My first scientific paper

Week 3 — State your problem

m1p.org

Significant increase in complexity and modest increase in accuracy

	train	test	out-of-time	# parameters
Logistic regression	53,08%	55,18%	57,50%	= 12
Neural network	59,85%	57,04%	58,27%	~ 240
Regression forest	61,85%	57,01%	59,61%	> 1000
Gradient boosting	**63,58%**	**58,31%**	**59,50%**	> 10,000

Model selection is an important problem!

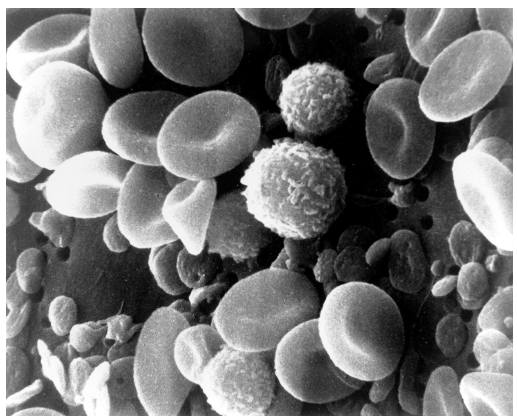
... it was a banking credit scoring model

Classification of patients in immunology

Collect the Cardio Immune Data for patients with Cardio-Vascular Disease.

Classes	$\longrightarrow$	Patient groups	Class lables "A1 (operated group) " и "A3 (risk group)".
Objects	$\longrightarrow$	Patients	Examined 14 patients in "A1" and 17 patients in"A3".
Features	$\longrightarrow$	Markers	20 biomarkers: K, L, K/M, L/M, K/N, K/O, L/O, K/P, L/P,...

- Quality criterion: number of misclassified patients
- Model: generalized linear
- Hypothesis: independed



https://en.wikipedia.org/wiki/Blood_cell

Design matrix, the patients and markers table, an example

Class	Patient	K	L	K/M	L/M	
A1	C001	58.3	16.7	0.52	0.00	
A1	C004	40.2	6.0	NaN	NaN	
A1	C005	54.3	13.1	NaN	NaN	
A1	C008	48.7	9.8	0.05	0.02	etc.
A3	023	46.6	21.2	0.40	0.08	
A3	026	50.7	26.2	0.12	0.00	
A3	027	45.3	24.5	0.05	0.02	
A3	D037	46.3	13.1	1.23	0.13	
				etc.		

Can we show that these groups are significantly different using these data?

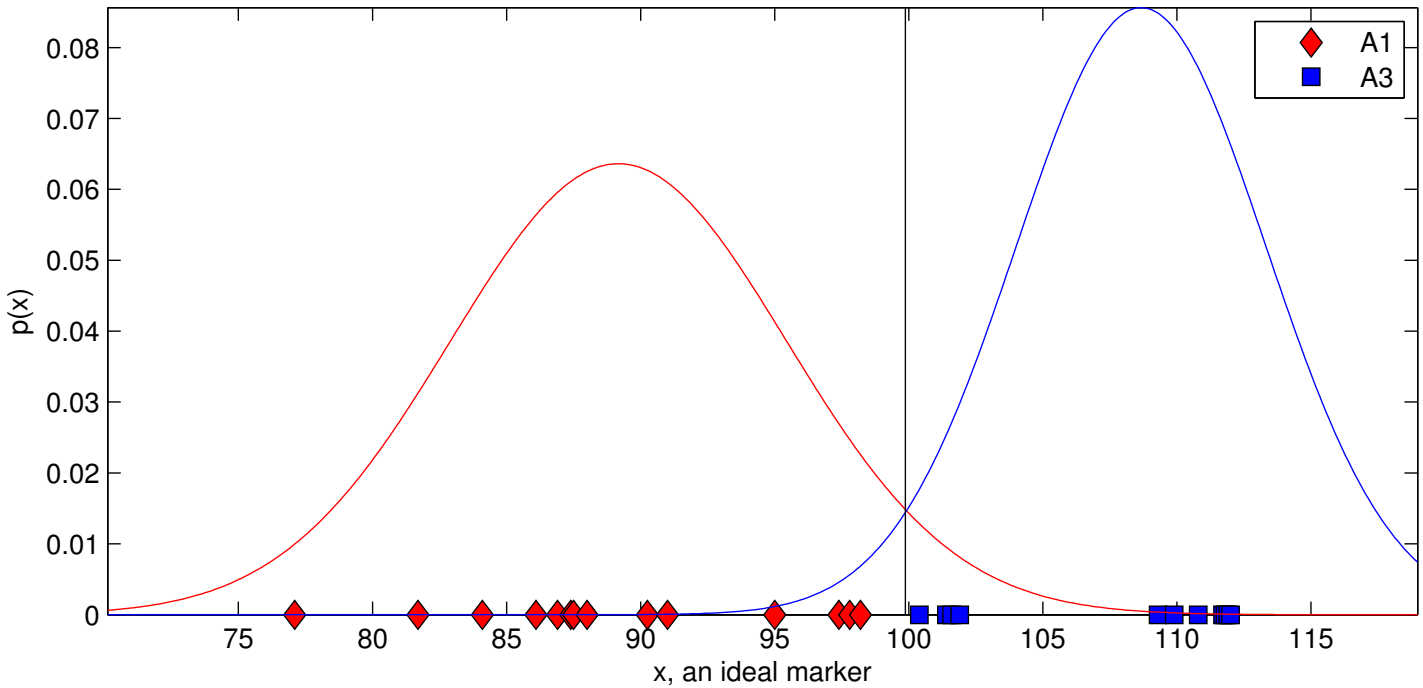
Problem statement

Consider a problem of classification.

