ITMO

Natural Language Processing

(for Bioinformatics)

About me

ITMO

Applied LLM Engineer @ ecom.tech

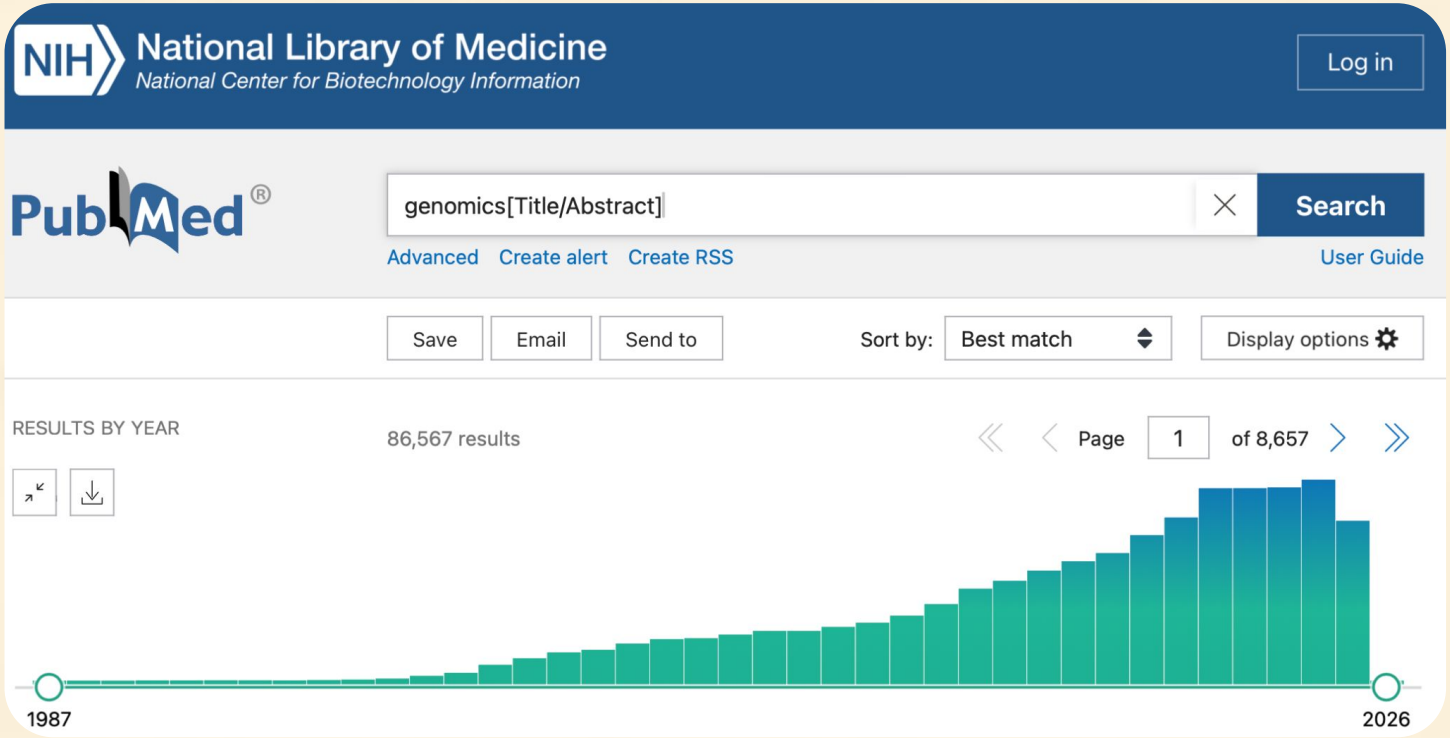
Building production-grade multi-agent & multimodal systems.

Specializing in:
- Agentic Orchestration
- Multimodal RAG & Code Generation
- High-Performance Inference

Previously: Sber AI, Tochka, MTS AI
ITMO University (M.Sc. in Artificial Intelligence)

Oleg Zagorulko

The Data Explosion in genomics



https://pubmed.ncbi.nlm.nih.gov/?term=genomics%5BTitle%2FAbstract%5D&timeline=expanded

Why NLP in Bioinformatics?

