Вычислительная сложность восстановления обобщенных линейных моделей зависимостей

Моттль Вадим Вячеславович
Вычислительный центр РАН, Москва

Сулимова Валентина Вячеславовна
Тульский государственный университет

Морозов Алексей Олегович, Пугач Илья Владимирович
Московский физико-технический институт

Татарчук Александр Игоревич
Вычислительный центр РАН, Москва

Pattern recognition and Regression: Two particular cases of Dependence estimation

The generalized scenario:

$\mathbf{x} \in \mathbb{R}^n$ – real-world objects observable through real-valued features

$y \in \mathbb{Y}$ – a hidden property of each object

$y = f(\mathbf{x}) : \mathbb{R}^n \to \mathbb{Y}$ – the unknown dependence that exists if reality

$\{(\mathbf{x}_j, y_j),\ j = 1, \dots, N\}$ – the set of precedents (training set)

$\hat{y} = \hat{f}(\mathbf{x}) : \mathbb{R}^n \to \mathbb{Y}$ – it is required to generate a decision rule applicable to each $\mathbf{x} \in \mathbb{R}^n$

$\hat{y} \approx y$ (to approximately restore the dependence)

If $\mathbb{Y} = \mathbb{R}$ this is regression estimation $y = f(\mathbf{x}) : \mathbb{R}^n \to \mathbb{R}$

If $\mathbb{Y} = \{-1, 1\}$ this is two-class pattern recognition $y = f(\mathbf{x}) : \mathbb{R}^n \to \{-1, 1\}$

Generalized Linear Model (GLM) of the hidden dependence

John Nelder. Generalized Linear Models. Journal of the Royal Statistical Society. Series A, Vol. 135, Issue 3, 1972, pp. 370-384.

$z(\mathbf{x}|\mathbf{a}, b) = \mathbf{a}^T \mathbf{x} + b : \mathbb{R}^n \to \mathbb{R}$ the generalized linear model of the dependence

Parameters of the model: $\mathbf{a} \in \mathbb{R}^n$ – direction vector, $b \in \mathbb{R}$ – bias

Pattern recognition and Regression: Two particular cases of Dependence estimation

The generalized scenario:

$\mathbf{x} \in \mathbb{R}^n$ – real-world objects observable through real-valued features

$y \in \mathbb{Y}$ – a hidden property of each object

$y = f(\mathbf{x}) : \mathbb{R}^n \to \mathbb{Y}$ – the unknown dependence that exists if reality

$\left\{(\mathbf{x}_j, y_j),\ j = 1,\ldots,N\right\}$ – the set of precedents (training set)

$\hat{y} = \hat{f}(\mathbf{x}) : \mathbb{R}^n \to \mathbb{Y}$ – it is required to generate a decision rule applicable to each $\mathbf{x} \in \mathbb{R}^n$

$\hat{y} \approx y$ (to approximately restore the dependence)

If $\mathbb{Y} = \mathbb{R}$ this is regression estimation $y = f(\mathbf{x}) : \mathbb{R}^n \to \mathbb{R}$

If $\mathbb{Y} = \{-1, 1\}$ this is two-class pattern recognition $y = f(\mathbf{x}) : \mathbb{R}^n \to \{-1, 1\}$

Generalized Linear Model (GLM) of the hidden dependence

John Nelder. Generalized Linear Models. Journal of the Royal Statistical Society. Series A, Vol. 135, Issue 3, 1972, pp. 370-384.

$z(\mathbf{x}|\mathbf{a}, b) = \mathbf{a}^T \mathbf{x} + b : \mathbb{R}^n \to \mathbb{R}$ the generalized linear model of the dependence

Pattern recognition and Regression: Two particular cases of Dependence estimation

The generalized scenario:

$\mathbf{x} \in \mathbb{R}^n$ – real-world objects observable through real-valued features

$y \in \mathbb{Y}$ – a hidden property of each object

$y = f(\mathbf{x}) : \mathbb{R}^n \to \mathbb{Y}$ – the unknown dependence that exists if reality

$\left\{(\mathbf{x}_j, y_j),\ j = 1, \ldots, N\right\}$ – the set of precedents (training set)

$\hat{y} = \hat{f}(\mathbf{x}) : \mathbb{R}^n \to \mathbb{Y}$ – it is required to generate a decision rule applicable to each $\mathbf{x} \in \mathbb{R}^n$

$\hat{y} \approx y$ (to approximately restore the dependence)

If $\mathbb{Y} = \mathbb{R}$ this is regression estimation $y = f(\mathbf{x}) : \mathbb{R}^n \to \mathbb{R}$

If $\mathbb{Y} = \{-1, 1\}$ this is two-class pattern recognition $y = f(\mathbf{x}) : \mathbb{R}^n \to \{-1, 1\}$

Generalized Linear Model (GLM) of the hidden dependence

John Nelder. Generalized Linear Models. Journal of the Royal Statistical Society. Series A, Vol. 135, Issue 3, 1972, pp. 370-384.

