



Л е к ц и я № 7

Метод k-ближайших соседей

Разработал профессор П. М. Васильев
Кафедра фармакологии и биоинформатики

Для студентов, обучающихся по направлению 06.03.01 «Биология»
профили Биохимия, Генетика
при изучении дисциплины «Цифровые технологии в биологии»

П л а н л е к ц и и

- Что такое метод k-ближайших соседей
- Кусочно-линейные разделяющие функции
- Алгоритм k-NN
- Метрики расстояния
- Метод k-NN в программе Statistica

Метод k-ближайших соседей

k- Nearest Neighbors, k-NN

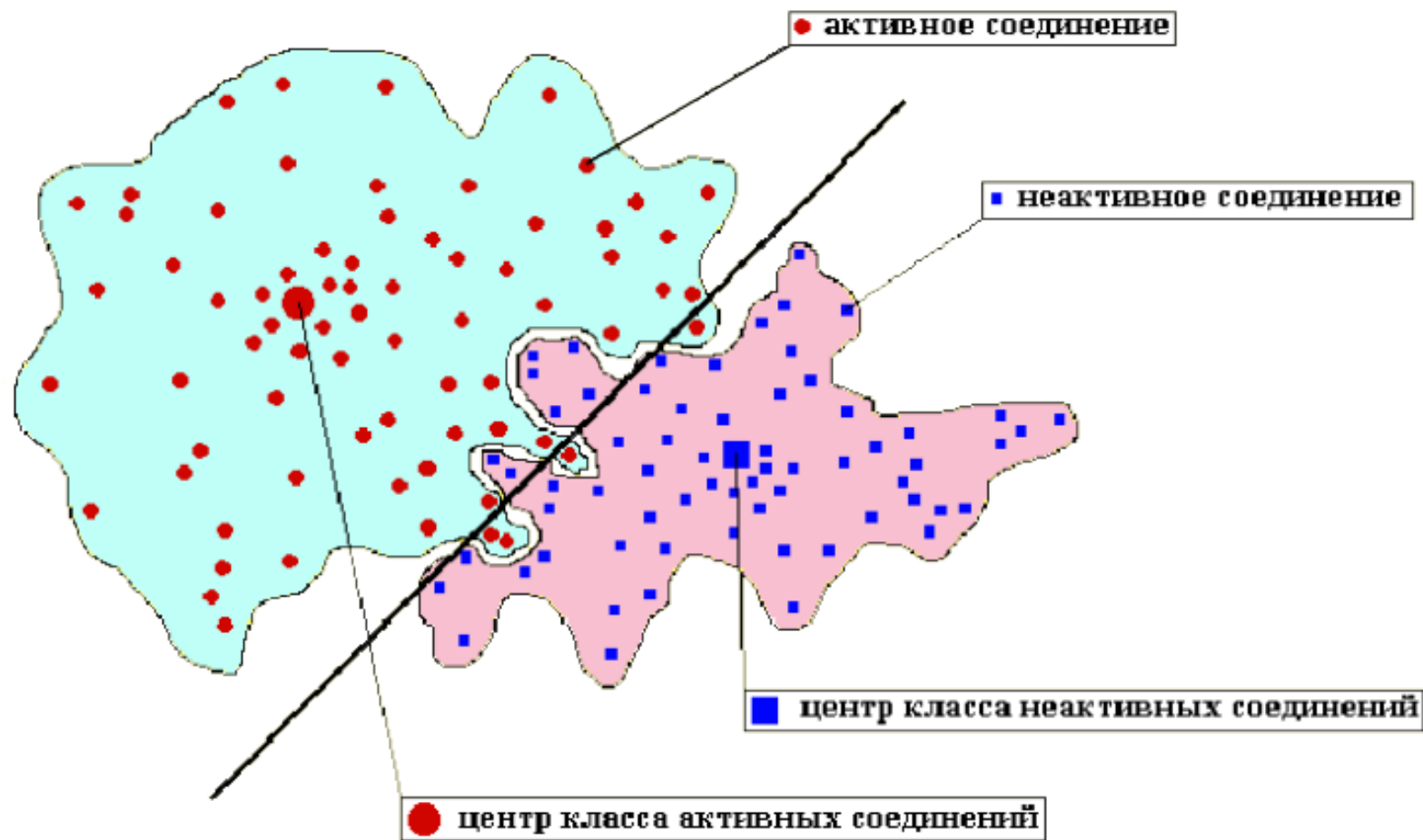
Основан на гипотезе компактности и предполагает, что расположенные близко друг к другу объекты в пространстве признаков принадлежат к одному классу или имеют схожие значения целевой переменной.