ITMO

Data Flood →

PubMed

RCSB PDB
PROTEIN DATA BANK

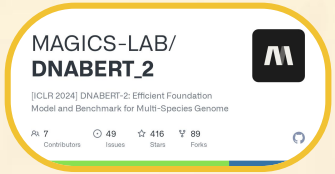
Why NLP in Bioinformatics?

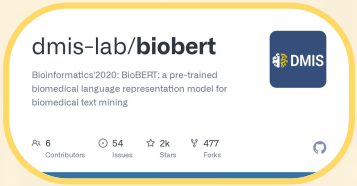
ITMO

Data Flood → **NLP Models** →

PubMed

RCSB PDB PROTEIN DATA BANK





https://github.com/MAGICS-LAB/DNABERT_2

https://github.com/dmis-lab/biobert

Why NLP in Bioinformatics?

ITMO

Data Flood → **NLP Models** → **Insights**

PubMed

RCSB PDB PROTEIN DATA BANK

MAGICS-LAB/**DNABERT_2**

[ICLR 2024] DNABERT-2: Efficient Foundation Model and Benchmark for Multi-Species Genome

7 Contributors · 49 Issues · 416 Stars · 89 Forks

dmis-lab/**biobert**

Bioinformatics'2020: BioBERT: a pre-trained biomedical language representation model for biomedical text mining

6 Contributors · 54 Issues · 2k Stars · 477 Forks

https://github.com/MAGICS-LAB/DNABERT_2

https://github.com/dmis-lab/biobert

Why NLP in Bioinformatics?

ITMO

Data Flood → **NLP Models** → **Insights**

PubMed

RCSB PDB
PROTEIN DATA BANK

MAGICS-LAB/
DNABERT_2
[ICLR 2024] DNABERT-2: Efficient Foundation Model and Benchmark for Multi-Species Genome
7 Contributors · 49 Issues · 416 Stars · 89 Forks

dmis-lab/**biobert**
Bioinformatics'2020: BioBERT: a pre-trained biomedical language representation model for biomedical text mining
6 Contributors · 54 Issues · 2k Stars · 477 Forks

Or not...