$z(\mathbf{x} | \mathbf{a}, b) = \mathbf{a}^T \mathbf{x} + b : \mathbb{R}^n \to \mathbb{R}$ the generalized linear model of the dependence

$q(y, z) : \mathbb{Y} \times \mathbb{R} \to \mathbb{R}^+$ link function (loss function)

Pattern recognition and Regression: Two particular cases of Dependence estimation

The generalized scenario:

$\mathbf{x} \in \mathbb{R}^n$ – real-world objects observable through real-valued features

$y \in \mathbb{Y}$ – a hidden property of each object

$y = f(\mathbf{x}) : \mathbb{R}^n \to \mathbb{Y}$ – the unknown dependence that exists if reality

$\{(\mathbf{x}_j, y_j),\ j = 1, \dots, N\}$ – the set of precedents (training set)

$\hat{y} = \hat{f}(\mathbf{x}) : \mathbb{R}^n \to \mathbb{Y}$ – it is required to generate a decision rule applicable to each $\mathbf{x} \in \mathbb{R}^n$

$\hat{y} \approx y$ (to approximately restore the dependence)

If $\mathbb{Y} = \mathbb{R}$ this is regression estimation $y = f(\mathbf{x}) : \mathbb{R}^n \to \mathbb{R}$

If $\mathbb{Y} = \{-1, 1\}$ this is two-class pattern recognition $y = f(\mathbf{x}) : \mathbb{R}^n \to \{-1, 1\}$

Generalized Linear Model (GLM) of the hidden dependence

John Nelder. Generalized Linear Models. Journal of the Royal Statistical Society. Series A, Vol. 135, Issue 3, 1972, pp. 370-384.

$$\begin{cases} z(\mathbf{x}|\mathbf{a}, b) = \mathbf{a}^T\mathbf{x} + b : \mathbb{R}^n \to \mathbb{R} & \text{the generalized linear model of the dependence} \\ q(y, z) : \mathbb{Y} \times \mathbb{R} \to \mathbb{R}^+ & \text{link function (loss function)} \end{cases}$$

Pattern recognition and Regression: Two particular cases of Dependence estimation

The generalized scenario:

$\mathbf{x}\in\mathbb{R}^n$ – real-world objects observable through real-valued features

$y\in\mathbb{Y}$ – a hidden property of each object

$y=f(\mathbf{x}):\mathbb{R}^n\to\mathbb{Y}$ – the unknown dependence that exists if reality

$\{(\mathbf{x}_j,y_j),\ j=1,...,N\}$ – the set of precedents (training set)

$\hat{y}=\hat{f}(\mathbf{x}):\mathbb{R}^n\to\mathbb{Y}$ – it is required to generate a decision rule applicable to each $\mathbf{x}\in\mathbb{R}^n$

$\hat{y}\approx y$ (to approximately restore the dependence)

If $\mathbb{Y}=\mathbb{R}$ this is regression estimation $y=f(\mathbf{x}):\mathbb{R}^n\to\mathbb{R}$

If $\mathbb{Y}=\{-1,1\}$ this is two-class pattern recognition $y=f(\mathbf{x}):\mathbb{R}^n\to\{-1,1\}$

Generalized Linear Model (GLM) of the hidden dependence

John Nelder. Generalized Linear Models. Journal of the Royal Statistical Society. Series A, Vol. 135, Issue 3, 1972, pp. 370-384.

$$\begin{cases} z(\mathbf{x}|\mathbf{a},b)=\mathbf{a}^T\mathbf{x}+b:\mathbb{R}^n\to\mathbb{R} & \text{the generalized linear model of the dependence} \\ q(y,z):\mathbb{Y}\times\mathbb{R}\to\mathbb{R}^+ & \text{link function (loss function)} \end{cases}$$

$$\hat{y}(\mathbf{x}|\mathbf{a},b)=\arg\min_{y\in\mathbb{Y}} q\big(y,\ z(\mathbf{x}|\mathbf{a},b)\big)$$ decision rule