1. We have two groups of patients.
2. Each group is labelled for a class.
3. Each patient is described by a set of markers.

We have to find a model which is based on measured data that shows the groups are significantly different.

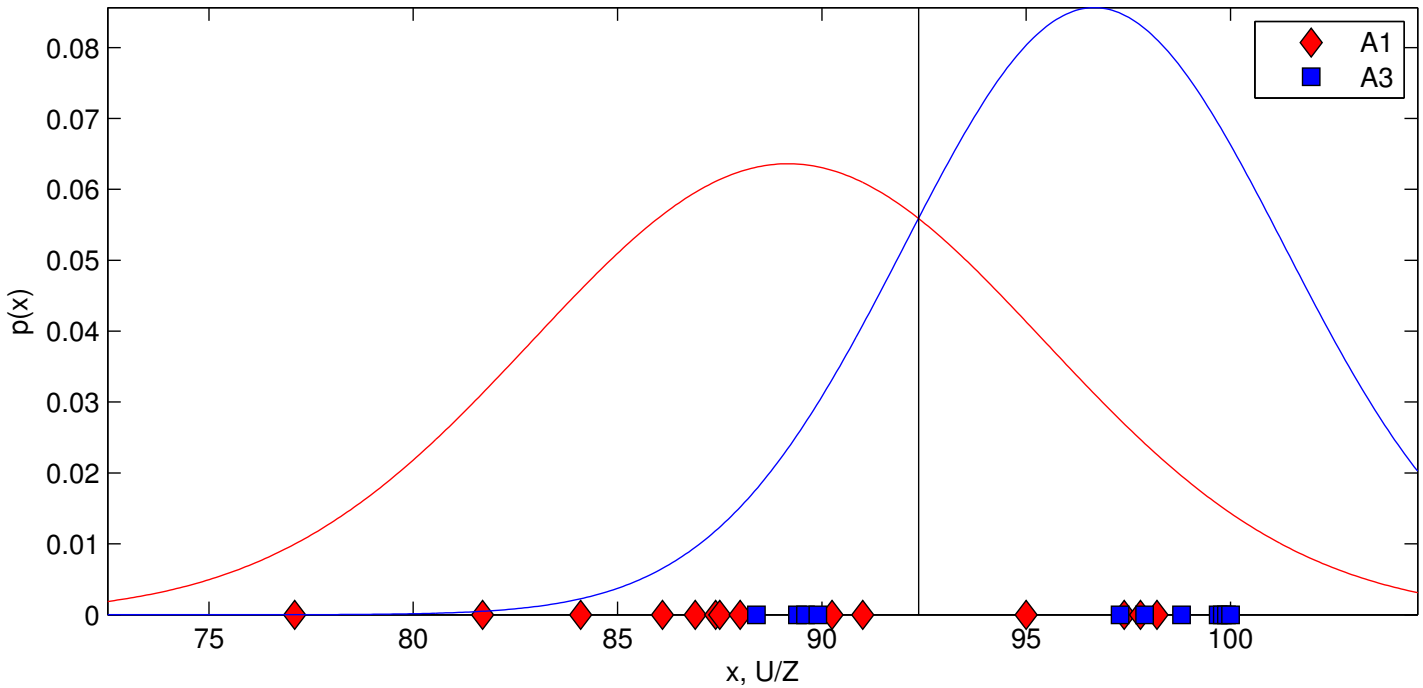
Bias of expert: one-dimensional statistics



Assume that two groups are separate if they accept the null-hypothesis in one of the statistical tests.

