Final project presentation

Machine Learning course 2024

Skoltech
Skolkovo Institute of Science and Technology

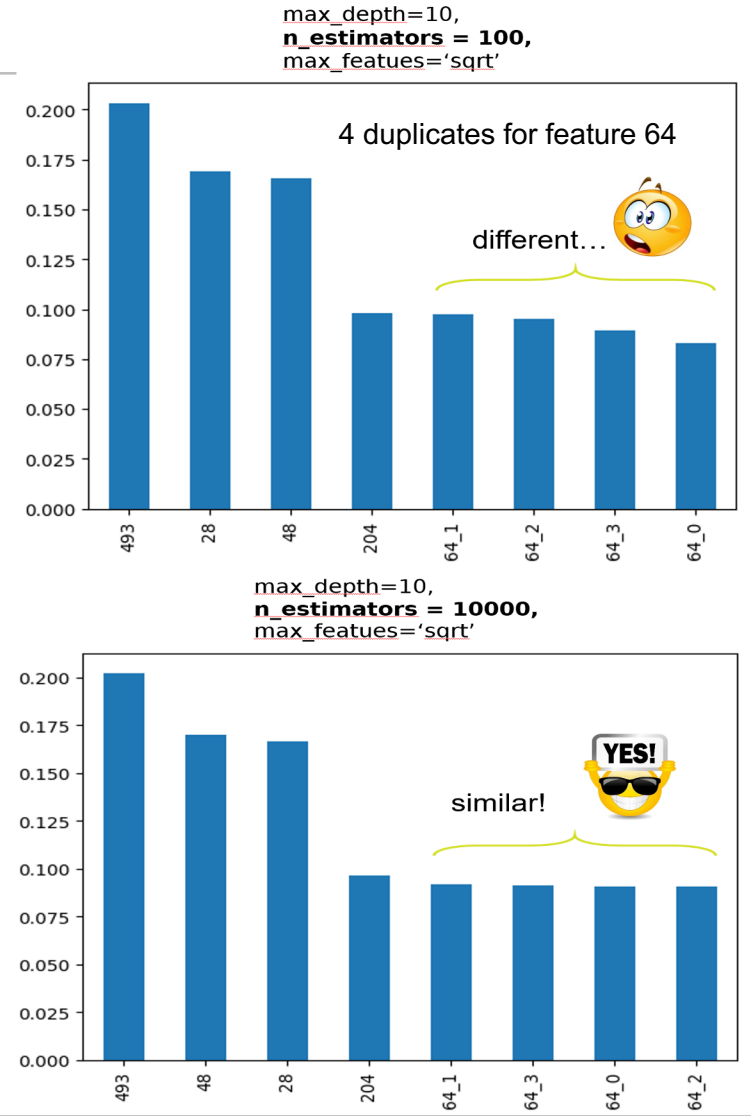
Stable feature importance in Random Forest: estimating the number of decision trees

Andrey Lange, @lange_am

Problem: When the number of trees in RF is insufficient, the Feature Importance histogram may not be plausible and stabilises as the number of trees increases

Tasks:

1. To propose statistically sound criteria for selecting the number of trees in RF
2. To demonstrate that FI stops changing visually when the number of trees increases beyond the found number
3. Assess the degree of stability of the sorted FI order, for example, by analysing each pair of neighbouring features (other ideas are also possible)



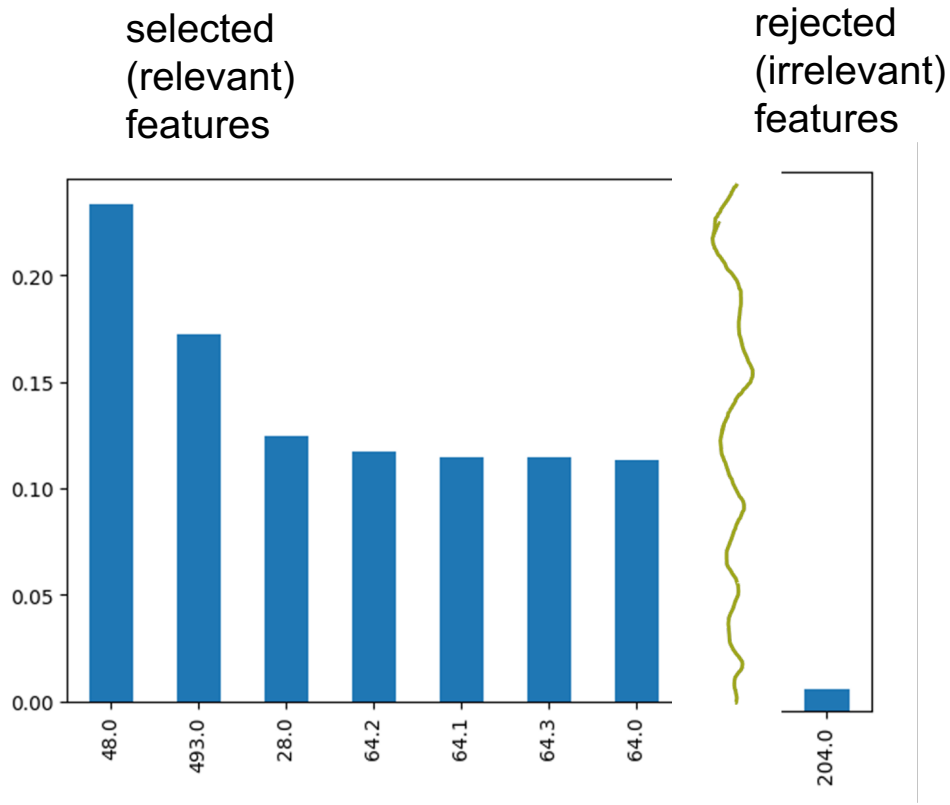
Error probability estimation and feature ranking in All-Relevant Feature Selection based on Random Forest