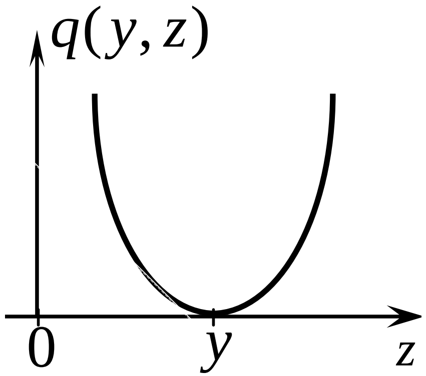
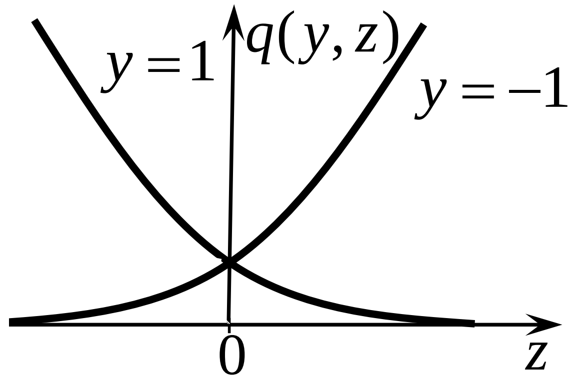
The Generalized Linear Model (GLM) of the dependence

$\begin{cases} z(\mathbf{x}\|\mathbf{a},b)=\mathbf{a}^T\mathbf{x}+b: \mathbb{R}^n \to \mathbb{R} \\ q(y,z): \mathbb{Y}\times\mathbb{R} \to \mathbb{R}^+ \end{cases}$	generalized linear feature of the entity represented by its initial feature vector $\mathbf{x}\in\mathbb{R}^n$
	link function, convex in $z\in\mathbb{R}$ for any $y\in\mathbb{Y}$
$\hat{y}(\mathbf{x}\|\mathbf{a},b)=\arg\min_{y\in\mathbb{Y}} q\big(y,\, z(\mathbf{x}\|\mathbf{a},b)\big)$	decision rule

The Generalized Linear Model (GLM) of the dependence

$\begin{cases} z(\mathbf{x}\|\mathbf{a},b)=\mathbf{a}^T\mathbf{x}+b: \mathbb{R}^n \to \mathbb{R} \\ q(y,z): \mathbb{Y}\times\mathbb{R} \to \mathbb{R}^+ \end{cases}$	generalized linear feature of the entity represented by its initial feature vector $\mathbf{x}\in\mathbb{R}^n$
	link function, convex in $z\in\mathbb{R}$ for any $y\in\mathbb{Y}$
$\hat{y}(\mathbf{x}\|\mathbf{a},b) = \arg\min_{y\in\mathbb{Y}} q\big(y,\, z(\mathbf{x}\|\mathbf{a},b)\big)$	decision rule

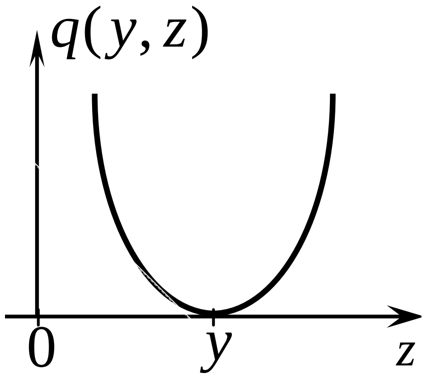
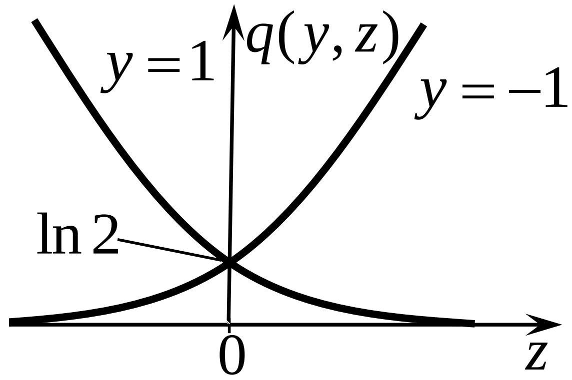
Particular cases

Regression $\mathbb{Y}=\mathbb{R}$	Two-class pattern recognition $\mathbb{Y}=\{-1,1\}$
$q(y,z)=(y-z)^2: \mathbb{R}\times\mathbb{R} \to \mathbb{R}^+$	$\lim_{z\to-\infty} q(y=+1,\, z)=\infty,\quad \lim_{z\to\infty} q(y=+1,\, z)=0,$ $\lim_{z\to-\infty} q(y=-1,\, z)=0,\quad \lim_{z\to\infty} q(y=-1,\, z)=\infty.$
	

The Generalized Linear Model (GLM) of the dependence

$\begin{cases} z(\mathbf{x}\vert\mathbf{a},b)=\mathbf{a}^T\mathbf{x}+b: \mathbb{R}^n \to \mathbb{R} \\ q(y,z): \mathbb{Y}\times\mathbb{R}\to\mathbb{R}^+ \end{cases}$	generalized linear feature of the entity represented by its initial feature vector $\mathbf{x}\in\mathbb{R}^n$ link function, convex in $z\in\mathbb{R}$ for any $y\in\mathbb{Y}$
$\hat{y}(\mathbf{x}\vert\mathbf{a},b)=\arg\min_{y\in\mathbb{Y}} q\big(y,\, z(\mathbf{x}\vert\mathbf{a},b)\big)$	decision rule