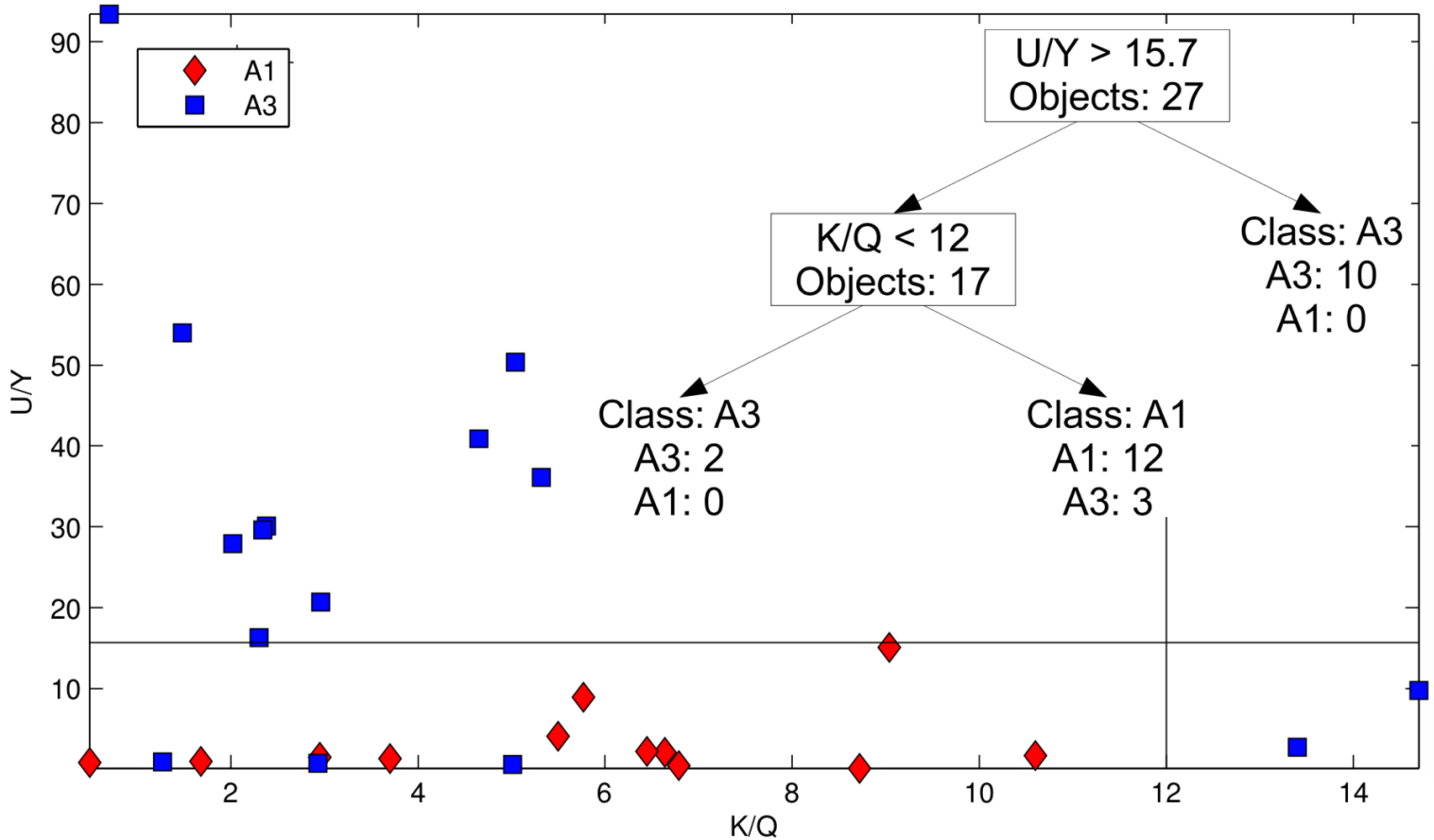
Namely, Student t-test, Welch t-test or Mann-Whitney U-test.

Real data: one-dim is impossible to use



- ✓ It is very simple to visualize one-dimensional data.
- ✓ One-dimensional statistics is well-developed and recognized.
- × And give poor results if one deals with a complex problem.

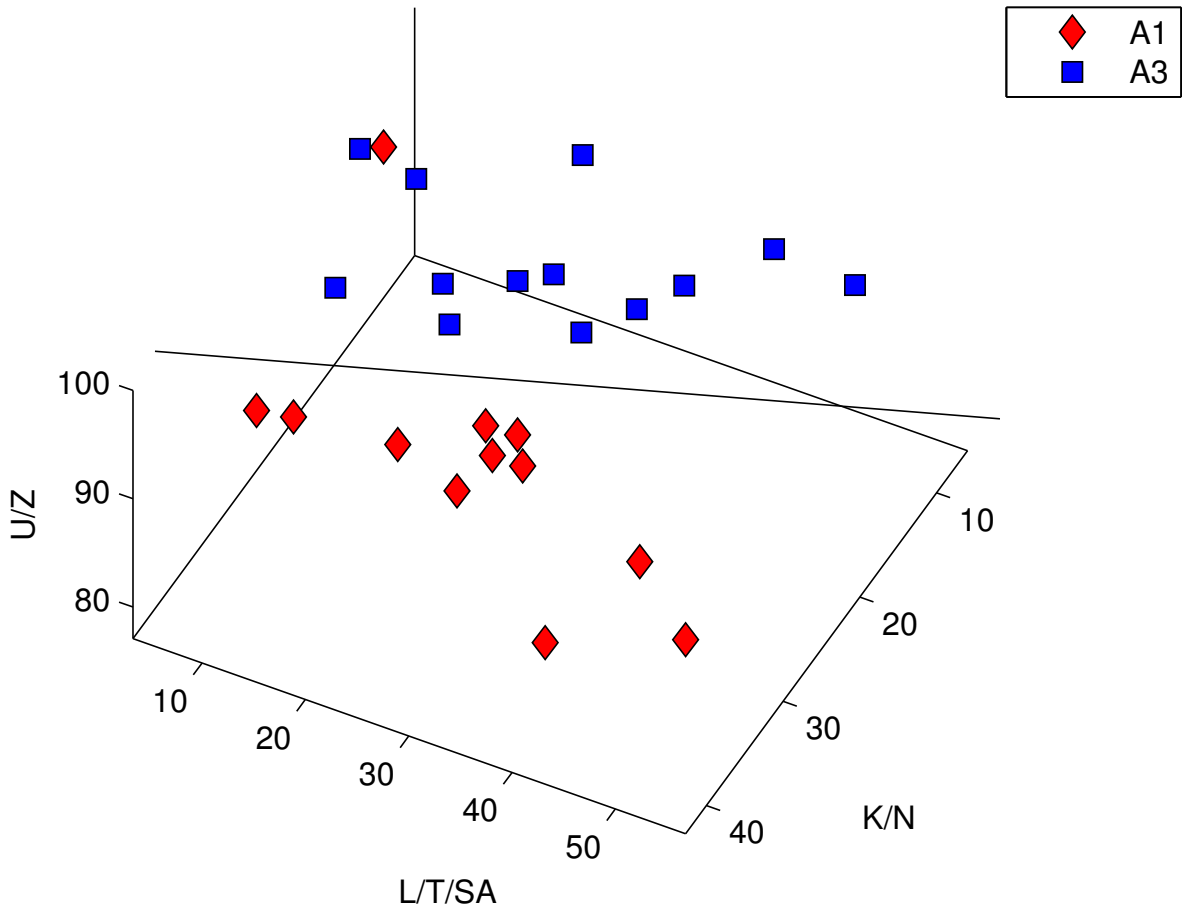
A classification rule in a decision tree