Гипотеза компактности

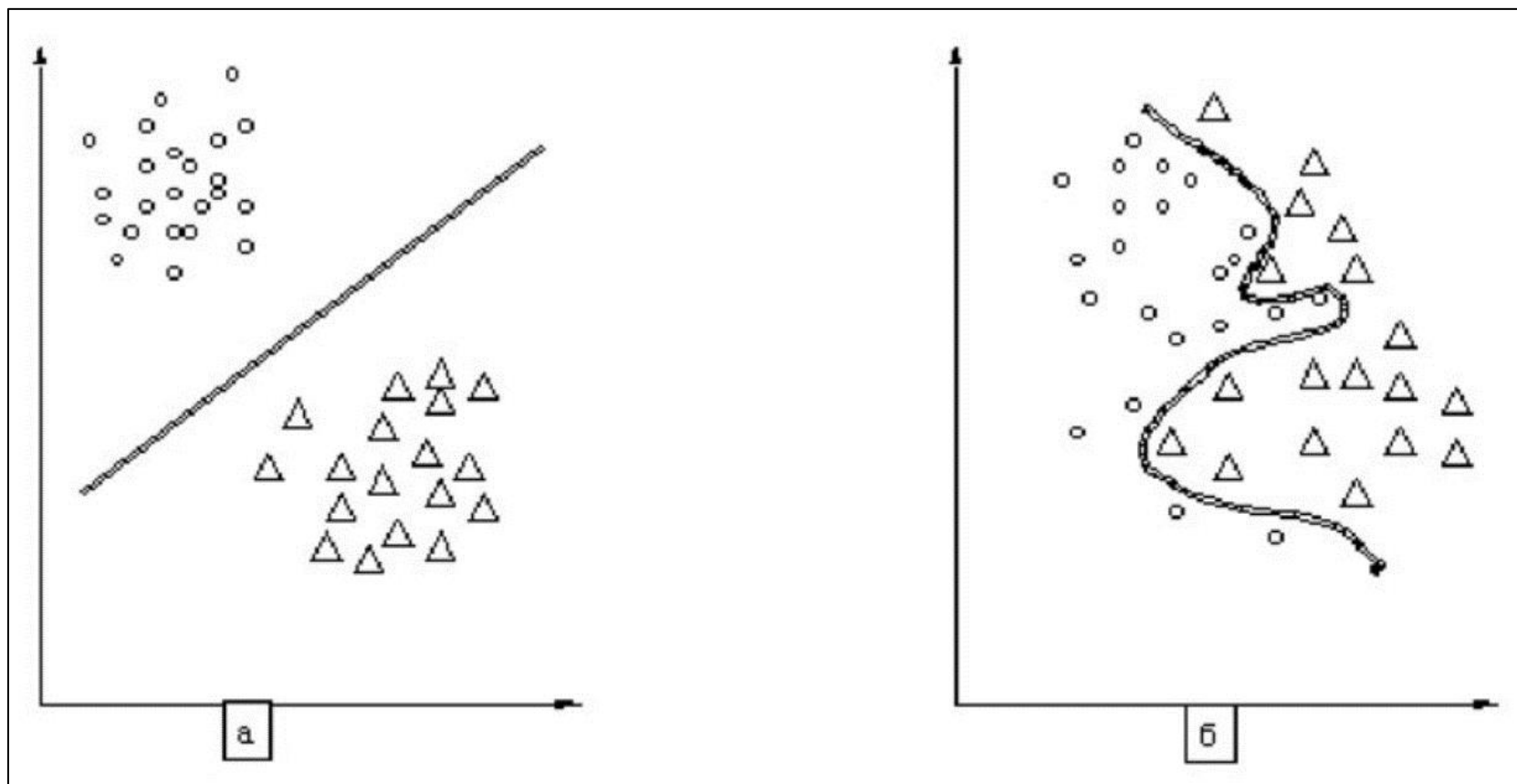
Расстояние между двумя объектами, принадлежащими к одному и тому же кластеру, всегда меньше расстояния между двумя объектами, принадлежащими к разным кластерам.