Particular cases

Regression $\mathbb{Y}=\mathbb{R}$	Two-class pattern recognition $\mathbb{Y}=\{-1,1\}$
$q(y,z)=(y-z)^2: \mathbb{R}\times\mathbb{R}\to\mathbb{R}^+$	$\lim_{z\to-\infty} q(y=+1,\, z)=\infty,\quad \lim_{z\to\infty} q(y=+1,\, z)=0,$ $\lim_{z\to-\infty} q(y=-1,\, z)=0,\quad \lim_{z\to\infty} q(y=-1,\, z)=\infty.$
	

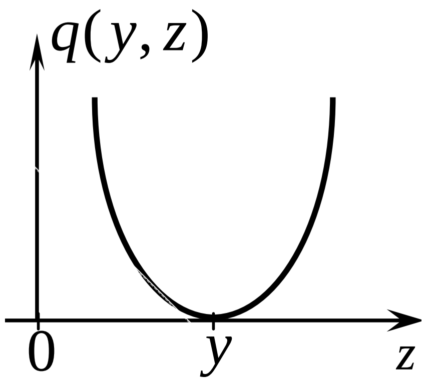
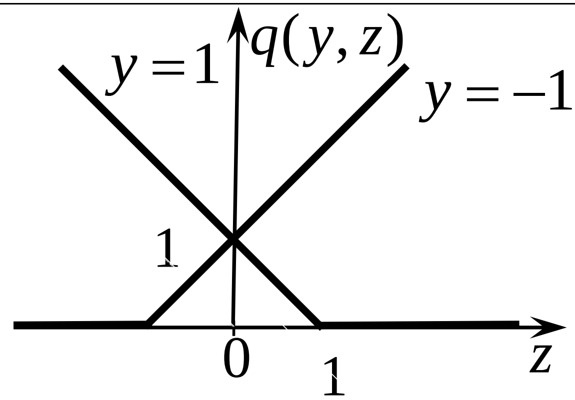
Logistic Regression

$$q(y,z)=\ln\left[1+\exp(-yz)\right]$$

The Generalized Linear Model (GLM) of the dependence

$\begin{cases} z(\mathbf{x}\|\mathbf{a},b)=\mathbf{a}^T\mathbf{x}+b: \mathbb{R}^n\to\mathbb{R} \\ q(y,z): \mathbb{Y}\times\mathbb{R}\to\mathbb{R}^+ \end{cases}$	generalized linear feature of the entity represented by its initial feature vector $\mathbf{x}\in\mathbb{R}^n$
	link function, convex in $z\in\mathbb{R}$ for any $y\in\mathbb{Y}$
$\hat{y}(\mathbf{x}\|\mathbf{a},b)=\arg\min_{y\in\mathbb{Y}} q\big(y,\ z(\mathbf{x}\|\mathbf{a},b)\big)$	decision rule

Particular cases

Regression $\mathbb{Y}=\mathbb{R}$	Two-class pattern recognition $\mathbb{Y}=\{-1,1\}$
$q(y,z)=(y-z)^2: \mathbb{R}\times\mathbb{R}\to\mathbb{R}^+$	$\lim_{z\to-\infty} q(y=+1,\ z)=\infty,\quad \lim_{z\to\infty} q(y=+1,\ z)=0,$ $\lim_{z\to-\infty} q(y=-1,\ z)=0,\quad \lim_{z\to\infty} q(y=-1,\ z)=\infty.$
	

Support Vector Machine (SVM)

$$q(y,z)=\max\big(0,\ 1-yz\big)$$

The commonly adopted principle of learning from precedents: Regularized empirical risk minimization

Set of precedents (training set): $\{(\mathbf{x}_j, y_j),\ j=1,\dots,N\}$

It is required to choose two parameters $(\mathbf{a}\in\mathbb{R}^n, b\in\mathbb{R})$ of the linear model

Criterion: Minimization of the loss within the bounds of the training set

$EmpR(\mathbf{a},b)=$		$\sum_{j=1}^{N} q(y_j,\ \mathbf{a}^T\mathbf{x}_j+b)\to\min$	empirical risk in the training set, instead of the average risk over “all the feasible” real-world entities

However, if $n>N$, there exist a continuum of minimum points $(\mathbf{a},b)\in\mathbb{R}^{n+1}$.