https://github.com/MAGICS-LAB/DNABERT_2

https://github.com/dmis-lab/biobert

Roadmap

ITMO

Lecture 2

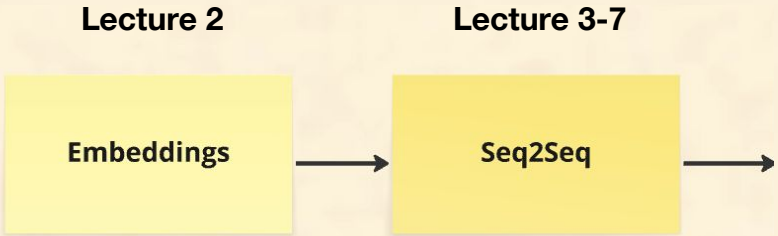
Embeddings

Distributional semantics

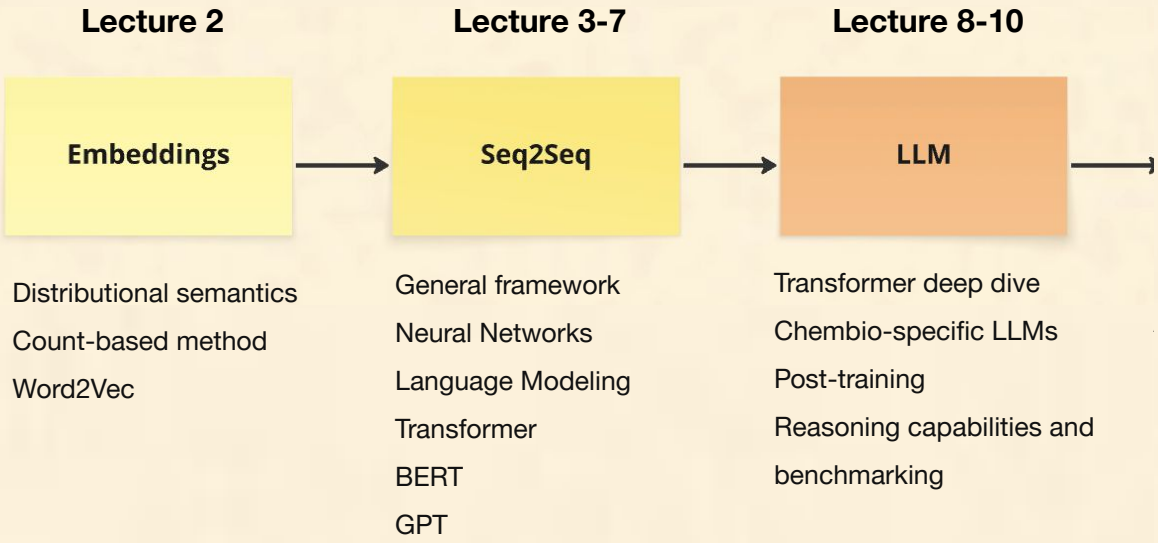
Count-based method

Word2Vec

Roadmap

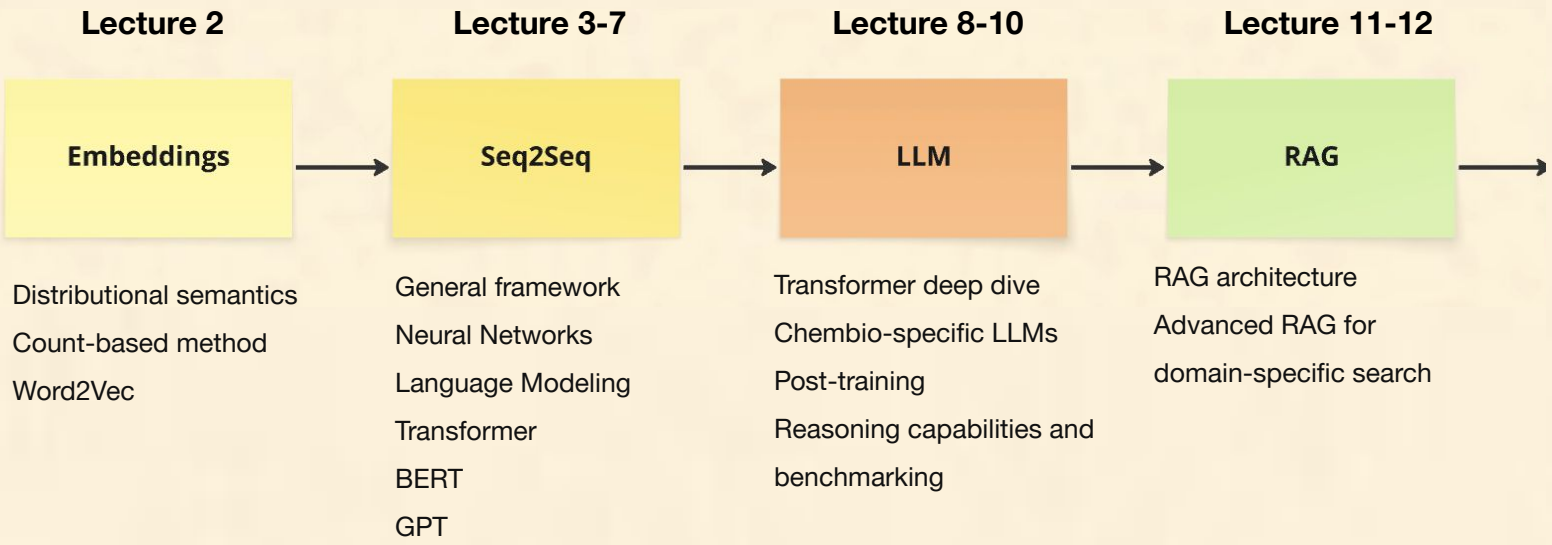

Lecture 2
Lecture 3-7
Embeddings
Seq2Seq
Distributional semantics
Count-based method
Word2Vec
General framework
Neural Networks
Language Modeling
Transformer
BERT
GPT

Roadmap


Lecture 2
Lecture 3-7
Lecture 8-10
Embeddings
Seq2Seq
LLM
Distributional semantics
Count-based method
Word2Vec
General framework
Neural Networks
Language Modeling
Transformer
BERT
GPT
Transformer deep dive
Chembio-specific LLMs
Post-training
Reasoning capabilities and benchmarking