Задача классификации

Обучение с учителем



Разделяющие функции

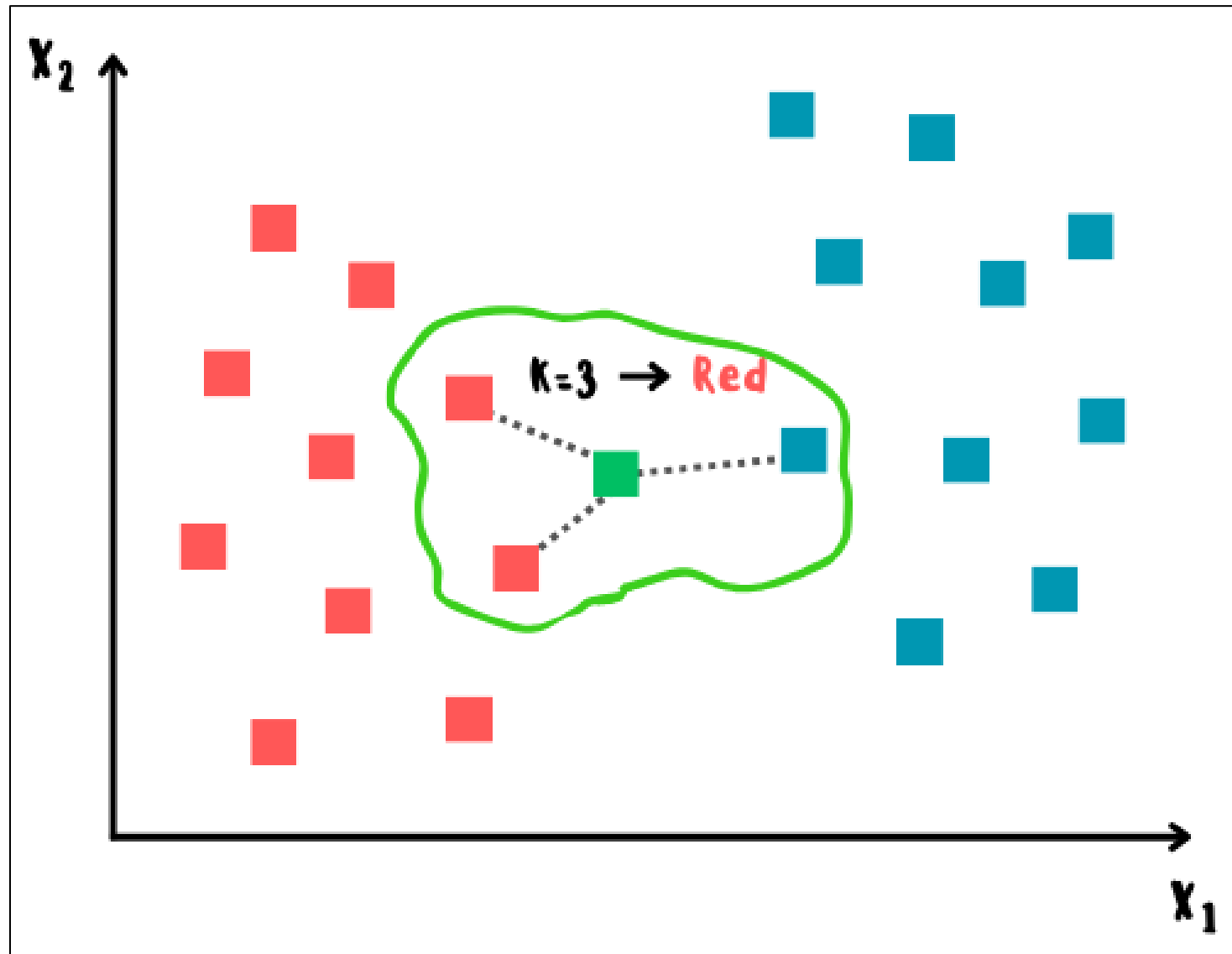


Линейная
Кусочно-линейная

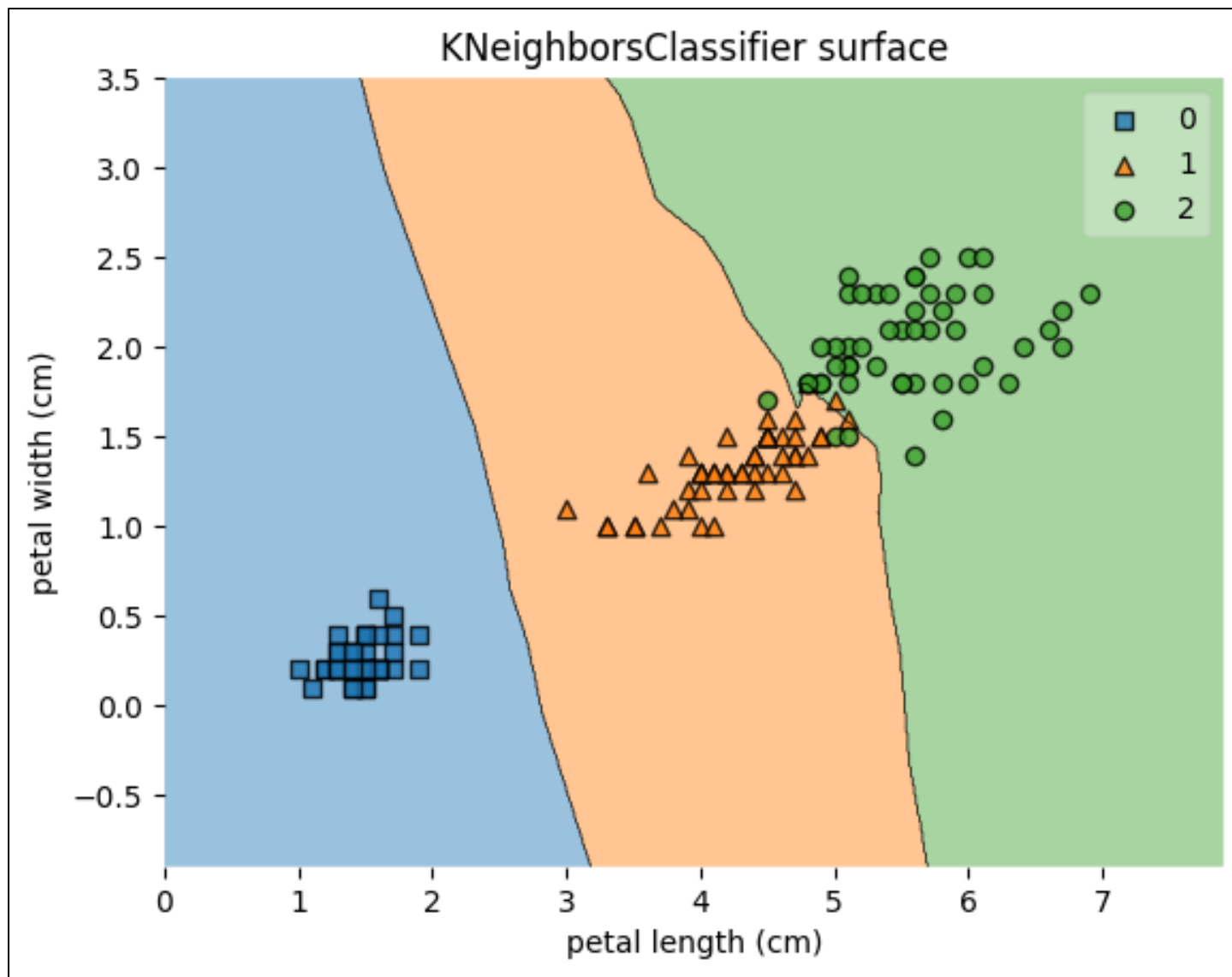
Алгоритм k-NN

- Задают число соседей k
- Вычисляют расстояние между прогнозируемым объектом и всеми обучающими объектами
- Выбирают k -ближайших обучающих объектов (соседей)
- Относят прогнозируемый объект к классу с максимальным числом ближайших обучающих объектов

