Machine Learning and Data Mining
Classification techniques

Dmitry I. Ignatov

Computer Science Faculty
Department of Data Analysis and Artificial Intelligence

2019

Methods' review

In this lecture:

1. Classification problem
2. 1-Rule algorithm
3. Distance-based classifiers
   - k Nearest Neighbours approach
4. Bayes classifier
   - Naïve Bayes Classifier
5. Logistic regression
6. Quality metrics
   - Precision, Recall, F-measure
   - ROC curve and AUC

Later:

1. Support Vector Machines (SVM)
2. Artificial Neural Networks

Outline

Learning by examples (LE)

Problem statement

Input:
- A set of **objects** X
- A set of associated class **labels** Y
- A **Target function** $y : X \to Y$ defined over a finite set of examples (**training set**)

Task: Find (learn) an algorithm $a : X \to Y$ that recovers $y(x)$ (i.e. predicts its values not only for training set, but for the whole).

Classification

If $Y = \{1, 2, \ldots, I\}$, then the corresponding problem LE is called a classification problem over I disjoint classes.

Examples of applications

- Document classification
- Opinion Mining, Sentiment Analysis
- Spam detection (Spam/Ham)
- Medical diagnostics: disease diagnosis, treatment prescription, duration and outcome prediction, etc.
- Drug design, toxicity prediction
- Churn prediction (clients)
- Searching for mineral deposits in Geology
- ...

Loss function

Loss function

A function $L(a(x), y(x))$ is the error of the algorithm a for the object x. Usually, in classification:

$$L(a(x), y(x)) = \begin{cases} 1, & a(x) \neq y(x); \\ 0, & a(x) = y(x). \end{cases}$$

Empirical risk

Classification task (learning by examples) as an optimisation problem:

$$\frac{1}{n} \sum_{t=1}^{n} L(a(x_t), y(x_t)) \underset{a}{\rightarrow} min$$

In case of known probability density $p(x, y)$:

$$\iint_{X \times Y} L(a(x), y(x)) p(x, y) dx dy \underset{a}{\rightarrow} min$$

Learning, parameters' tuning, quality validation

Cross-validation

An input dataset is split into 2 (or 3) disjoint subsets:

- X^{train} Training Set
  - For learning algorithms
  - Parameter estimation
- X^{test} Test Set
  - Quality assessment
- X^{valid} Validation Set
  - Estimation of the algorithm's hyperparameters

NP: During the splitting it is important to keep «proportion of classes».

Outline

One-Rule

[Witten et al., Data Mining, 2011]

Algorithm

For each attribute,
  For each value of that attribute, make a rule as follows:
    count how often each class appears
    find the most frequent class
    make the rule assign that class to this attribute value.
  Calculate the error rate of the rules.
Choose the rules with the smallest error rate.

Outline

k Nearest Neighbours (kNN)

Let objects $x_i \in X$ be described by a feature matrix

$$F_{n\times m} = [f_j(x_i)] = \begin{pmatrix} f_1(x_1) & \cdot & f_m(x_1) \\ \cdot & \cdot & \cdot \\ f_1(x_n) & \cdot & f_m(x_n) \end{pmatrix}$$

kNN

Let $d : X \times X \to [0, \infty)$ be a distance function.
For any training set we know that $y : X^{train} \to Y$
`Hypothesis:` Similar objects belong to the same class.