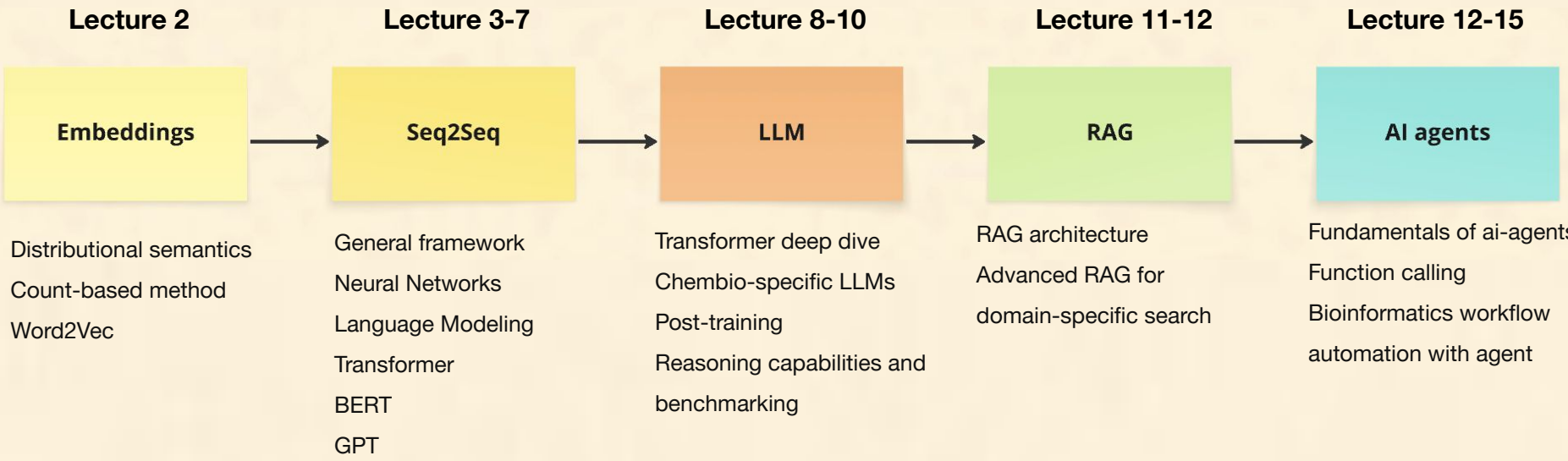
Roadmap

ITMO



Roadmap

ITMO



Format and Grading

ITMO

Tests (theory checks) (max ***+ 25%***)

- Short multiple-choice or open-ended questions after selected lectures
- Assess understanding of theoretical concepts and key terminology

Homework Assignments (max ***+75%***)

- Three practical tasks, focused on implementing NLP techniques for bioinformatics.
- Evaluated on correctness, clarity of code, and relevance of results

Optional Homework (max ***+25%***)