Regularized empirical risk minimization – finding the shortest vector among them

$$J(\mathbf{a},b)=\gamma\mathbf{a}^T\mathbf{a}+\sum_{j=1}^{N} q(y_j,\ \mathbf{a}^T\mathbf{x}_j+b)\to\min(\mathbf{a}\in\mathbb{R}^n, b\in\mathbb{R})$$

the ***simplest ridge*** regularization,
$0<\gamma\ll 1$, i.e., $\gamma\to 0$

The commonly adopted principle of learning from precedents: Regularized empirical risk minimization

Set of precedents (training set): $\{(\mathbf{x}_j, y_j),\ j=1,\dots,N\}$

It is required to choose two parameters $\left(\mathbf{a}\in\mathbb{R}^n, b\in\mathbb{R}\right)$ of the linear model

Criterion: Minimization of the loss within the bounds of the training set

$EmpR(\mathbf{a},b)=$		$\sum_{j=1}^{N} q(y_j,\ \mathbf{a}^T\mathbf{x}_j+b)\to\min$	empirical risk in the training set, instead of the average risk over “all the feasible” real-world entities

However, if $n>N$, there exist a continuum of minimum points $(\mathbf{a},b)\in\mathbb{R}^{n+1}$.

Regularized empirical risk minimization – finding the shortest vector among them

$$J(\mathbf{a},b)=\gamma\mathbf{a}^T\mathbf{a}+\sum_{j=1}^{N} q(y_j,\ \mathbf{a}^T\mathbf{x}_j+b)\to\min(\mathbf{a}\in\mathbb{R}^n, b\in\mathbb{R})$$

the ***simplest ridge*** regularization, $0<\gamma\ll 1$, i.e., $\gamma\to 0$

$$J(\mathbf{a},b|\mu)=\gamma\sum_{i=1}^{n}\begin{pmatrix} 2\mu\,|a_i|, & |a_i|\le\mu \\ \mu^2+a_i^2, & |a_i|>\mu \end{pmatrix}+\sum_{j=1}^{N} q(y_j,\ \mathbf{a}^T\mathbf{x}_j+b)\to\min$$

a more sophisticated ***selective ridge*** regularization

The commonly adopted principle of learning from precedents: Regularized empirical risk minimization

Set of precedents (training set): $\{(\mathbf{x}_j, y_j),\ j=1,...,N\}$

It is required to choose two parameters $\left(\mathbf{a}\in\mathbb{R}^n, b\in\mathbb{R}\right)$ of the linear model

Criterion: Minimization of the loss within the bounds of the training set

$EmpR(\mathbf{a},b)=$		$\sum_{j=1}^{N} q(y_j, \mathbf{a}^T\mathbf{x}_j+b)\to\min$	empirical risk in the training set, instead of the average risk over “all the feasible” real-world entities

However, if $n>N$, there exist a continuum of minimum points $(\mathbf{a},b)\in\mathbb{R}^{n+1}$.

Regularized empirical risk minimization – finding the shortest vector among them

$$J(\mathbf{a},b)=\gamma\mathbf{a}^T\mathbf{a}+\sum_{j=1}^{N} q(y_j, \mathbf{a}^T\mathbf{x}_j+b)\to\min(\mathbf{a}\in\mathbb{R}^n, b\in\mathbb{R})$$

the ***simplest ridge*** regularization, $0<\gamma\ll 1$, i.e., $\gamma\to 0$

$$J(\mathbf{a},b|\mu)=\gamma\sum_{i=1}^{n}\begin{pmatrix} 2\mu\,|a_i|,\ |a_i|\le\mu \\ \mu^2+a_i^2,\,|a_i|>\mu \end{pmatrix}+\sum_{j=1}^{N} q(y_j, \mathbf{a}^T\mathbf{x}_j+b)\to\min$$

a more sophisticated ***selective ridge*** regularization

Selectivity parameter $0\le\mu<\infty$. As μ grows, the penalty $\mu\,|a_i|$ drives to zero the coefficients at redundant features, which weakly contribute to diminishing of the empirical risk.

Result of optimization – a small subset of active features: $\hat{\mathbb{I}}(\mu)=\{i: a_i\neq 0\}\subseteq\{1,...,n\}$

Selective regularized empirical risk minimization

$$J(\mathbf{a},b|\mu)=\gamma\sum_{i=1}^{n}\begin{pmatrix}2\mu\,|a_i|, & |a_i|\le\mu\\ \mu^2+a_i^2, & |a_i|>\mu\end{pmatrix}+\sum_{j=1}^{N}q(y_j,\,\mathbf{a}^T\mathbf{x}_j+b)\to\min.$$ In scalar form:

$$J(a_1,...,a_n,b|\mu)=\gamma\sum_{i=1}^{n}\begin{pmatrix}2\mu\,|a_i|, & |a_i|\le\mu\\ \mu^2+a_i^2, & |a_i|>\mu\end{pmatrix}+\sum_{j=1}^{N}q\left(y_j,\sum_{i=1}^{n}a_i x_{j,i}+b\right)\to\min$$ $\begin{cases}\text{problem of convex}\\ \text{optimization with}\\ (n+1)\text{ variables}\end{cases}$

What will be interesting to us is the computational complexity of dependence estimation
In the general case, the computational complexity is polynomial relative to n.