Andrey Lange, @lange_am

Problem: Supplement the existing All-Relevant FS method with misclassification probability assessment and feature ranking

Tasks:

1. Propose a way of ranking the features by their relevance
2. Propose a way of estimating the probability of misclassification (irrelevant feature selected or relevant rejected)
3. Compare by experiments how your feature ranking/assessment is in line with Boruta's (RF-based) classification



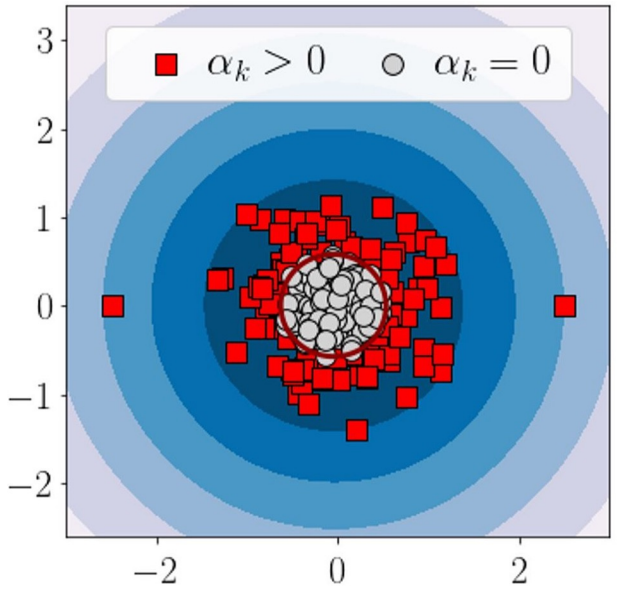
Anomaly detection using One-Class SVM with Vapnik's Privileged Information: find caching rules with best performance

Andrey Lange, @lange_am

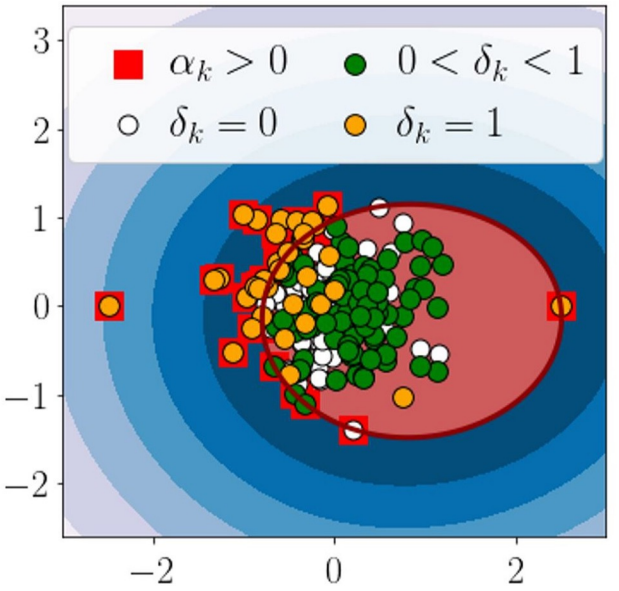
Problem: Find optimal caching rules for Anomaly Detection with Privileged Information.
Research question: does optimal caching rule depend on RAM available?

Tasks:

1. Performance testing of OC-SVM+ with large number of UCI datasets
2. Performance testing of OC-SVM+ with limited RAM
3. Propose new combinations of caching rules and study the performance