if U/Y $>$ 15.7 then A3 else (if K/Q $<$ 12 then A1 else A3)

- ✓ Different subsets of markers produce trees of different quality.
- ✓ One can use several trees to make a voting algorithm.

Linear classification model



The result of model selection (classify A1 versus A3) are the feature names and parameters of the machine learning model

$$\hat{y}_i = f(\mathbf{x}_i, \mathbf{w}) = \text{sign}(\mathbf{w}^\mathsf{T}\mathbf{x}_i - b) = \text{sign}([0.35, 0.72, 0.29]^\mathsf{T}\mathbf{x}_i - 34.16).$$

Each algorithm implies a hypothesis:

✓	markers do not depend on each other	$\longrightarrow$	decision trees,
✓	objects can be separated by hyperplane	$\longrightarrow$	linear classifier,
✓	objects can be mapped into a separable space	$\longrightarrow$	support vector machines,
✓	classes are compact	$\longrightarrow$	radial basis functions,
✓	marker space are complex	$\longrightarrow$	voting algorithms.

To interpret the results

we use the parametric algorithms. They are based on a mathematical model and a set of parameters.

Profs and coins

Decision tree

- ✓ uses several markers for classification; so, it makes less mistakes,
- ✓ helps to select the most informative markers,
- ✓ can be used in a voting algorithm,
- × assumes the markers do not depend on each other.

Linearity

However in this problem we are using the markers do depend on each other. For example, three markers K, N and K/N are linear dependent, since we assume

$$K/N = \alpha K + (1 - \alpha)N.$$

Problem statement for machine learning

Formal problem statement, an analyst has to set

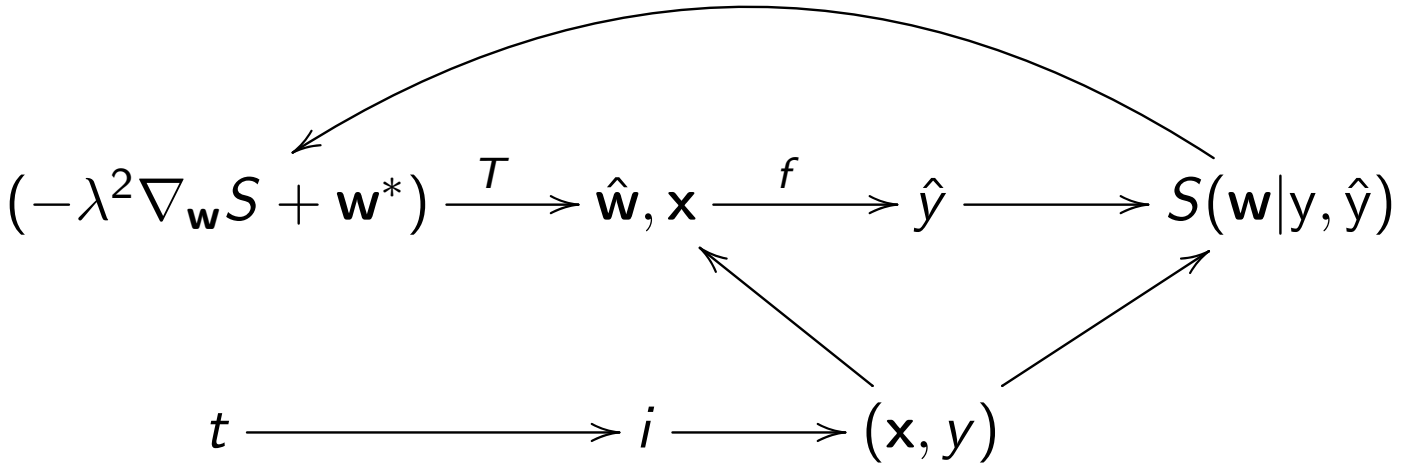
1) an algebraic structure for the dataset from measurements
2) a data generation hypothesis from 1)
3) a model, or a mixture from 2)
4) an error function (quality criteria with restrictions) from 2)
5) an optimization algorithm from 3) and 4)

The result of the model construction is a Cartesian product

$$\{\textbf{models} \times \textbf{data sets} \times \textbf{quality critea}\}.$$

Def: *Big data rejects the i.i.d. (independent and identically distributed random variables) data generation hypothesis from 2). It requests a mixture model.*

The simplest problem statement in machine learning



f is the forecasting model,

S is the criterion,

T is an optimization algorithm,

$\hat{\mathbf{w}}$ is some solution,

$$\hat{\mathbf{w}} = \arg\min S(\mathbf{w}|y, f).$$

[1] These notations are equivalent: $x_i, x(i), i \to x$.