In practice, the number of features is often much greater than the training set size $n\gg N$
If n is large, the polynomial computational complexity relative to n is inadmissible.

We are going to prove that this is not the case for dependence estimation.
The computational complexity will be polynomial with respect to N and linear to n.
To show this, it is enough to put the problem of selective regularized empirical risk minimization in the so-called ***disjoint form***:

$$\begin{cases}\gamma\sum_{i=1}^{n}\begin{pmatrix}2\mu\,|a_i|, & |a_i|\le\mu\\ \mu^2+a_i^2, & |a_i|>\mu\end{pmatrix}+\sum_{j=1}^{N}q(y_j,z_j)\to\min(a_1,...,a_n,b,z_1,...,z_N\,|\mu),\\ z_j=\sum_{i=1}^{n}a_i x_{j,i}+b,\quad j=1,...,N.\end{cases}$$

Such a disjoint writing allows for a dual formulation of the problem

The dual formulation and numerical solution of the disjoint empirical risk minimization problem

$$\begin{cases} \gamma \sum_{i=1}^{n} \begin{pmatrix} 2\mu\,|a_i|, & |a_i| \le \mu \\ \mu^2 + a_i^2, & |a_i| > \mu \end{pmatrix} + \sum_{j=1}^{N} q(y_j, z_j) \to \min(a_1, \ldots, a_n, b, z_1, \ldots, z_N \,|\, \mu), \\ z_j = \sum_{i=1}^{n} a_i x_{j,i} + b, \; j = 1, \ldots, N, \text{ Lagrange multipliers } \lambda_j; \end{cases}$$

disjoint writing of the empirical risk minimization problem

n – number of features, N – number of training objects.

The dual formulation and numerical solution of the disjoint empirical risk minimization problem

$$\begin{cases} \gamma\sum_{i=1}^{n}\begin{pmatrix} 2\mu\,|a_i|, & |a_i|\le\mu \\ \mu^2+a_i^2, & |a_i|>\mu \end{pmatrix}+\sum_{j=1}^{N}q(y_j,z_j)\to\min(a_1,...,a_n,b,z_1,...,z_N\,|\,\mu), \\ z_j=\sum_{i=1}^{n}a_i x_{j,i}+b,\ j=1,...,N,\ \text{Lagrange multipliers } \lambda_j; \end{cases}$$

disjoint writing of the empirical risk minimization problem

n – number of features, N – number of training objects.

Theorem. The solution of the disjoint problem is completely defined by:

1) solution $(\hat\lambda_1,...,\hat\lambda_N)$ of the convex dual problem

$$(\hat\lambda_1,...,\hat\lambda_N)=\arg\min\left\{\frac{1}{2}\sum_{i=1}^{n}\left\{\max\left[0,\left(\sum_{j=1}^{N}\lambda_j x_{j,i}\right)^2-\mu^2\right]\right\}+\sum_{j=1}^{N}\left[-\inf_{z\in\mathbb{R}}\left(\frac{1}{2\gamma}q(y_j,z)+\lambda_j z\right)\right]\right\},$$

$$\sum_{j=1}^{N}\lambda_j=0,\quad -\frac{1}{2\gamma}g_{\sup}(y_j)\le\lambda_j\le-\frac{1}{2\gamma}g_{\inf}(y_j),\ j=1,...,N;$$

Polynomial computational complexity in the number of raining objects N

2) independent computing $i=1,...,n$ $\quad \hat a_i=\begin{cases} 0, & \left(\sum_{j=1}^{N}(\hat\lambda_j x_{j,i}+\hat\xi_j\tilde x_{j,i})\right)^2\le\mu^2, \\ \sum_{j=1}^{N}\hat\lambda_j x_{j,i}, & \left(\sum_{j=1}^{N}(\hat\lambda_j x_{j,i}+\hat\xi_j\tilde x_{j,i})\right)^2>\mu^2. \end{cases}.$

Linear computational complexity in the number of features n

Regularization path along the selectivity axis

The dual problem once again:

$$(\hat{\lambda}_1,\ldots,\hat{\lambda}_N)=\arg\min\left\{\frac{1}{2}\sum_{i=1}^{n}\left\{\max\left[0,\left(\sum_{j=1}^{N}\lambda_j x_{j,i}\right)^2-\mu^2\right]\right\}+\sum_{j=1}^{N}\left[-\inf_{z\in\mathbb{R}}\left(\frac{1}{2\gamma}q(y_j,z)+\lambda_j z\right)\right]\right\},$$

$$\sum_{j=1}^{N}\lambda_j=0,\quad -\frac{1}{2\gamma}g_{\sup}(y_j)\le\lambda_j\le-\frac{1}{2\gamma}g_{\inf}(y_j),\ j=1,\ldots,N;$$

$$\hat{a}_i=\begin{cases}0, & \left(\sum_{j=1}^{N}(\hat{\lambda}_j x_{j,i}+\hat{\xi}_j\tilde{x}_{j,i})\right)^2\le\mu^2,\\ \sum_{j=1}^{N}\hat{\lambda}_j x_{j,i}, & \left(\sum_{j=1}^{N}(\hat{\lambda}_j x_{j,i}+\hat{\xi}_j\tilde{x}_{j,i})\right)^2>\mu^2.\end{cases}$$

The selectivity parameter $0\le\mu<\infty$ – the main hyperparameter of the dependence estimation problem.