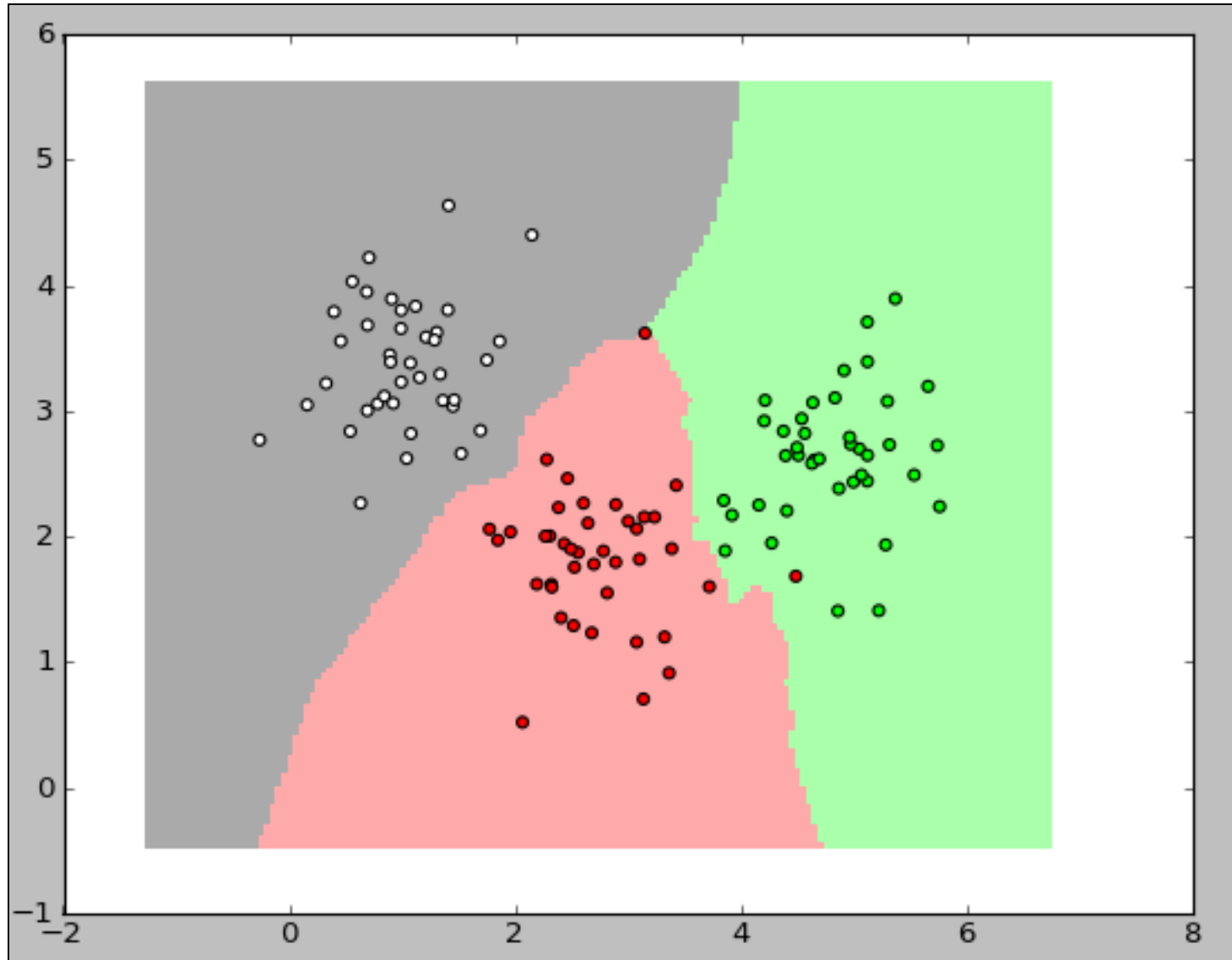
Принцип k-NN



k-NN разделяющие функции



k-NN разделяющие функции



Метрики расстояния

Эвклидово расстояние

$$d_{ij} = \sqrt{\sum_{k=1}^p (x_{ik} - x_{jk})^2}$$

Требует обязательного нормирования показателей

Метрики расстояния

Расстояние городских кварталов

$$d_{ij} = \sum_{k=1}^p |x_{ik} - x_{jk}|$$

Менее чувствительно к выбросам

Метрики расстояния

Расстояние Чебышева

$$d_{ij} = \max_k |x_{ik} - x_{jk}|$$

Хорошо учитывает различия объектов

Метрики расстояния

Расстояние Минковского

$$d_{ij} = \left(\sum_{k=1}^p |x_{ik} - x_{jk}|^p \right)^{1/r}$$

Взвешивает расстояние между переменными

Метод k-ближайших соседей

Пример исходных данных

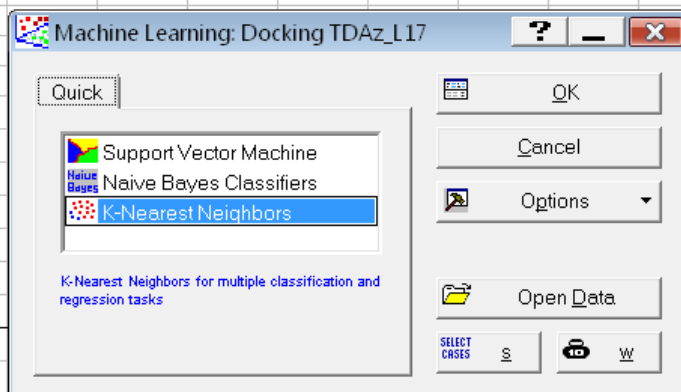
Code	Y3	Y2	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13
LT-47	high	h	-8.5	-6.5	-7.9	-7.1	-8.2	-8.7	-8.3	-6.2	-6.7	-7.5	-8.5	-6.4	-6.3
LT-51	high	h	-8.4	-6.7	-7.4	-6.6	-8.1	-8.6	-8.3	-6.1	-6.5	-7.4	-8.4	-6.8	-6.1
LT-52	high	h	-8.6	-7.1	-8.2	-7.0	-9.4	-9.2	-8.6	-6.4	-6.8	-8.2	-8.6	-7.5	-6.4
L-91	high	h	-8.6	-6.9	-6.9	-8.3	-9.2	-8.7	-8.9	-6.1	-6.4	-7.5	-8.2	-7.0	-6.3
LT-13	high	h	-8.6	-7.2	-8.2	-7.2	-9.0	-8.9	-8.7	-6.3	-6.5	-7.5	-8.2	-6.4	-5.8
LT-23	high	h	-8.5	-6.8	-7.0	-7.7	-8.8	-8.2	-8.4	-6.1	-6.4	-7.6	-8.1	-7.0	-6.0
LT-48	high	h	-8.5	-6.4	-7.4	-6.8	-8.4	-8.6	-8.2	-6.2	-6.6	-7.6	-8.0	-6.6	-6.1
LT-46	high	h	-8.2	-6.2	-7.8	-7.8	-7.9	-8.3	-7.9	-6.0	-6.4	-7.4	-8.2	-6.2	-6.0
H-16	high	h	-8.3	-6.5	-7.1	-6.6	-8.0	-8.1	-8.1	-5.7	-6.1	-7.3	-8.0	-6.6	-6.4
H-88	high	h	-8.6	-7.0	-7.3	-8.0	-9.3	-8.4	-8.0	-6.1	-6.4	-7.2	-7.9	-6.8	-6.3
LT-22	high	h	-8.5	-6.9	-7.4	-8.3	-9.0	-8.3	-8.4	-6.0	-6.3	-7.5	-8.2	-6.8	-6.5
LT-49	high	h	-9.0	-6.6	-7.0	-6.8	-8.4	-8.5	-8.4	-6.0	-6.3	-7.6	-8.0	-6.3	-6.3
L-87	moderate	nh	-8.5	-7.0	-7.5	-7.8	-8.6	-8.5	-8.9	-6.2	-6.4	-7.3	-8.2	-6.9	-6.3
LT-55	moderate	nh	-8.6	-7.0	-7.5	-7.3	-8.5	-8.3	-9.1	-6.1	-6.4	-7.8	-8.1	-6.7	-6.0
LT-9	moderate	nh	-8.6	-6.8	-7.4	-8.0	-8.4	-8.2	-8.4	-6.1	-6.4	-6.9	-7.7	-6.3	-6.0
L-20	moderate	nh	-8.4	-6.6	-7.1	-7.3	-8.2	-8.3	-8.1	-6.0	-6.3	-7.4	-8.1	-6.6	-6.1