- Advanced task

Format and Grading

ITMO

A: 90-100% - Excellent

B: 80-89% - Good

C: 70-79% - Satisfactory

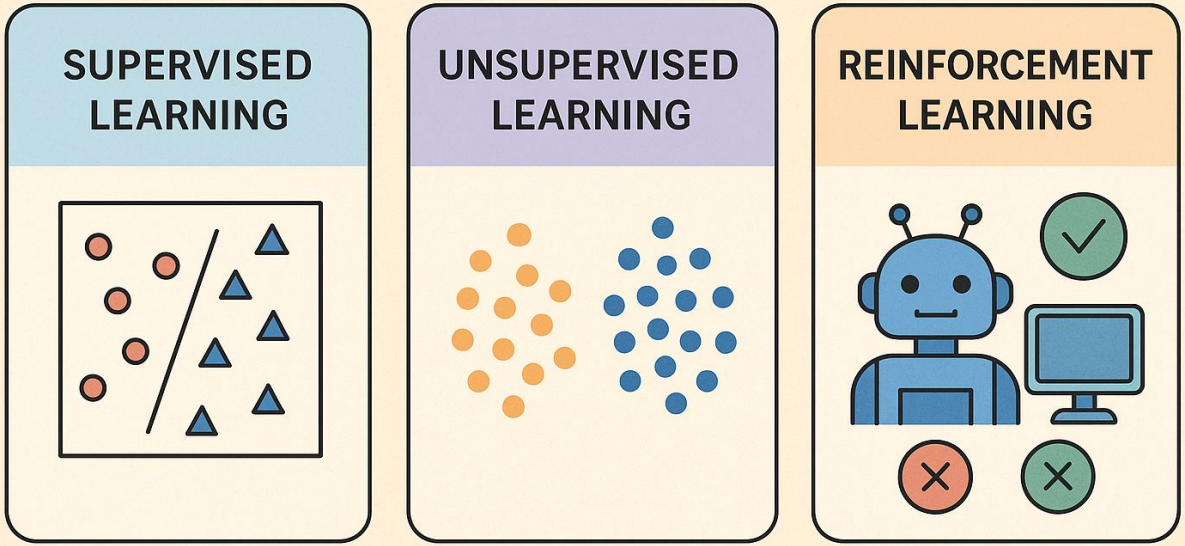
D: 60-69% - Poor

F: <60% - Fail

ML recap

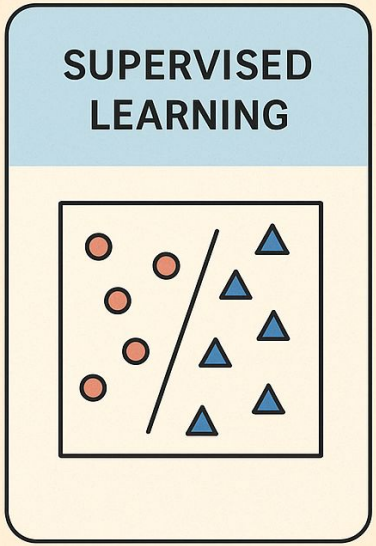
Types of Machine Learning Tasks

ITMO



Types of Machine Learning Tasks

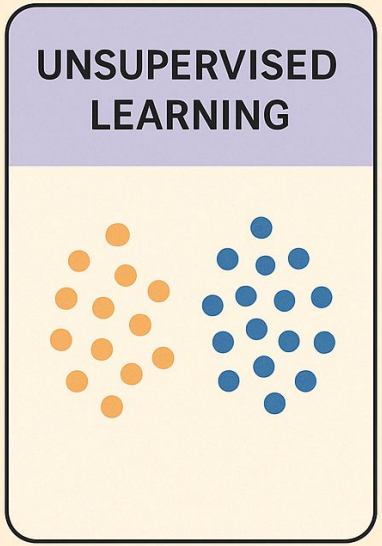
ITMO



TL;DR: regression, classification, ranking

Types of Machine Learning Tasks

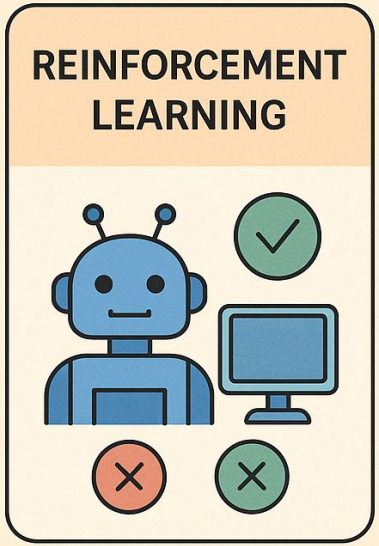
ITMO



TL;DR: cluster analysis

Types of Machine Learning Tasks

ITMO



TL;DR: training an agent to act in an environment in order to maximize reward

Linear models

ITMO

TL;DR: Linear models are the simplest and most interpretable class of functions. They are a natural starting point for both classification and regression tasks, because they provide a fast, transparent, and mathematically tractable way to map objects to targets.

To assign each object (e.g., a card transaction, a mining site) a target value.

Classification: $X \to \{0,1,\ldots,K\}$
Regression: $X \to R$

What makes them useful ?

ITMO

- The **simplest** parameterized family of functions.
- **Easy** to **compute** and **interpret**.
- Provide a **clear performance** baseline.
- Serve as a foundation for more complex (nonlinear) models.

Linear models: weighted sum of features + bias

ITMO

TL;DR: A linear model predicts the target as a weighted sum of features plus a bias. It is called *linear* because it is linear with respect to the numerical features. In regression, it approximates values with a line (or hyperplane); in classification, it defines a separating rule between classes.