Regularization path along the selectivity axis

The dual problem once again:

$$(\hat{\lambda}_1,\dots,\hat{\lambda}_N)=\arg\min\left\{\frac{1}{2}\sum_{i=1}^{n}\left\{\max\left[0,\left(\sum_{j=1}^{N}\lambda_j x_{j,i}\right)^2-\mu^2\right]\right\}+\sum_{j=1}^{N}\left[-\inf_{z\in\mathbb{R}}\left(\frac{1}{2\gamma}q(y_j,z)+\lambda_j z\right)\right]\right\},$$

$$\sum_{j=1}^{N}\lambda_j=0,\quad -\frac{1}{2\gamma}g_{\sup}(y_j)\le\lambda_j\le-\frac{1}{2\gamma}g_{\inf}(y_j),\ j=1,\dots,N;$$

$$\hat{a}_i=\begin{cases}0, & \left(\sum_{j=1}^{N}(\hat{\lambda}_j x_{j,i}+\hat{\xi}_j\tilde{x}_{j,i})\right)^2\le\mu^2,\\ \sum_{j=1}^{N}\hat{\lambda}_j x_{j,i}, & \left(\sum_{j=1}^{N}(\hat{\lambda}_j x_{j,i}+\hat{\xi}_j\tilde{x}_{j,i})\right)^2>\mu^2.\end{cases}$$

The selectivity parameter $0\le\mu<\infty$ – the main hyperparameter of the dependence estimation problem. If $\mu=0$, the criterions possess no selectivity property at all, and all the estimated components of the direction vector remain active. On the contrary, when the selectivity grows $\mu\to\infty$, all the direction vector components become zero. It is easy to find the maximal value of selectivity μ_0 that completely suppresses all the features.

Regularization path along the selectivity axis

The dual problem once again:

$$(\hat{\lambda}_1,\ldots,\hat{\lambda}_N)=\arg\min\left\{\frac{1}{2}\sum_{i=1}^{n}\left\{\max\left[0,\left(\sum_{j=1}^{N}\lambda_j x_{j,i}\right)^2-\mu^2\right]\right\}+\sum_{j=1}^{N}\left[-\inf_{z\in\mathbb{R}}\left(\frac{1}{2\gamma}q(y_j,z)+\lambda_j z\right)\right]\right\},$$

$$\sum_{j=1}^{N}\lambda_j=0,\quad -\frac{1}{2\gamma}g_{\sup}(y_j)\le\lambda_j\le-\frac{1}{2\gamma}g_{\inf}(y_j),\ j=1,\ldots,N;$$

$$\hat{a}_i=\begin{cases}0, & \left(\sum_{j=1}^{N}(\hat{\lambda}_j x_{j,i}+\hat{\xi}_j\tilde{x}_{j,i})\right)^2\le\mu^2,\\ \sum_{j=1}^{N}\hat{\lambda}_j x_{j,i}, & \left(\sum_{j=1}^{N}(\hat{\lambda}_j x_{j,i}+\hat{\xi}_j\tilde{x}_{j,i})\right)^2>\mu^2.\end{cases}$$

It is enough to vary selectivity in the interval $0\le\mu\le\mu_0$.

The idea: To divide this interval into a number of subintervals in logarithmic scale

