\mathbf{x}), \text{ where } \mathbf{w} = [\mathbf{b}_{i(1)}^{\mathsf{T}}, \dots, \mathbf{b}_{i(K)}^{\mathsf{T}}]^{\mathsf{T}}.$$

The admissible superposition f

is the superposition, which satisfies

$$\text{cod}(g_{i(k+1)}) \subseteq \text{dom}(g_{i(k)}), \text{ for any } k = 1, \dots, K-1.$$

- The vertex V_i corresponds to the primitive function $g_{s(i)}$.
- The number of outgoing nodes from the vertex V_i equal the number of arguments of $v(g_{s(i)})$.
- The order of the outgoing nodes from the vertex V_i equals the order of the arguments of $g_{s(i)}$.
- The leaves of the tree Γ_f corresponds to the independent variables x_i and constants; they are treated as the primitives $g(\emptyset)$.



The tree for the superposition $\sin(\ln x_1) + \frac{x_2^3}{2}$

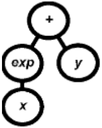
The structural density and depth

The superposition depth $d(f)$ is

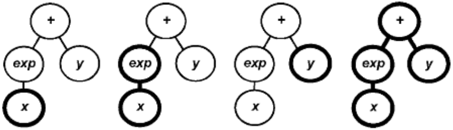
maximum depth of the tree Γ_f, number of the nodes V from the root to the most distanced leaf.

The superposition complexity $C(f)$ is

the number of all admissible subtrees of the tree Γ_f.



$$: \quad 1 + 2 + 1 + 4 = 8$$



Given: $G = \{g_u, h_v | u \in \mathcal{U}, v \in \mathcal{V}\}$ is a set of the primitive functions of one and two arguments, $\mathbf{x} = \{x_j | j \in \mathcal{J}\}$ — independent variables.

Step 1: $\mathcal{F}_1 = \left\{ f_s^{(1)} \right\} = \{g_u(x_j)\} \cup \{h_v(x_j, x_k)\}$, $k \in \mathcal{J}$, $s \in \left\{1, \dots, |\mathcal{U}| \cdot |\mathcal{J}| + |\mathcal{V}| \cdot |\mathcal{J}|^2 \right\}$.

Step k:

(Gen) Append to $\mathcal{F}$ the set

$$\mathcal{F}^{(k)} = \left\{ f_s^{(k)} \right\} = \left\{ g_u \left(f_{s'}^{(k-1)} \right) \right\} \cup \left\{ h_v \left(f_{s''}^{(k-1)}, f_{s'''}^{(k-1)} \right) \right\},$$

(Rem) which does not contain the superpositions, isomorphic to $g_u \left(f_s^{(k)} \right)$ and $h_v \left(f_s^{(k)}, f_{s'}^{(k)} \right)$ form the sets $\mathcal{F}^{(k)} \dots \mathcal{F}^{(1)}$.

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

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

1. There are set of binary vectors $\{\mathbf{a}_1, \dots, \mathbf{a}_P\}$, $\mathbf{a} \in \{0, 1\}^n$;
2. get two vectors $\mathbf{a}_p, \mathbf{a}_q$, $p, q \in \{1, \dots, P\}$;
3. chose random number $\nu \in \{1, \dots, n-1\}$;
4. split both vectors and change their parts:

$$[a_{p,1}, \dots, a_{p,\nu}, a_{q,\nu+1}, \dots, a_{q,n}] \to \mathbf{a'}_p,$$

$$[a_{q,1}, \dots, a_{q,\nu}, a_{p,\nu+1}, \dots, a_{p,n}] \to \mathbf{a'}_q;$$

5. choose random numbers $\eta_1, \dots, \eta_Q \in \{1, \dots, n\}$;
6. invert positions $\eta_1, \dots, \eta_Q$ of the vectors $\mathbf{a'}_p, \mathbf{a'}_q$;
7. repeat items 2-6 $P/2$ times;
8. evaluate the obtained models.

Repeat R times; here P, Q, R are the parameters of the algorithm and n is the number of the corresponding model features.

Discrete genetic algorithm for grouping

1. There are set of binary vectors $\{\mathbf{a}_1, \dots, \mathbf{a}_P\}$, $\mathbf{a} \in \{1, \dots, k\}^n$;
2. get two vectors $\mathbf{a}_p, \mathbf{a}_q$, $p, q \in \{1, \dots, P\}$;
3. chose random number $\nu \in \{1, \dots, n-1\}$;
4. split both vectors and change their parts:

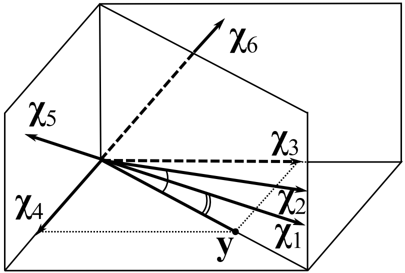
$$[a_{p,1}, \dots, a_{p,\nu}, a_{q,\nu+1}, \dots, a_{q,n}] \to \mathbf{a'}_p,$$

$$[a_{q,1}, \dots, a_{q,\nu}, a_{p,\nu+1}, \dots, a_{p,n}] \to \mathbf{a'}_q;$$

5. choose random numbers $\eta_1, \dots, \eta_Q \in \{1, \dots, n\}$;
6. replace values in positions $\eta_1, \dots, \eta_Q$ of the vectors $\mathbf{a'}_p, \mathbf{a'}_q$ for random values from $\{1, \dots, k\}$;
7. repeat items 2-6 $P/2$ times;
8. evaluate the obtained models.