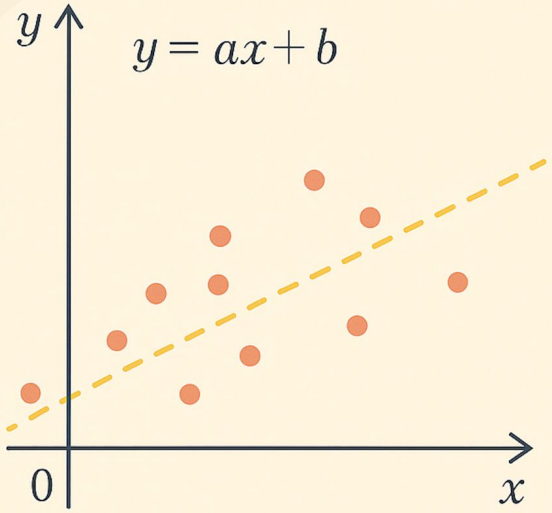
Linear functions: $y = w_1 x_1 + \ldots + w_D x_D + w_0$, or more compactly $y = \langle x, w \rangle + w_0$

Works directly with numerical features.

Regression: fits a line (or hyperplane) to approximate target values.

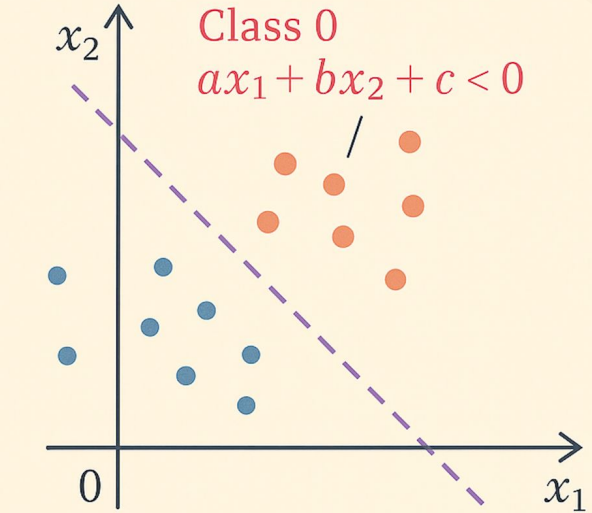
Classification: defines a separating rule (positive side → one class, negative side → another).

Linear regression and classification



x – (only) feature

y – target



$ax_1 + bx_2 + c > 0$

Decision boundary

x_1, x_2 – features

Logistic regression

TL;DR: Logistic regression maps linear model outputs to probabilities using the sigmoid function, making it ideal for binary classification tasks such as click-through prediction.

Classes: 0 and 1

Goal: predict the probability of an event, not just a label

Linear model outputs values on $(-\infty,+\infty)$

We need mapping to [0,1]

Use logit (log-odds)

Model estimates the probability of the positive class

$$\langle w, x_i \rangle = \log \frac{p}{1-p}$$

$$p = \frac{1}{1 + e^{-\langle w, x_i \rangle}} = \sigma(\langle w, x_i \rangle)$$

Regression Metrics

ITMO

$$MSE(y^{true}, y^{pred}) = \frac{1}{N}\sum_{i=1}^{N}(y_i - f(x_i))^2$$

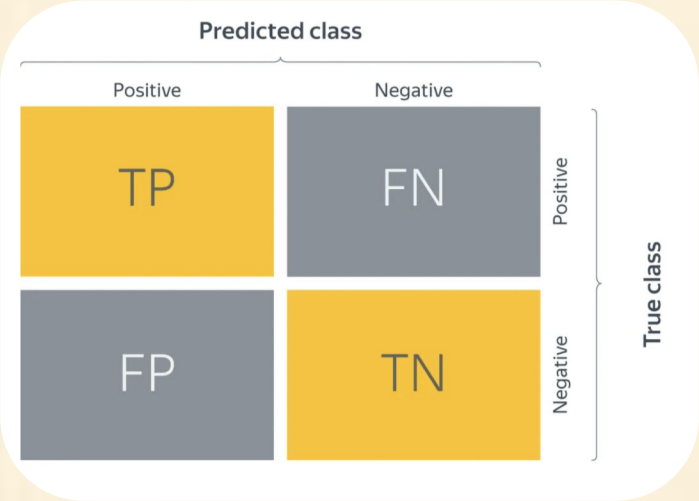
$$MAPE(y^{true}, y^{pred}) = \frac{1}{N}\sum_{i=1}^{N}\frac{|y_i - f(x_i)|}{|y_i|}$$