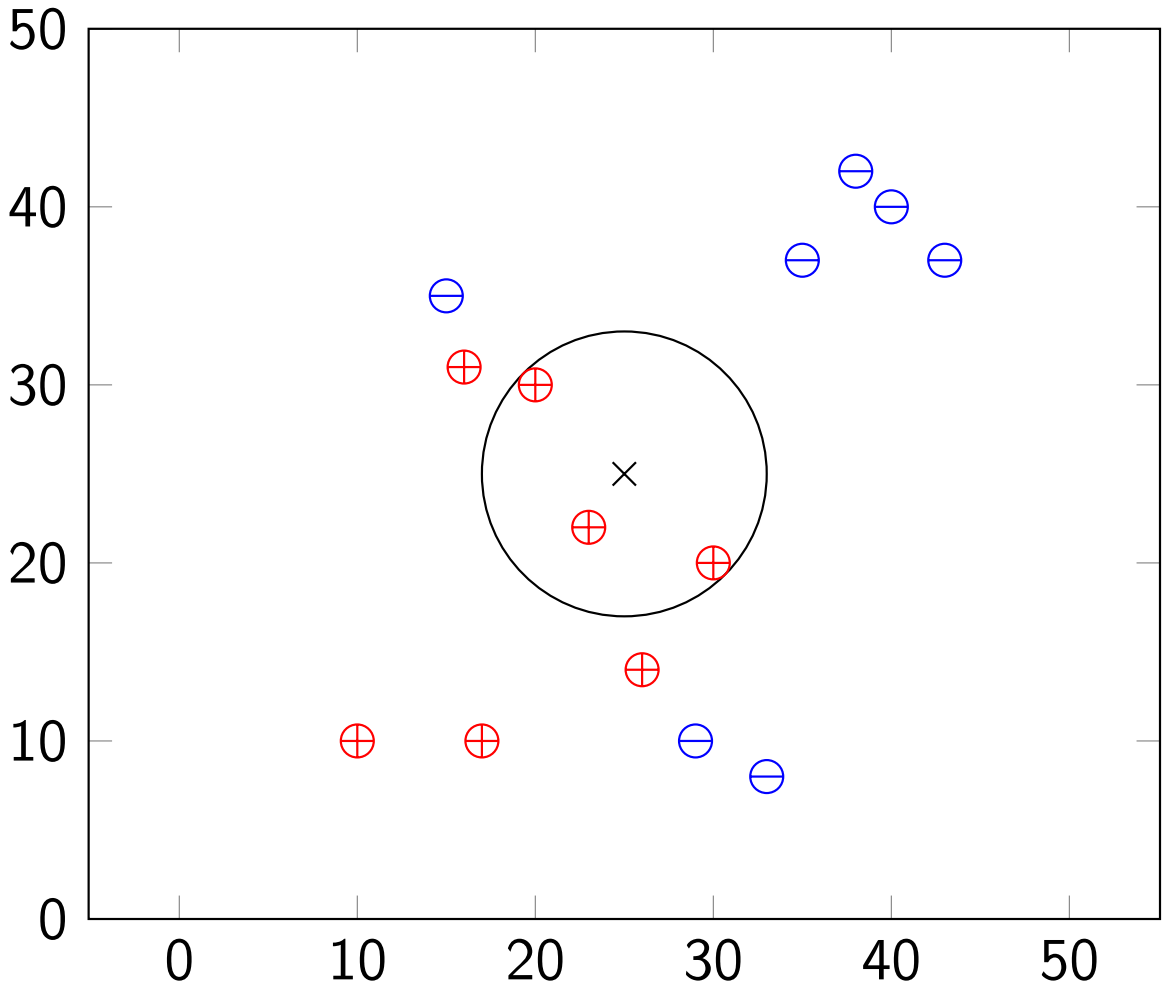
For an object $v \in X$ rank all the objects from X^{train} in increasing order w.r.t. to their distance v:

d	$d(v, x^{(1)})$	$\leq$	$d(v, x^{(2)})$	$\leq$	$\ldots$	$\leq$	$d(v, x^{(k)})$	$\leq$	$\ldots$
$y(x)$	1		2		$\ldots$		2		$\ldots$

The object v is assigned to the class such that its elements forms a majority among the neighbours of k

Example

kNN with $k = 3$



x is an undetermined object

k Nearest Neighbours

Algorithm's variations

- kNN with ($k = 1$)
- Weighted kNN
  - Linear decay: $w_i = \frac{k+1-i}{k}$

Pro et Contra

+ Simple implementation
– Storing the whole input dataset
– High computational complexity
– Low noise- and error-tolerance $\rightarrow$ reference objects

Outline

Bayes formula, Law of total probability

Bayes formula

$$\mathrm{P}(A|B) = \frac{\mathrm{P}(B|A)\,\mathrm{P}(A)}{\mathrm{P}(B)}$$

Law of total probability

$$\mathrm{P}(B) = \sum_i \mathrm{P}(A_i)\,\mathrm{P}(B|A_i)$$

Problem 1

According medical statistics, one half of meningitis cases causes neck pain. The fraction of patients with meningitis is about 1/50000; The neck pain symptom is peculiar to 1/20 of all patients. Is "neck pain" a strong symptom of meningitis?