L-31	moderate	nh	-8.4	-6.9	-7.4	-7.8	-8.0	-8.4	-8.1	-6.2	-6.6	-7.4	-8.0	-6.4	-6.6
L-86	moderate	nh	-8.3	-6.8	-6.8	-8.5	-8.9	-8.2	-8.6	-6.1	-6.2	-7.2	-8.0	-6.9	-6.2
LT-21	moderate	nh	-8.6	-6.8	-7.1	-8.2	-8.8	-8.2	-8.1	-6.0	-6.2	-7.3	-8.0	-6.9	-6.1
LTG-2	moderate	nh	-7.8	-6.8	-6.9	-7.3	-7.9	-8.0	-7.4	-6.2	-6.5	-8.1	-8.4	-6.9	-6.8
L-19	moderate	nh	-8.3	-6.6	-7.1	-6.7	-8.3	-8.1	-7.9	-6.2	-6.4	-7.6	-7.9	-6.6	-6.1
LT-24	moderate	nh	-8.8	-7.1	-6.8	-7.9	-8.8	-8.5	-8.2	-5.9	-6.3	-7.4	-7.9	-6.5	-6.4
LT-33	moderate	nh	-8.0	-6.1	-7.2	-7.0	-7.9	-7.9	-7.7	-5.8	-5.9	-7.1	-8.0	-6.1	-5.5
L-17	moderate	nh	-8.0	-6.6	-7.4	-8.3	-8.1	-7.9	-7.8	-6.1	-6.3	-7.4	-8.0	-6.5	-6.0
LT-34	moderate	nh	-8.2	-6.3	-7.6	-6.5	-8.3	-8.1	-7.9	-6.0	-6.3	-7.1	-8.1	-6.3	-6.1
LT-36	moderate	nh	-8.2	-6.3	-7.1	-6.2	-7.8	-8.1	-7.7	-5.8	-6.0	-7.0	-8.1	-6.7	-5.9
LT-50	moderate	nh	-8.7	-6.6	-7.4	-6.8	-8.3	-8.5	-7.9	-6.2	-6.4	-7.5	-7.9	-5.9	-6.4
LT-5	moderate	nh	-8.5	-6.5	-7.1	-8.0	-8.1	-7.9	-8.0	-5.8	-6.2	-7.0	-7.5	-6.1	-5.6
H-18	moderate	nh	-8.6	-6.7	-6.8	-6.6	-8.2	-8.0	-8.2	-6.1	-6.3	-7.4	-7.9	-6.3	-6.4
H-69	low	nh	-8.2	-6.5	-7.1	-6.5	-8.0	-8.1	-8.1	-5.9	-6.2	-7.4	-7.8	-6.3	-6.4
LT-30	low	nh	-7.6	-6.1	-6.5	-7.2	-7.8	-7.2	-7.5	-5.6	-6.0	-7.0	-7.5	-6.0	-6.1
LT-31	low	nh	-7.8	-6.1	-6.1	-5.9	-7.8	-7.2	-7.3	-5.6	-5.9	-7.1	-7.6	-6.1	-5.7
LT-35	low	nh	-8.1	-6.2	-7.0	-6.5	-8.5	-8.2	-8.0	-6.0	-6.3	-7.0	-7.8	-6.6	-5.9
LT-39	low	nh	-8.2	-6.2	-6.7	-6.3	-8.4	-8.1	-8.2	-5.7	-6.1	-7.3	-7.6	-6.2	-5.8
LT-29	low	nh	-7.3	-5.9	-6.3	-6.8	-7.7	-6.8	-7.2	-5.6	-5.8	-7.1	-7.2	-6.0	-5.8
LT-32	low	nh	-7.7	-6.2	-6.0	-6.7	-7.6	-7.3	-7.6	-5.3	-5.7	-7.0	-7.4	-6.1	-5.8
LT-38	low	nh	-9.3	-6.4	-7.1	-6.0	-8.2	-8.2	-8.1	-5.9	-6.2	-7.2	-7.6	-6.3	-5.8
LT-37	low	nh	-9.3	-6.3	-7.0	-6.2	-8.3	-8.1	-8.1	-5.9	-6.2	-7.1	-7.7	-6.5	-5.7
H-42	low	nh	-7.4	-6.0	-6.2	-6.4	-7.5	-7.3	-7.0	-5.4	-5.6	-7.0	-7.3	-6.3	-5.8
LT-53	low	nh	-7.5	-6.1	-6.3	-6.8	-7.4	-7.4	-7.2	-5.6	-5.9	-7.0	-7.4	-6.3	-5.9
H-29	low	nh	-7.1	-5.5	-6.2	-7.6	-7.1	-7.1	-6.7	-5.3	-5.6	-6.6	-6.8	-5.9	-5.5