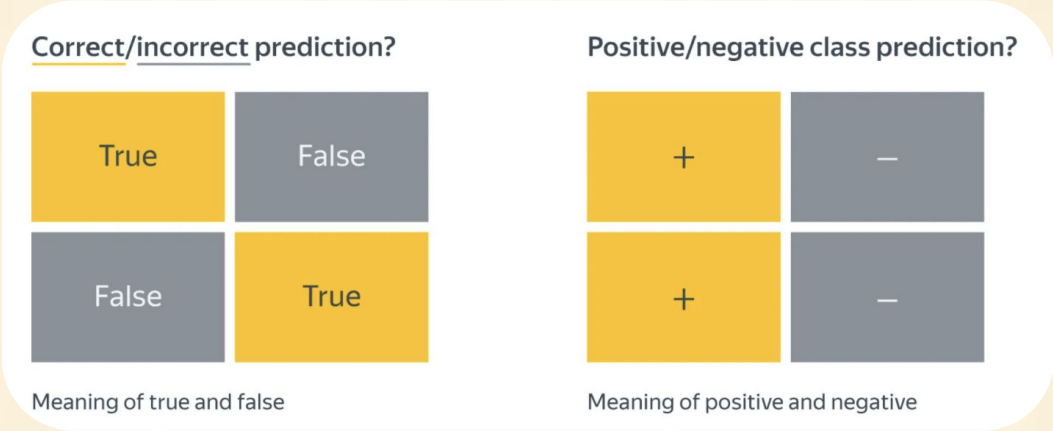
$$MAE(y^{true}, y^{pred}) = \frac{1}{N}\sum_{i=1}^{N}|y_i - f(x_i)|$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - f(x_i))^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}$$

https://education.yandex.ru/handbook/ml/article/metriki-klassifikacii-i-regressii

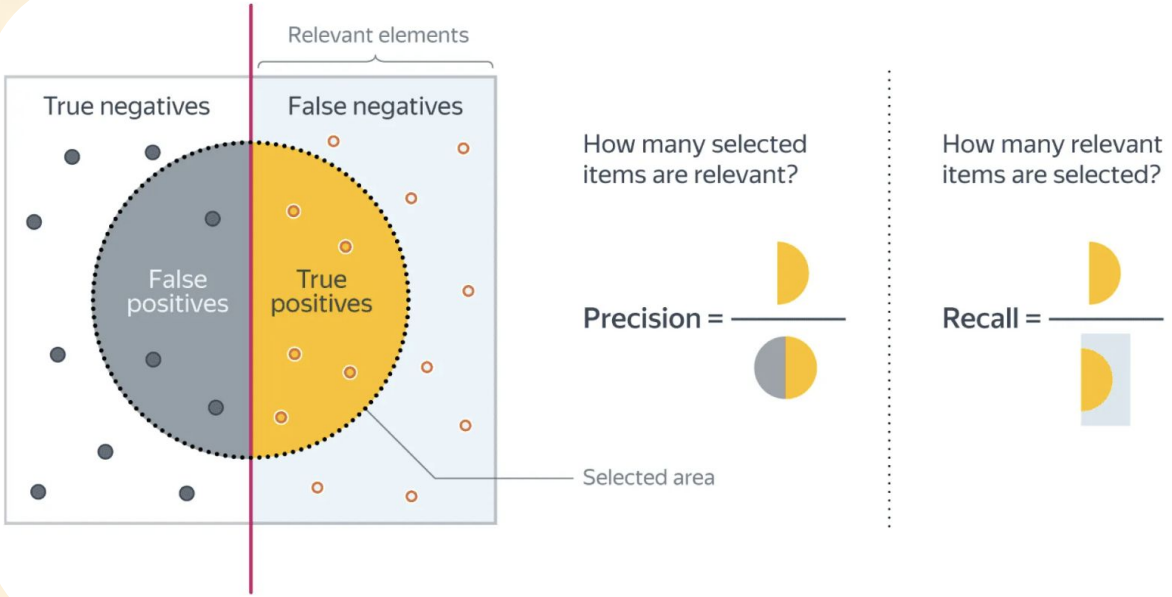
Classification Metrics





https://education.yandex.ru/handbook/ml/article/metriki-klassifikacii-i-regressii
https://en.wikipedia.org/wiki/Precision_and_recall

Classification Metrics



$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

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

QA

ITMO