Bayes formula, Law of total probability

Problem 2 Monty Hall problem

Assume that you are a participant of a TV-show. There are three closed doors in front of you. One of them hides a cool motorbike and behind the others are goats. You have to choose one of the doors. After a while you pick one of the doors. The showman, in his turn, opens one of the remaining doors showing a goat behind it, and let you the chance to ~~ride it~~ switch your choice. What is more beneficial to do?

Bayesian classification techniques

Naïve Bayes

Let X be a set of objects with class labels $Y = \{y_1, y_2, \ldots, y_k\}$ and attributes $\{a_1, a_2, \ldots, a_m\}$.
Hypothesis (Inductive bias): attributes $\{a_1, a_2, \ldots, a_m\}$ are independent.
We need to classify $x_i \in X$ with the attribute values $\{a_1^*, a_2^*, \ldots, a_m^*\}$:

$$y^* = \arg\max_{c \in Y} \mathrm{P}(y(x_i) = c | a_1^*, a_2^*, \ldots, a_m^*)$$

The value of $\mathrm{P}(y(x_i) = c | a_1^*, a_2^*, \ldots, a_m^*)$ is computed on X by Bayes rule under independence assumption:

$$\mathrm{P}(y(x_i) = c | a_1^*, a_2^*, \ldots, a_m^*) = \frac{\mathrm{P}(a_1^*, a_2^*, \ldots, a_m^* | y(x_i) = c)\, \mathrm{P}(y(x_i) = c)}{\mathrm{P}(a_1^*, a_2^*, \ldots, a_m^*)}$$

$$\mathrm{P}(a_1^*, a_2^*, \ldots, a_m^* | y(x_i) = c) = \prod_j \mathrm{P}(a_j^* | y(x_i) = c)$$

Problem

Given a data table below

Day	Outlook	Temperature	Humidity	Wind	Play Tennis
Dl	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
Dl0	Rain	Mild	Normal	Weak	Yes
Dl1	Sunny	Mild	Normal	Strong	Yes
Dl2	Overcast	Mild	High	Strong	No

answer the question: Is it worth to play tennis today:
$\{sunny, cool, high, strong\}$?

Smoothing

What if $\prod_j \mathrm{P}(a_j^*|y(x_i) = c)$ there exists j' such that

Additive Smoothing

$$\mathrm{P}_{add}(a_j|y(x_i)) = \frac{N_{a_j}^{y(x_i)} + \lambda}{N^{y(x_i)} + \lambda V}$$

where $N_{a_j}^{y(x_i)}$ is the number of objects with the attribute value a_j in class $y(x_i)$, V is the number of different values of a_j, and $\lambda > 0$.

- Probabilities are biased.
- λ needs tuning, but usually $\lambda = 1$

Outline

Logistic regression

Let X is split in two classes $Y = \{0, 1\}$. For example, there are two kinds of tumor, benign ($y = 1$) and malignant ($y = 0$).

Problem

Having attribute description of x estimate benign probability, i.e. $p_i = \mathrm{P}(y(x_i) = 1 | x_i)$.
Define logistic transformation as $logit(p_i) = \log(\frac{p_i}{1-p_i})$.

Logistic regression

Let X is split in two classes $Y = \{0, 1\}$. For example, there are two kinds of tumor, benign ($y = 1$) and malignant ($y = 0$).

Problem

Having attribute description of x estimate benign probability, i.e. $p_i = \mathrm{P}(y(x_i) = 1|x_i)$.
Define logistic transformation as $logit(p_i) = \log(\frac{p_i}{1-p_i})$.

$$\log(\frac{p_i}{1-p_i}) = \beta^T x_i \Leftrightarrow \frac{p_i}{1-p_i} = \exp(\beta^T x_i) \Leftrightarrow p_i = \frac{1}{1+\exp(-\beta^T x_i)},$$

$$h(\beta, x_i) = \frac{1}{1+\exp(-\beta^T x_i)}$$

$h(\beta, x_i) = 0.7 \Leftrightarrow$ the label of i-th object is $y(x_i) = 1$ with probability 0.7.