[1] $*$

Определение 1. *Шкала $\mathbb{L}$ — алгебраическая структура* [19] *с заданным набором операций и отношений, удовлетворяющая фиксированному набору аксиом.*

Определение 2. *Номинальная шкала $\mathbb{C}$ — шкала с заданным на ней бинарным отношением равенства:*

1. $x = y \vee x \neq y$;
2. $x, y : x = y \Rightarrow y = x$;
3. $x, y, z : x = y \wedge y = z \Rightarrow x = z$,

где x, y, z — объекты, представленные в шкале $\mathbb{C}$: $x, y, z \in \mathbb{C}$.

Определение 3. *Порядковая шкала $\mathbb{O}$ — номинальная шкала с заданным на ней бинарным отношением R, для которого выполнены следующие свойства:*

1. xRx,
2. $xRy \wedge yRx \Rightarrow x = y$;
3. $xRy \wedge yRz \Rightarrow xRz$;

где $x, y, z \in \mathbb{O}$.

Определение 4. *Линейная шкала $\mathbb{W}$ — порядковая шкала с отношением полного порядка и определенными операциями сложения и вычитания.*

Условия Гаусса-Маркова

Нахождение параметров $\mathbf{w}$ линейной модели при предположении о нормальном распределении зависимой переменной $\mathbf{y}$ заключается в минимизации евклидовой нормы вектора регрессионных остатков

$$S(\mathbf{w}) = \|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2 = \|\boldsymbol{\varepsilon}\|^2.$$

Предполагается выполнение следующих условий:

1) независимые переменные $\mathbf{x}$ не являются случайными величинами,

2) математическое ожидание $\mathsf{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$,

3) дисперсия $\mathsf{D}(\boldsymbol{\varepsilon}) = \sigma_{\boldsymbol{\varepsilon}}^2\mathbf{I}$ (условие гомоскедастичности),

4) при $i \neq k$ математическое ожидание $\mathsf{E}(\varepsilon_i, \varepsilon_k) = 0$,

5) $\operatorname{rank}(\mathbf{X}) = n \leqslant m$.

Некторые задачи машинного обучения

- Задача оценки параметров модели,
- задача выбора признаков или объектов выборки,
- задача выбора модели оптимальной сложности,
- задача построения и выбора структуры модели,
- задача проверки гипотезы порождения данных.

Предполагается, что функция ошибки $S(\mathrm{w}|D, f)$ задана исходя из

- гипотезы порождения данных,
- либо из практических соображений.

Задача нахождения наиболее правдоподобных параметров

Задана выборка $D = \{(\mathrm{x}_i, y_i)\}$, $i \in \mathcal{I}$, функция ошибки модели S и модель — параметрическое семейство функций $f(\mathrm{w}, \mathrm{x})$. Требуется найти такие параметры w модели, которые бы доставляли минимум функции ошибки

$$\mathrm{w}^* = \arg\min_{\mathrm{w}\in\mathbb{W}} S(\mathrm{w}|D, f). \tag{1}$$

Функция ошибки, определенная посредством логарифмической функции правдоподобия

$$S(\mathrm{w}) = -\ln\big(p(D|\mathrm{w}, f)\big),$$

обеспечивает максимизацию правдоподобия параметров. Параметры, найденные минимизацией этой функции ошибок, будут называться наиболее правдоподобными.

Примеры функции ошибки в регрессии и классификации

Регрессия

Гипотеза порождения данных: $\mathsf{y} \sim \mathcal{N}(\mathsf{f}, \mathsf{I})$.
Функция ошибки:

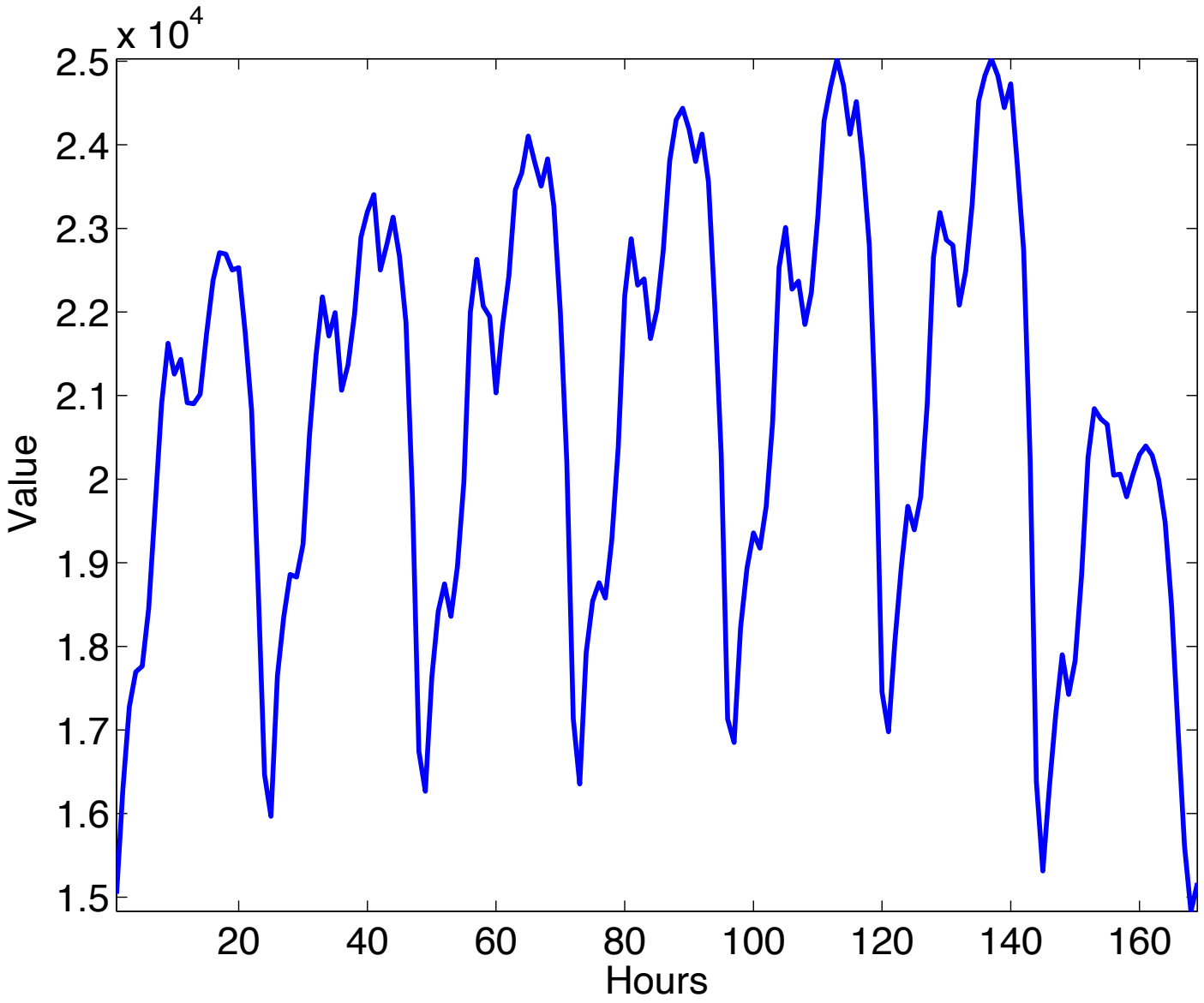
$$S(\mathsf{w}) = \|\mathsf{y} - \mathsf{f}\|_2^2.$$

Классификация

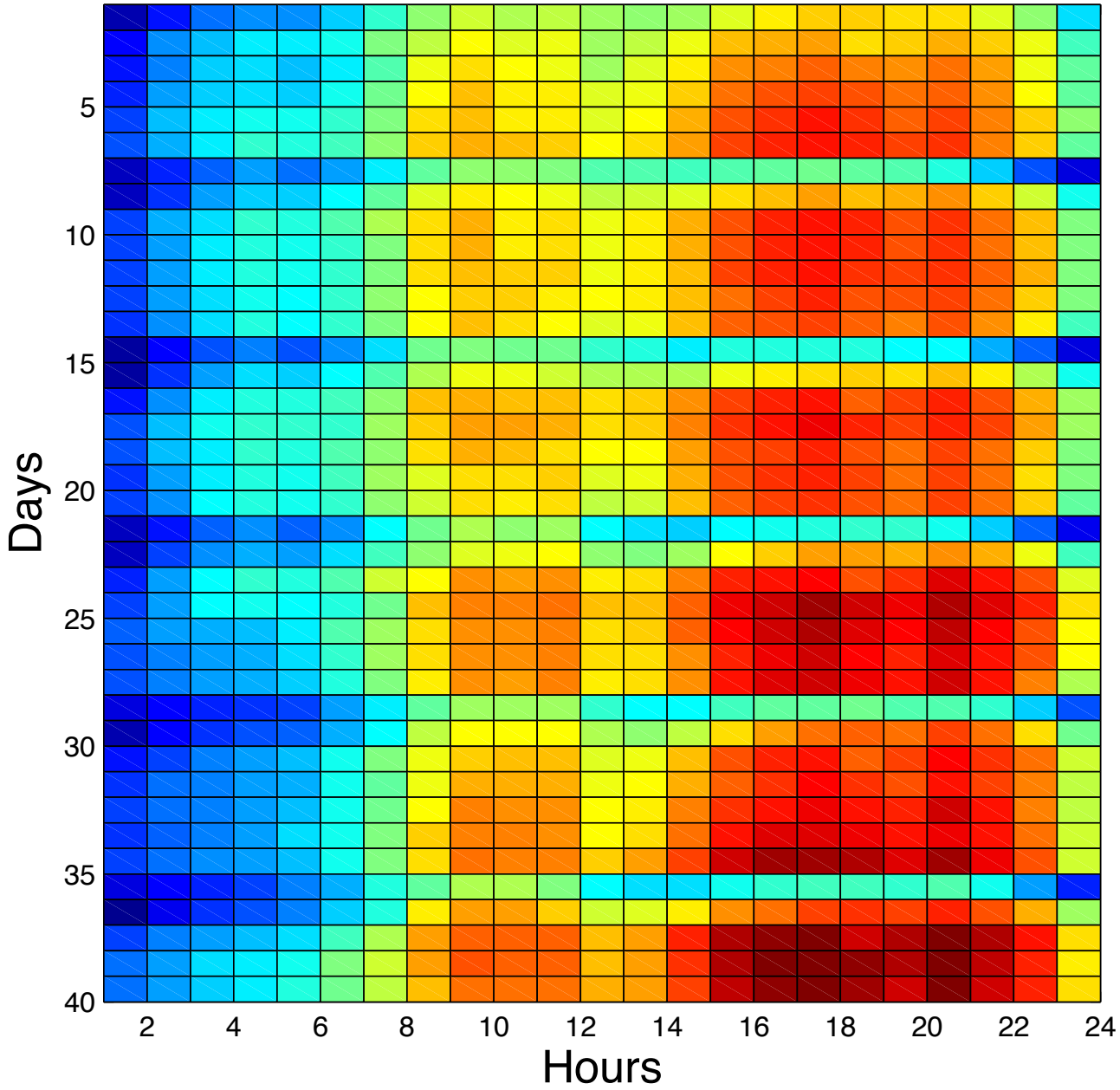
Гипотеза порождения данных: $\mathsf{y} \sim \mathcal{B}(f, 1 - f)$.
Функция ошибки:

$$S(\mathsf{w}) = \sum_{i \in \mathcal{I}} y_i \ln f(\mathsf{w}^\mathsf{T}\mathsf{x})_i + (1 - y_i) \ln\bigl(1 - f(\mathsf{w}^\mathsf{T}\mathsf{x})_i\bigr).$$

Source time series, one week



The autoregressive matrix, five week-ends



The autoregressive matrix and the linear model

$$\underset{(m+1)\times(n+1)}{X^*} = \left(\begin{array}{c|ccc} s_T & s_{T-1} & \cdots & s_{T-\kappa+1} \\ \hline s_{(m-1)\kappa} & s_{(m-1)\kappa-1} & \cdots & s_{(m-2)\kappa+1} \\ \cdots & \cdots & \cdots & \cdots \\ s_{n\kappa} & s_{n\kappa-1} & \cdots & s_{n(\kappa-1)+1} \\ \cdots & \cdots & \cdots & \cdots \\ s_{\kappa} & s_{\kappa-1} & \cdots & s_1 \end{array}\right).$$

In a nutshell,

$$X^* = \left[\begin{array}{c|c} \underset{1\times 1}{s_T} & \underset{1\times n}{\mathbf{x}_{m+1}} \\ \hline \underset{m\times 1}{\mathbf{y}} & \underset{m\times n}{X} \end{array}\right].$$

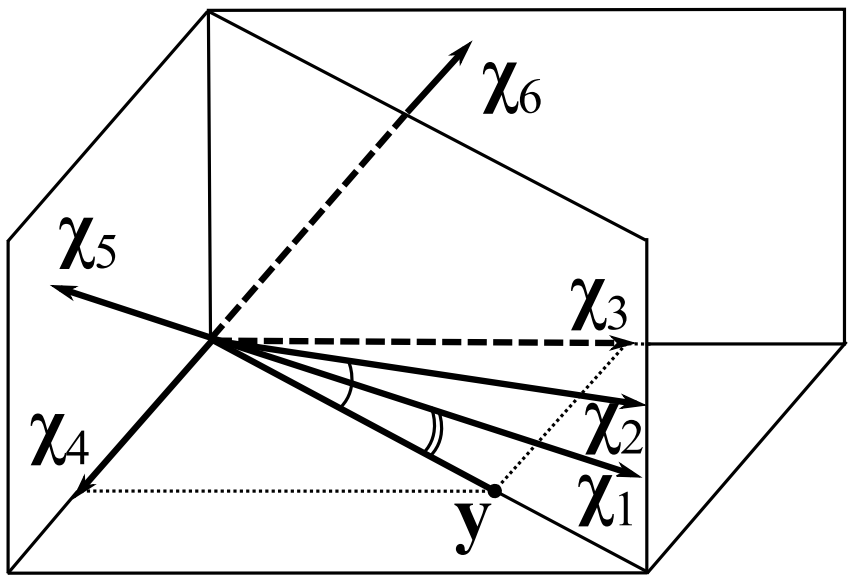
In terms of linear regression:

$$\mathbf{y} = X\mathbf{w},$$

$$y_{m+1} = s_T = \mathbf{w}^{\mathsf{T}}\mathbf{x}_{m+1}^{\mathsf{T}}.$$

Выбор устойчивой и точной модели

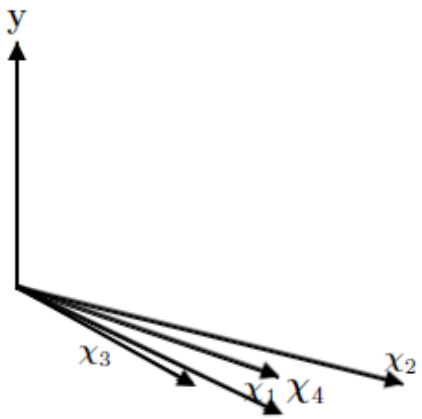
Выборка содержит мультикоррелирующие χ_1, χ_2 и устойчивые χ_5, χ_6 признаки — столбцы матрицы «объект-признак» $\mathbf{X}$. Требуется выбрать два признака из шести.