Repeat R times; here P, Q, R are the parameters of the algorithm and k is desired number of categories.

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 $\mathbf{x}_j$.

Make singular values decomposition of the design matrix X

$$X = U\Lambda V^\mathsf{T},$$

where $UU^\mathsf{T} = I_m$, $V^\mathsf{T}V = I_n$, a Λ is the diagonal matrix with elements $\lambda_1 > \lambda_2 > \ldots > \lambda_r$, r is the rank of X, in our case $r = n$. The matrix $X^\mathsf{T}X$ is considered as the estimation of the correlation matrix.

$$X^\mathsf{T}X = V\Lambda^\mathsf{T}U^\mathsf{T}U\Lambda V^\mathsf{T} = V\Lambda^2V^\mathsf{T},$$

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

Find the conditional indexes

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

Obtain the variances of the parameters **w**

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

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}

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

Таблица: The decomposition of $\mathbf{var}(w_i)$

Conditional index	$\mathbf{var}(w_1)$	$\mathbf{var}(w_2)$	$\dots$	$\mathbf{var}(w_n)$
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.

Add stage:
Add the feature, which brings minimum to the error function $S(\mathbf{w})$

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

$$\mathcal{A}_k = \mathcal{A}_{k-1} \cup \{j^*\}$$

until $S(f_{\mathcal{A}_k} | \mathbf{w}^*, \mathcal{D})$ exceeds its minimum value on this stage but no more than a given ΔS_{Add}.

Del stage:
Dell the feature according to the Belsley method:

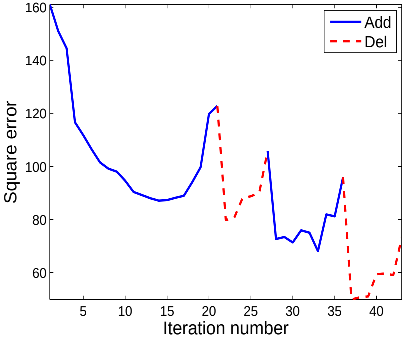
$$i^* = \sum_{g=1}^{t} \left[\eta_g^2 > \eta_t\right]$$

$$j^* = \arg\max_{j \in \mathcal{A}_{k-1}} \sum_{g=t-i^*+1}^{t} q_g^j$$

$$\mathcal{A}_k = \mathcal{A}_{k-1} \backslash j^*$$

until $S(f_{\mathcal{A}_k}|\mathbf{w}^*, \mathcal{D})$ exceeds its minimum value on this stage but no more than a given ΔS_{Del}.
Repeat Add and Del stages until the value of the error function $S(f_{\mathcal{A}_k}|\mathbf{w}^*, \mathcal{D})$ became stable.

The plot shows how the error function $S(\mathbf{w})$ varies during the steps. The control sample set $\mathcal{C}$ is used.



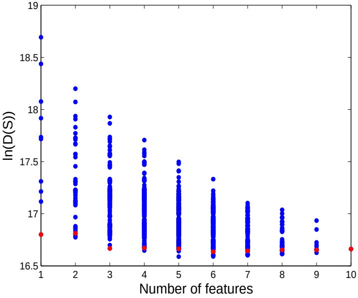
Expectation and variance of the error function, empirical approach

For given feature set $\mathcal{A} \in \mathcal{J}$ perform N-fold cross-validation procedure: $\mathcal{I} = \mathcal{L} \sqcup \mathcal{C}$. Treat the obtained values of the error function $S(\mathbf{w}|\mathcal{C})$ as realization of corresponding random variable. Estimate the expectation and variance:

$$ES = \frac{1}{N}\sum_{i=1}^{N} S_i,$$

$$DS = \frac{1}{N}\sum_{i=1}^{N}(S_i - ES)^2,$$

where N — number of folds (splits) and S_i is computed on the i-th split.



Red dots show the minimum expectation ES for the corresponding number of features $|\mathcal{A}|$.

Algorithms	$S_{\mathcal{L}}$	$S_{\mathcal{C}}$	AIC	BIC	C_p	$\lg \kappa$	k
Genetic	0.073	0.107	-1152	-1072	337	13	26
GMDH	0.146	0.194	-1076	-1045	745	6	10
Stepwise	0.128	0.154	-1092	-1055	644	7	12
Ridge	0.111	0.146	-819	-330	832	33	160
Lasso	0.121	0.147	-1089	-1034	611	5	18
Stage	0.071	0.096	-1157	-1077	324	9	26
FOS	0.106	0.135	-1105	-1044	527	7	20
LARS	0.098	0.095	-1102	-1017	492	7	28
Evidence	0.097	0.123	-1118	-1054	469	5	21

See

mvr.svn.sourceforge.net/viewvc/mvr/lectures/Strijov2012IAM.METU.Part3.pdf

or for short

bit.ly/K3i8zJ

The next

1. model comparison,
2. multimodelling.