Sigmoid

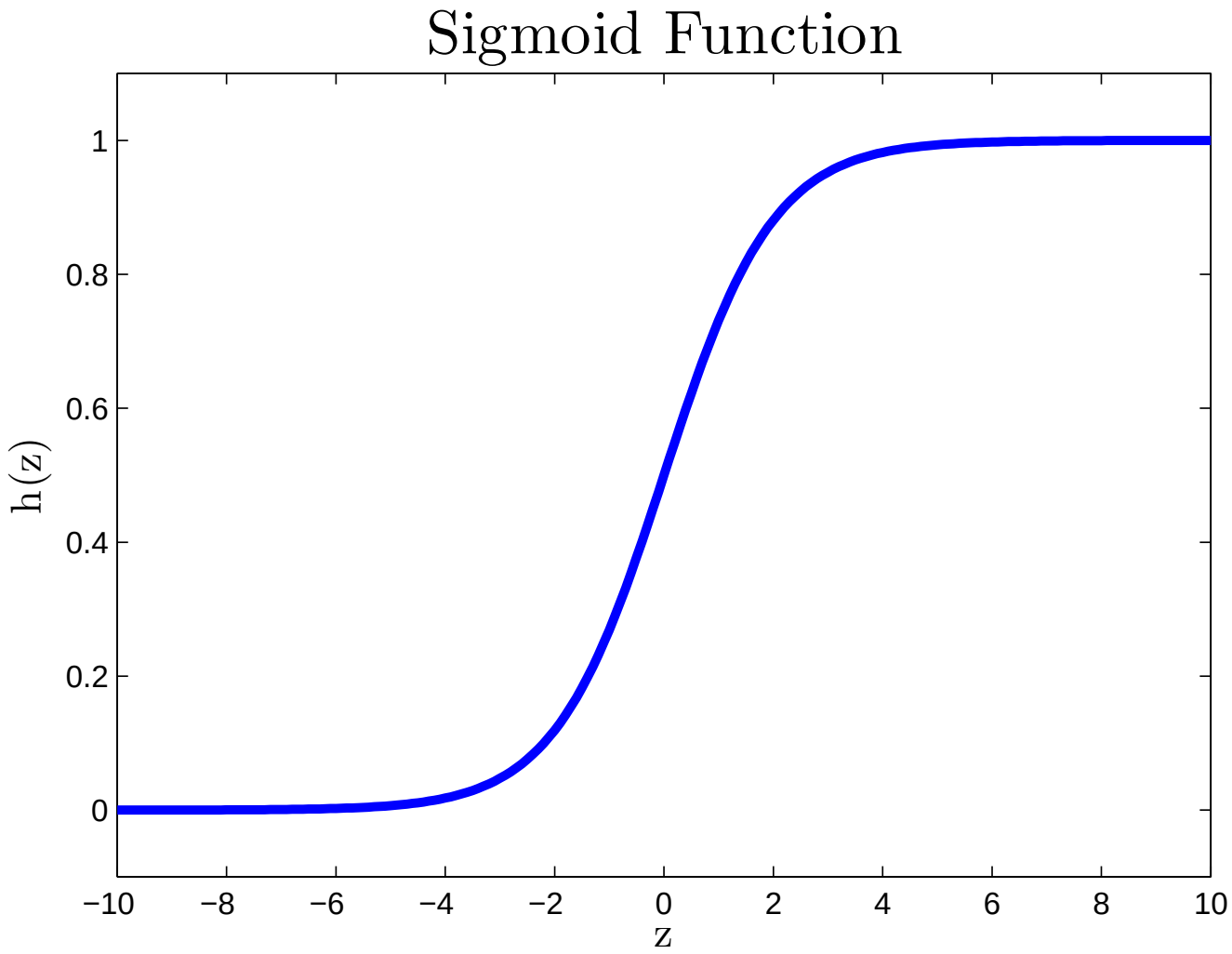


Рис. 1: *Sigmoid function* $h(z) = \frac{1}{1+\exp(-z)}$

Sigmoid

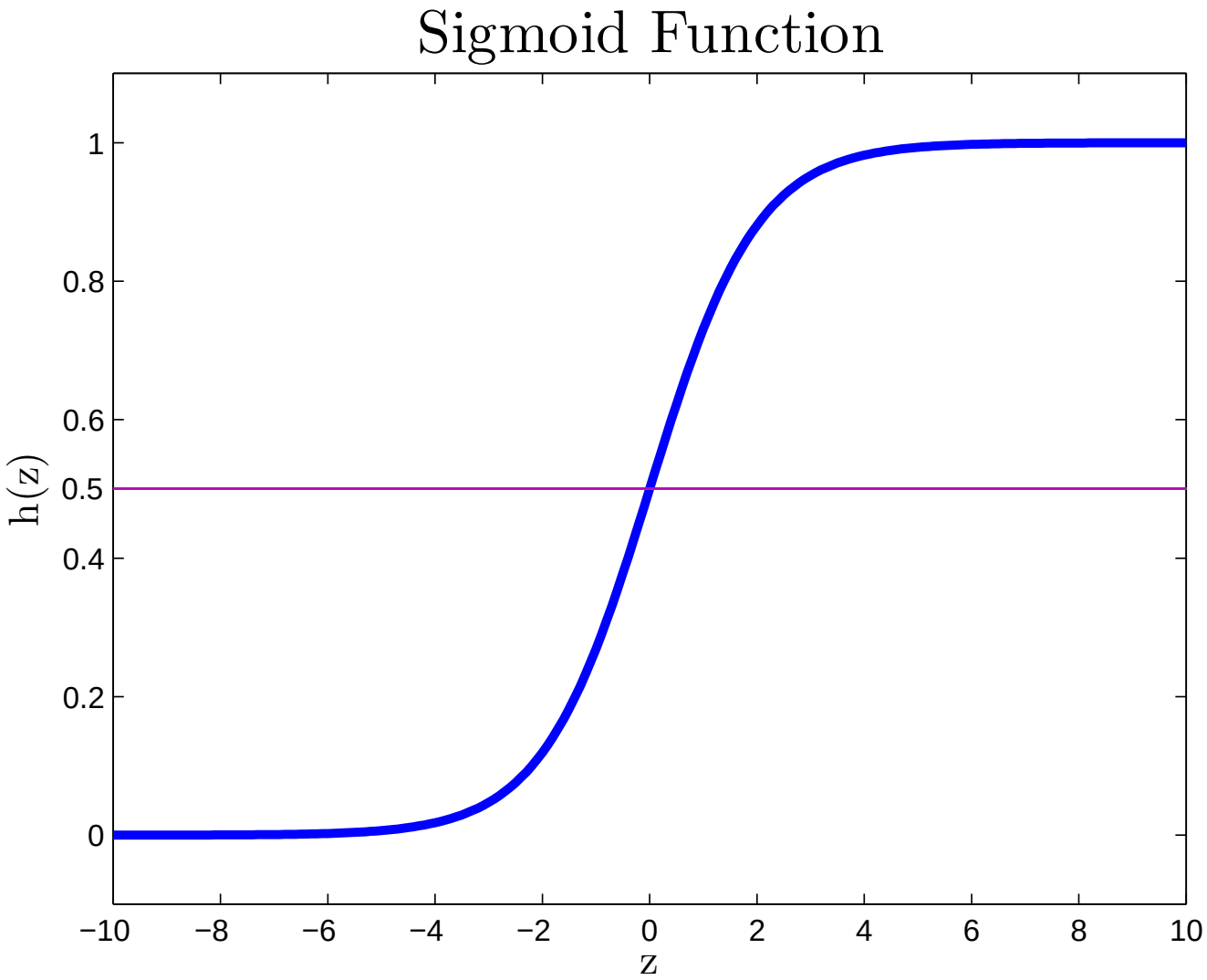


Рис. 1: *Sigmoid function* $h(z) = \frac{1}{1+\exp(-z)}$

Decision hyperplane

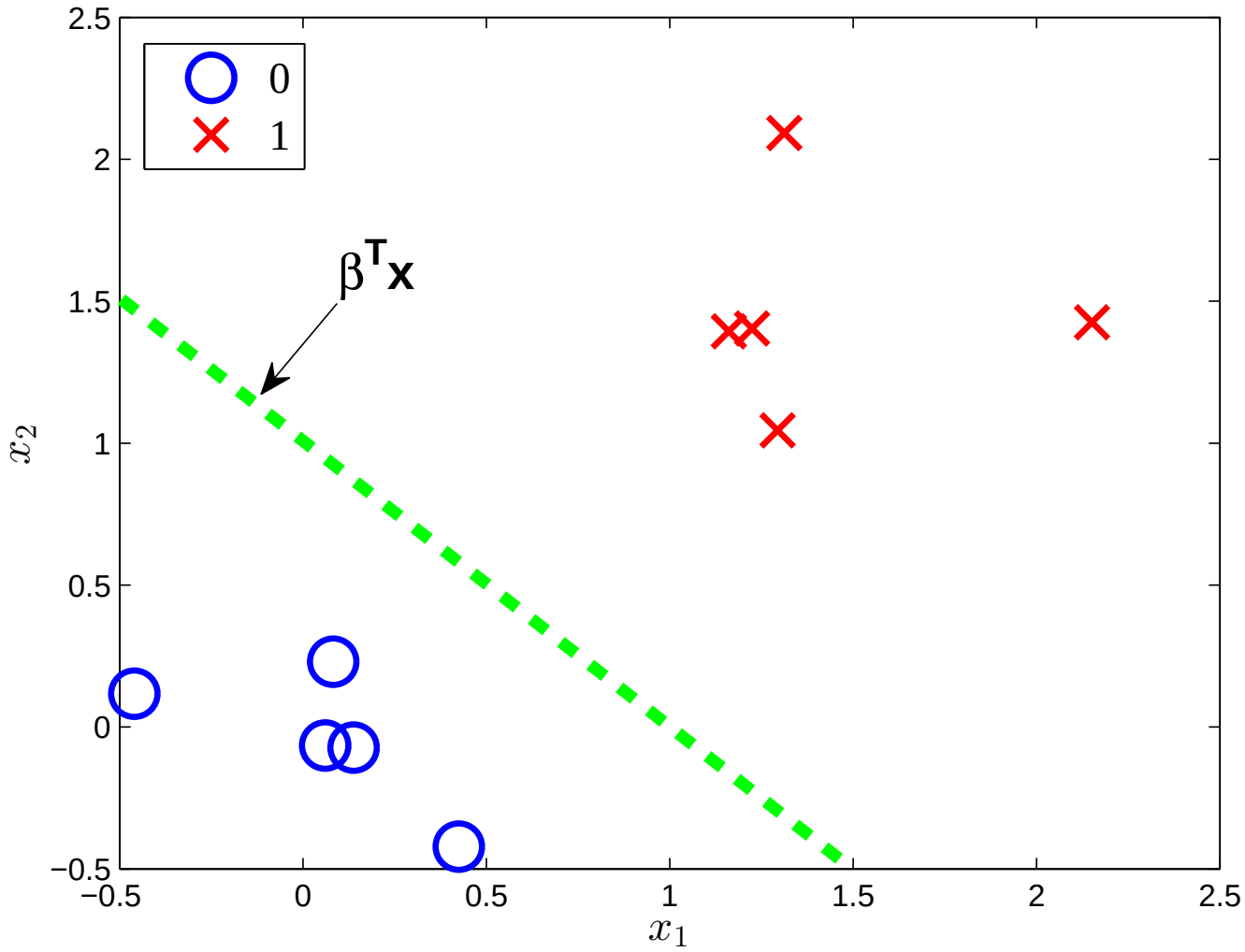


Рис. 2: *Decision hyperplane*

Tuning β by gradient descent

Loss function

Quadratic loss $L(h(\beta, x), y(x)) = \sum_i (h(\beta, x_i) - y(x_i))^2$ is not applicable since it is not convex for logistic function.

$$L(h(\beta, x), y(x)) = -\sum_i [y(x_i) \log(h(\beta, x_i)) + (1 - y(x_i)) \log(1 - h(\beta, x_i))]$$

Gradient descent

Iterative search $\min_\beta L(h(\beta, x), y(x))$:

$$\beta_j = \beta_j - \alpha \frac{\partial}{\partial \beta_j} L(h(\beta, x), y(x))$$

where α is learning rate. There are different halt criteria.