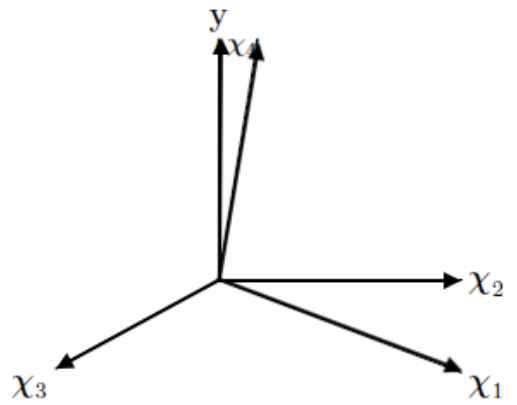
Точность и устойчивость при заданной сложности

Решение: χ_3, χ_4 — набор ортогональных признаков с наименьшим значением функции ошибки.

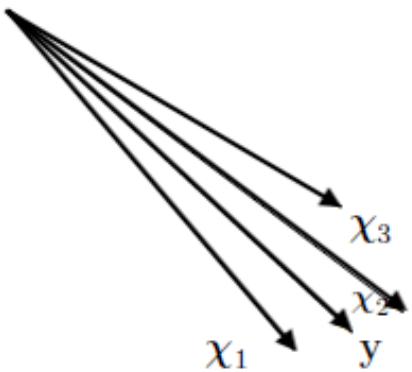
Configurations of design space



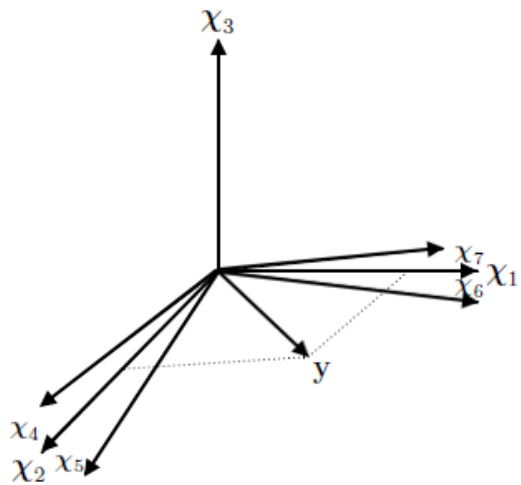
Non-adequate correlated



Adequate random



Adequate redundant



Adequate correlated

Katrutsa, Strijov. 2017. Comprehensive study of feature selection methods to solve multicollinearity problem // Expert Systems with Applications

Задача выбора оптимального набора признаков

- Задана выборка $D = \{(\mathrm{x}_i, y_i)\}$, $i \in \mathcal{I}$.
- Задано случайное разбиение множество индексов элементов выборки $\mathcal{I} = \mathcal{L} \sqcup \mathcal{C}$.
- Множество независимых переменных $\mathrm{x} = [x_1, \ldots, x_j, \ldots, x_n]$ проиндексировано $j \in \mathcal{J} = \{1, \ldots, n\}$.
- Задано множество моделей-претендентов $\mathfrak{F} = \{f(\mathrm{w}, \mathrm{x})\}$.
- Модель — параметрическое семейство функций $f(\mathrm{w}, \mathrm{x}) = \mu(\mathrm{w}^\mathsf{T}\mathrm{x})$, где μ — функция связи (в случае регрессии $\mu = \mathrm{id}$, в случае классификации $\mu = \frac{1}{1+\exp(-\mathrm{w}^\mathsf{T}\mathrm{x})}$.
- Структура модели $f_\mathcal{A}$ задана множеством индексов $\mathcal{A} \subseteq \mathcal{J}$ и означает включение переменных $\mathrm{x}_\mathcal{A}$. Иначе, используются только признаки-столбцы матрицы X с индексами из множества $\mathcal{A}$.
- Задана функция ошибки S.

Задача выбора оптимального набора признаков

Требуется найти такое подмножество индексов $\mathcal{A} \subseteq \mathcal{J}$, которое бы доставляло минимум функции

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

на разбиении выборки D, определенном множеством индексов $\mathcal{C}$.

При этом параметры w^* модели должны доставлять минимум функции

$$\mathrm{w}^* = \arg\min_{\mathrm{w} \in \mathbb{W}} S(\mathrm{w}|D_{\mathcal{L}}, f_{\mathcal{A}})$$

на разбиении выборки, определенном множеством $\mathcal{L}$.

Analyst creates an optimal model for expert to put it to operation

Quality criteria

- Accuracy: MAPE, AUC, F1 score
- Stability: forecasting variance, failure rate, parameter variance
- Complexity: number of parameters, Kolmogorov complexity

Origins of quality criteria

1. Theory: statistical hypotheses of data generation, algebraic structures of data, models of measurement
2. Computations: a criterion is useful to an optimisation procedure
3. Deployment: revenue, loss, failure rate