Метод k-ближайших соседей

STATISTICA - [Data: Docking TDAz_L17* (15v by 43c)]														
File Edit View Insert Format Statistics Data Mining Graphs Tools Data Window Help														
Automated Neural Networks														
Machine Learning (Bayesian, Support Vectors, Nearest Neighbor)														
General Classification/Regression Tree Models														
General CHAID Models														
Interactive Trees (C&RT, CHAID)														
Boosted Tree Classifiers and Regression														
Random Forests for Regression and Classification														
Generalized Additive Models														
MARSplines (Multivariate Adaptive Regression Splines)														
Generalized EM & k-Means Cluster Analysis														
Independent Components Analysis														
Text & Document Mining, Web Crawling - Methods														
Association Rules														
Sequence, Association, and Link Analysis														
Rapid Deployment of Predictive Models (PMML)														
Goodness of Fit, Classification, Prediction														
Feature Selection and Variable Screening														
Combining Groups (Classes) for Predictive Data Mining														
Data Mining - Workspaces														
QC Data Mining & Root Cause Analysis - Methods														
LDA														
	1	2	3											
	Y3	Y2	X1											
LT-47	high	h	-8.5											
LT-51	high	h	-8.4											
LT-52	high	h	-8.6											
L-91	high	h	-8.6											
LT-13	high	h	-8.6											
LT-23	high	h	-8.5											
LT-48	high	h	-8.5											
LT-46	high	h	-8.2											
H-16	high	h	-8.3											
H-88	high	h	-8.6											
LT-22	high	h	-8.5											
LT-49	high	h	-9.0											
L-87	moderate	nh	-8.5											
LT-55	moderate	nh	-8.6											
LT-9	moderate	nh	-8.6											
L-20	moderate	nh	-8.4											
L-31	moderate	nh	-8.4											
L-86	moderate	nh	-8.3											
LT-21	moderate	nh	-8.6	-6.8	-7.1	-8.2	-8.8	-8.2	-8.1	-6.0	-6.2	-7.3	-8.0	-6.9
LTG-2	moderate	nh	-7.8	-6.8	-6.9	-7.3	-7.9	-8.0	-7.4	-6.2	-6.5	-8.1	-8.4	-6.9
L-19	moderate	nh	-8.3	-6.6	-7.1	-6.7	-8.3	-8.1	-7.9	-6.2	-6.4	-7.6	-7.9	-6.6
LT-24	moderate	nh	-8.8	-7.1	-6.8	-7.9	-8.8	-8.5	-8.2	-5.9	-6.3	-7.4	-7.9	-6.5
LT-33	moderate	nh	-8.0	-6.1	-7.2	-7.0	-7.9	-7.9	-7.7	-5.8	-5.9	-7.1	-8.0	-6.1
L-17	moderate	nh	-8.0	-6.6	-7.4	-8.3	-8.1	-7.9	-7.8	-6.1	-6.3	-7.4	-8.0	-6.5
LT-34	moderate	nh	-8.2	-6.3	-7.6	-6.5	-8.3	-8.1	-7.9	-5.9	-6.3	-7.1	-8.1	-6.3
LT-36	moderate	nh	-8.2	-6.3	-7.1	-6.2	-7.8	-8.1	-7.7	-5.8	-6.0	-7.0	-8.1	-6.7
LT-50	moderate	nh	-8.7	-6.6	-7.4	-6.8	-8.3	-8.5	-7.9	-6.2	-6.4	-7.5	-7.9	-5.9
LT-5	moderate	nh	-8.5	-6.5	-7.1	-8.0	-8.1	-7.9	-8.0	-5.8	-6.2	-7.0	-7.5	-6.1