$$\mu_m=0<\mu_{m-1}<\ldots<\mu_1<\mu_0$$
$$\leftarrow - - - - - - - - - - -$$

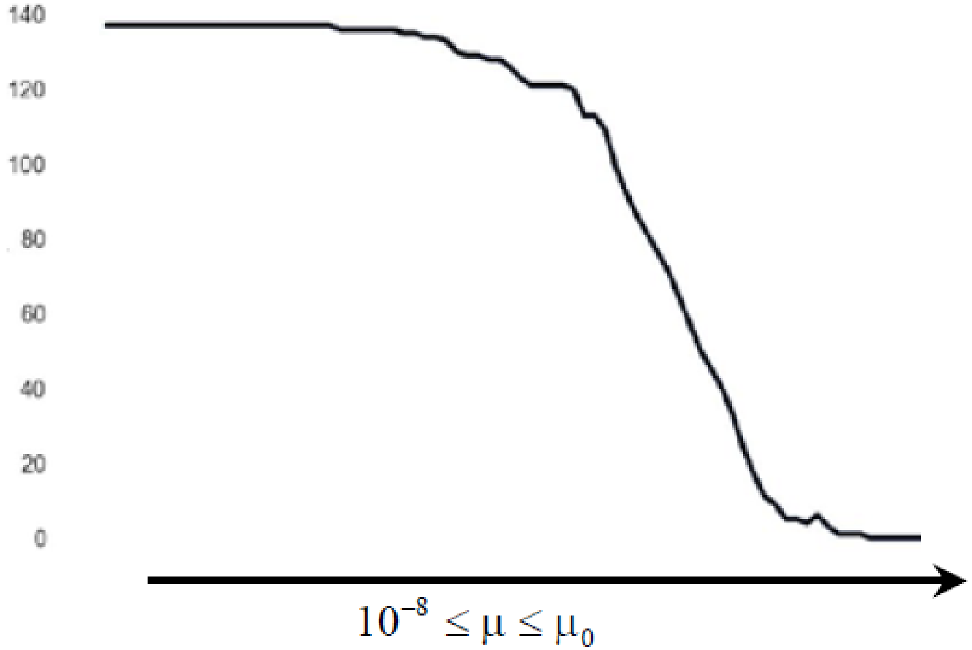
Each next value μ_k will almost coincide with the previous one μ_{k-1}, and the iteration process will converge at each step after one or two iterations.

The entire regularization path $0\le\mu\le\mu_0$ takes, as a rule, almost the same time as solving the dual problem for a single selectivity value μ.

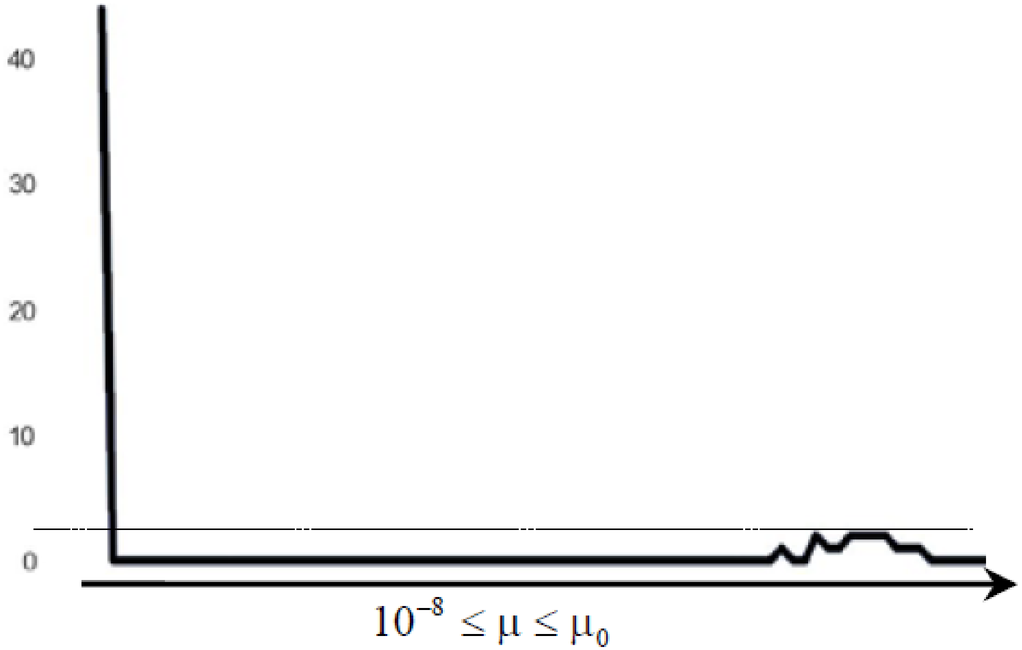
Regularization path along the selectivity axis

An experimental result. Regression estimation problem in a set of stock market data (Return-based analysis of an investment portfolio)

Number of observations $N = 240$
Number of features (known returns of stock market indexes)
The sought-for regression coefficients $n = 650$: capital sharing to be estimated



Number of active features



Number of iterations at each step

Conclusions

Under some quite lenient assumptions, the traditional formulation of the generalized linear dependence estimation problem results in the convex problem of regularized empirical risk minimization.

This problem inevitably has polynomial computational complexity in the number of features, what is in crucial conflict with the assumption on the huge dimension of the feature vectors.

We proposed an alternative disjoint formulation of the generalized linear dependence estimation problem, which is not only of linear computational complexity in the number of features, but also easily parallelizable.

Acknowledgement

We would like to acknowledge support from the Russian Foundation for Basic Research, projects 18-07-01087, 17-07-00993, 17-07-00436.

Thank you!

Questions?