Tuning β by gradient descent

Loss function

Quadratic loss $L(h(\beta, x), y(x)) = \sum_i (h(\beta, x_i) - y(x_i))^2$ is not applicable since it is not convex for logistic function.

$$L(h(\beta, x), y(x)) = -\sum_i [y(x_i) \log(h(\beta, x_i)) + (1 - y(x_i)) \log(1 - h(\beta, x_i))]$$

Gradient descent

Iterative search $\min_\beta L(h(\beta, x), y(x))$:

$$\beta_j = \beta_j - \alpha \frac{\partial}{\partial \beta_j} L(h(\beta, x), y(x))$$

where α is learning rate. There are different halt criteria.

$$\frac{\partial}{\partial \beta_j} L(h(\beta, x), y(x)) = \sum_i (h(\beta, x_i) - y(x_i)) x_{ij}$$

Outline

Precision, Recall, F-measure

Results of binary classification in terms of confusion matrix:

$$\begin{pmatrix} TP & FP \\ FN & TN \end{pmatrix} \text{ where}$$

TP (True Positive) is the number of true predictions for the positive class, FP (False Positive), its false predictions' number; FN (False Negative) and TN (True Negative) are defined similarly for the negative class.

$$Precision = \frac{TP}{TP + FP} \qquad Recall = \frac{TP}{TP + FN}$$

There is a trade-off between Precision and Recall, which can be captured by F-measure:

$$F_{\alpha,\beta}(Precision, Recall) = \frac{1}{\frac{1}{\alpha Precision} + \frac{1}{\beta Recall}}$$

What to do in multiclass setting?

ROC curve (ROC, Receiver Operating Characteristic)

X — False Positive Rate

Y — True Positive Rate

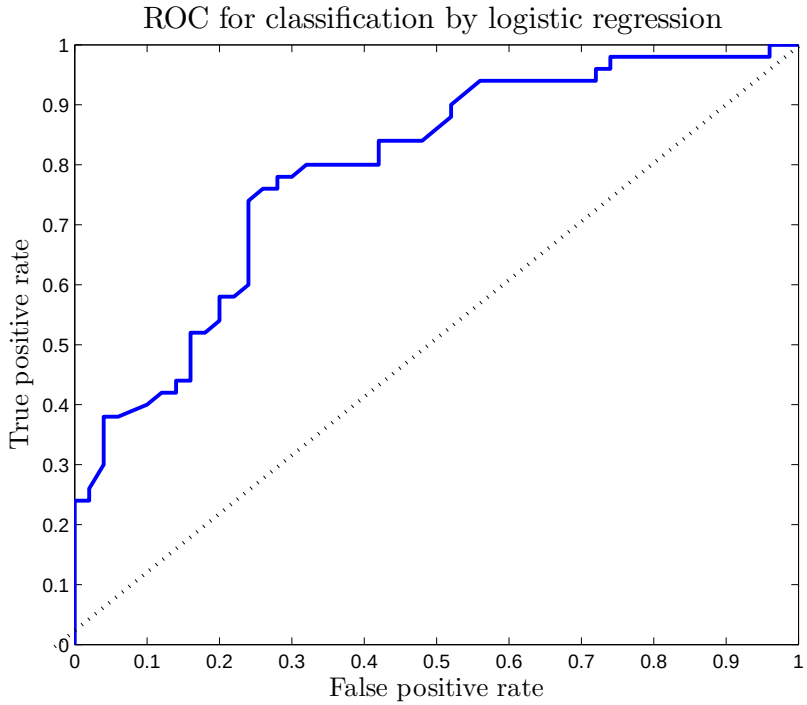


Рис. 3: *ROC curve for logistic regression*

Each point of the ROC curve is the result of the studied classification algorithm under a specific decision boundary value.

Generally, the square under the ROC curve characterise the quality of classification – AUC (Area Under Curve).

Questions and contacts

www.hse.ru/staff/dima

Thank you!

dmitrii.ignatov[at]gmail.com

dignatov[at]hse.ru