Метод k-ближайших соседей

STATISTICA - [Data: Docking TDAz_L17* (15v by 43c)]															
File Edit View Insert Format Statistics Data Mining Graphs Tools Data Window Help															
Add to Workbook ▾ Add to Report ▾ Add to MS Word ▾															
LDA	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
	Y3	Y2	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13
LT-47	high	h	-8.5	-6.5	-7.9	-7.1	-8.2	-8.7	-8.3	-6.2	-6.7	-7.5	-8.5	-6.4	-6.3
LT-51	high	h	-8.4	-6.7	-7.4	-6.6	-8.1	-8.6	-8.3	-6.1	-6.5	-7.4	-8.4	-6.8	-6.1
LT-52	high	h	-8.6												
L-91	high	h	-8.6												
LT-13	high	h	-8.6												
LT-23	high	h	-8.5												
LT-48	high	h	-8.5												
LT-46	high	h	-8.2												
H-16	high	h	-8.3												
H-88	high	h	-8.6												
LT-22	high	h	-8.5												
LT-49	high	h	-9.0												
L-87	moderate	nh	-8.5												
LT-55	moderate	nh	-8.6												
LT-9	moderate	nh	-8.6	-6.8	-7.4	-8.0	-8.4	-8.2	-8.4	-6.1	-6.4	-6.9	-7.7	-6.3	-6.0
L-20	moderate	nh	-8.4	-6.6	-7.1	-7.3	-8.2	-8.3	-8.1	-5.9	-6.3	-7.4	-8.1	-6.6	-6.1
L-31	moderate	nh	-8.4	-6.9	-7.4	-7.8	-8.0	-8.4	-8.1	-6.2	-6.5	-7.4	-8.0	-6.4	-6.6
L-86	moderate	nh	-8.3	-6.8	-6.8	-8.5	-8.9	-8.2	-8.6	-6.1	-6.2	-7.2	-8.0	-6.9	-6.2
LT-21	moderate	nh	-8.6	-6.8	-7.1	-8.2	-8.8	-8.2	-8.1	-6.0	-6.2	-7.3	-8.0	-6.9	-6.1
LTG-2	moderate	nh	-7.8	-6.8	-6.9	-7.3	-7.9	-8.0	-7.4	-6.2	-6.5	-8.1	-8.4	-6.9	-6.8
L-19	moderate	nh	-8.3	-6.6	-7.1	-6.7	-8.3	-8.1	-7.9	-6.2	-6.4	-7.6	-7.9	-6.6	-6.1
LT-24	moderate	nh	-8.8	-7.1	-6.8	-7.9	-8.8	-8.5	-8.2	-5.9	-6.3	-7.4	-7.9	-6.5	-6.4
LT-33	moderate	nh	-8.0	-6.1	-7.2	-7.0	-7.9	-7.9	-7.7	-5.8	-5.9	-7.1	-8.0	-6.1	-5.5
L-17	moderate	nh	-8.0	-6.6	-7.4	-8.3	-8.1	-7.9	-7.8	-6.1	-6.3	-7.4	-8.0	-6.5	-6.0
LT-34	moderate	nh	-8.2	-6.3	-7.6	-6.5	-8.3	-8.1	-7.9	-5.9	-6.3	-7.1	-8.1	-6.3	-6.1
LT-36	moderate	nh	-8.2	-6.3	-7.1	-6.2	-7.8	-8.1	-7.7	-5.8	-6.0	-7.0	-8.1	-6.7	-5.9
LT-50	moderate	nh	-8.7	-6.6	-7.4	-6.8	-8.3	-8.5	-7.9	-6.2	-6.4	-7.5	-7.9	-5.9	-6.4
LT-5	moderate	nh	-8.5	-6.5	-7.1	-8.0	-8.1	-7.9	-8.0	-5.8	-6.2	-7.0	-7.5	-6.1	-5.6



Метод k-ближайших соседей

Выбор переменных

K-nearest neighbors: Docking TDAz_L17

Quick | Sampling | Options | Cross-validation | Memory usage

Variables

Selected variables

Continuous dependent: none

Categorical dependent: Y2

Continuous independents: X1-X13

Categorical independents: none

Response codes: Selected

Predictor codes: none

OK

Cancel

Options

SELECT CASES s

MD deletion

☒ Casewise

☐ Mean substitution

Метод k-ближайших соседей

Выбор способа сэмплинга

K-nearest neighbors: Docking TDAz_L17

Quick **Sampling** Options | Cross-validation | Memory usage

Sampling method for selecting the example set

☒ Use random sampling:
Size of prototype sample (%): Seed:

☐ Use sample variable:
 off

☐ Use first N cases:
First N:

OK

Cancel

Options

SELECT CASES 

MD deletion

☒ Casewise

☐ Mean substitution

Метод k-ближайших соседей

Выбор числа соседей и метрики расстояния

K-nearest neighbors: Docking TDAz_L17

Quick | Sampling | **Options** | Cross-validation | Memory usage

Number of nearest neighbors: **1**

Distance measure

- ☒ **Euclidean**
- ☐ Euclidean squared
- ☐ Cityblock (Manhattan)
- ☐ Chebychev

☒ **Standardize distances** ☐ Distance weighted

OK

Cancel

Options

SELECT CASES s w

MD deletion

- ☒ Casewise
- ☐ Mean substitution

Метод k-ближайших соседей

Выбор перекрестной проверки

K-nearest neighbors: Docking TDAz_L17

Quick | Sampling | Options | **Cross-validation** | Memory usage

V-fold cross validation for model selection

☒ Apply v-fold cross-validation

V value: 10 Seed: 1000

Search range for nearest neighbors:

Minimum: 1

Maximum: 5

Increment: 1

OK

Cancel

Options

SELECT CASES s w

MD deletion

☒ Casewise

☐ Mean substitution

Метод k-ближайших соседей

Точность классификации

The screenshot displays a software window titled "K-Nearest Neighbor Results: Docking TDaz_L17". The window contains the following information:

Dataset Docking TDaz_L17:
Dependent: Y2
Independents: X1, X2, X3, X4, X5, X6, X7, X8, X9, X10, X11, X12, X13...
Sample size = 30 (Examples), 11 (Test), 41 (Overall)

KNN results:
Number of nearest neighbors = 1
Distance measure: Cityblock (Manhattan)
Input standardization: on
Averaging: distance weighted
Cross-validation accuracy (%) = 53.33333

The bottom section of the window features several tabs and controls:

- Top tabs: Quick (selected), Plots, Options, Custom predictions
- Left sidebar buttons: Summary, Descriptive statistics, Predictions, Histograms, Save
- Right sidebar buttons: Summary, Cancel, Options (dropdown), Code generator (dropdown)
- Bottom right: Sample selection (radio buttons for Examples, Test (selected), Overall)

In the "Predictions" section, the following options are visible:

- Include: ☐ Independents, ☐ Dependents, ☒ Predictions
- ☒ Accuracy, ☐ Confidence

Метод k-ближайших соседей

Классификация нового объекта

K-Nearest Neighbor Results: Docking TDAz_L17

Dataset Docking TDAz_L17:
Dependent: Y2
Independents: X1, X2, X3, X4, X5, X6, X7, X8, X9, X10, X11, X12, X13...
Sample size = 30 (Examples), 11 (Test), 41 (Overall)

KNN results:
Number of nearest neighbors = 1
Distance measure: Cityblock (Manhattan)
Input standardization: on
Averaging: distance weighted
Cross-validation accuracy (%) = 53.33333

Quick | Plots | Options | **Custom predictions**

 **User defined case**

 Predictions

 Summary

Cancel

 Options

 Code generator

Sample

☐ Examples

☒ Test

☐ Overall

Метод k-ближайших соседей

Задаем значения переменных

Custom Inputs Spreadsheet

Custom Inputs Spreadsheet
Double click on highlighted cell to edit

X1 -8

X2 -8

X3 -8

X4 -8

X5 -8

X6 -8

X7 -8

X8 -8

X9 -8

X10 -8

X11 -8

X12 -8

OK Cancel

Indeps.	Value
X1	-8
X2	-8
X3	-8
X4	-8
X5	-8
X6	-8
X7	-8
X8	-8
X9	-8
X10	-8
X11	-8
X12	-8

Метод k-ближайших соседей

Выполняем прогноз

K-Nearest Neighbor Results: Docking TDAz_L17

Dataset Docking TDAz_L17:
Dependent: Y2
Independents: X1, X2, X3, X4, X5, X6, X7, X8, X9, X10, X11, X12, X13...
Sample size = 30 (Examples), 11 (Test), 41 (Overall)

KNN results:
Number of nearest neighbors = 1
Distance measure: Cityblock (Manhattan)
Input standardization: on
Averaging: distance weighted
Cross-validation accuracy (%) = 53.33333

Quick | Plots | Options | Custom predictions

User defined case

Predictions

Summary

Cancel

Options

Code generator

Sample

- ☐ Examples
- ☒ Test
- ☐ Overall

Метод k-ближайших соседей

Результат прогноза

STATISTICA - [Workbook6* - K-Nearest Neighbors Custom predictions (Docking TDAz_L17)]

File Edit View Insert Format Statistics Data Mining Graphs Tools Data Workbook Window Help

Workbook6*

- Machine Learning (Docking TDAz_L17)
 - KNN Results Dialog
 - K-Nearest Neighbors Custom predictions (Docking TDAz_L17)

Independent and dependent variables	K-Nearest Neighbors Custom predictions (Docking TDAz_L17)
Value	
X1	-8.000000
X2	-8.000000
X3	-8.000000
X4	-8.000000
X5	-8.000000
X6	-8.000000
X7	-8.000000
X8	-8.000000
X9	-8.000000
X10	-8.000000
X11	-8.000000
X12	-8.000000
X13	-8.000000
Y2	nh

To be continued ...