v-SVM



v-SVM with a privileged feature

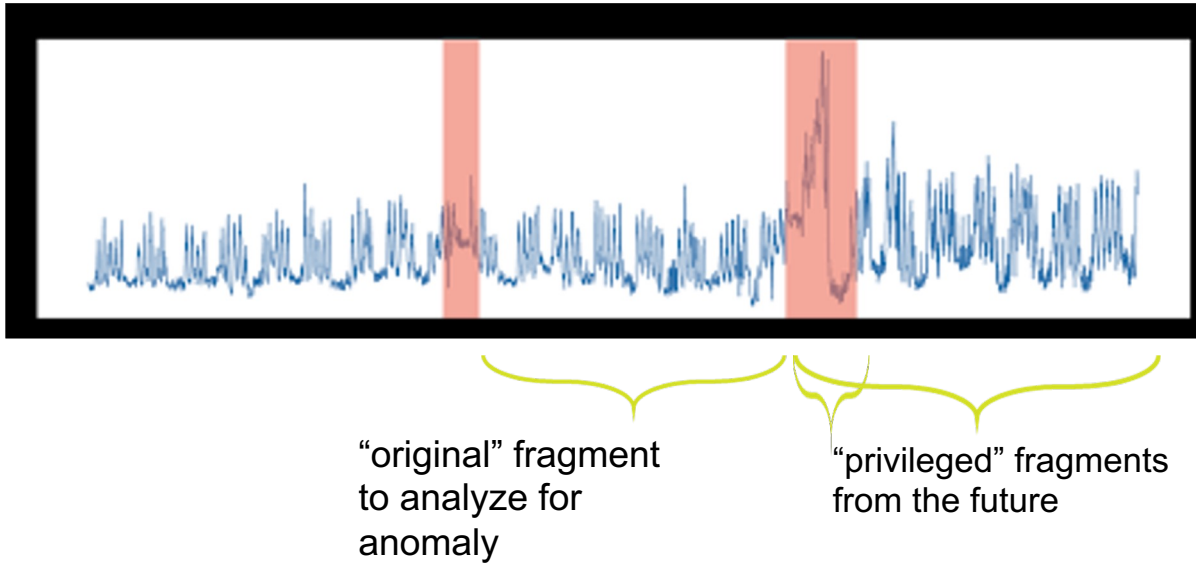
Anomaly detection in time series using One-Class SVM with Vapnik's Privileged Information

Andrey Lange, @lange_am

Problem: Apply OC-SVM (nu-SVM) to time series anomaly detection with "future behavior as Privileged Information".

Tasks:

1. Apply OC-SVM+ to time series anomaly detection.
2. Propose Feature Engineering that builds "original" feature vector based on moments t_{0}, t_{-1}, ..., t_{-n} (now and past) and "privileged" feature vector based on t_{k}, ..., t_{1}, t_{0}, t_{-1}, ..., t_{-n} (i.e. including future).
3. Show the results of experiments: How future (privileged) information improves anomaly detection; How PI + your Feature Engineering improves anomaly detection?



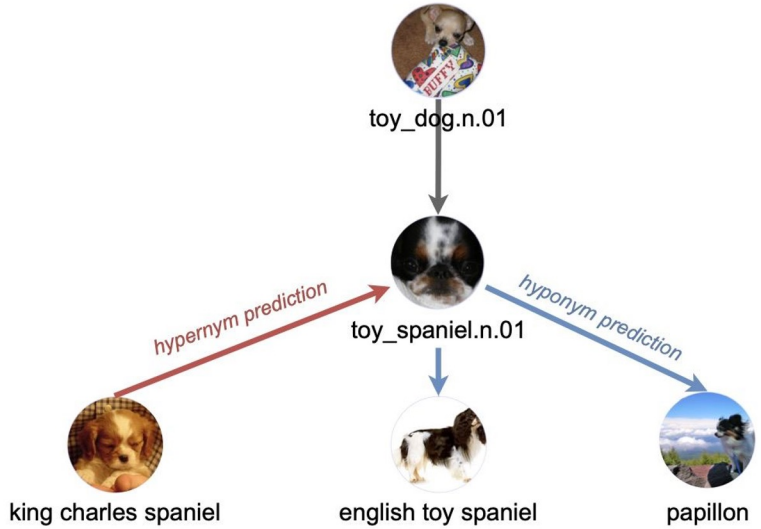
Taxonomy Enrichment via LLM

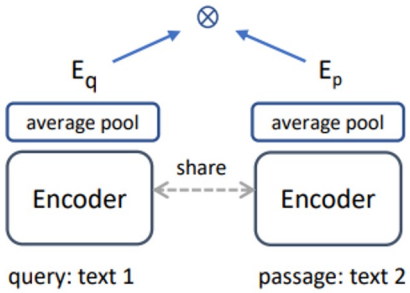
Viktor Moskvoretskii, @VityaVitalichDS

Problem: Solve taxonomy enrichment task with LLM as source of embeddings

Tasks:

1. Research on LLM as embedders
2. Research on how to sample contrastive pairs from DAG
3. Fine-tune LLM with contrastive loss on obtained pairs
4. Evaluate on Taxonomy Enrichment benchmark





$$\min \ \mathbb{L} = -\log \frac{\phi(q^+_{\text{inst}}, d^+)}{\phi(q^+_{\text{inst}}, d^+) + \sum_{n_i \in \mathbb{N}} (\phi(q^+_{\text{inst}}, n_i))}$$

Is classical ML all you need for computer vision?

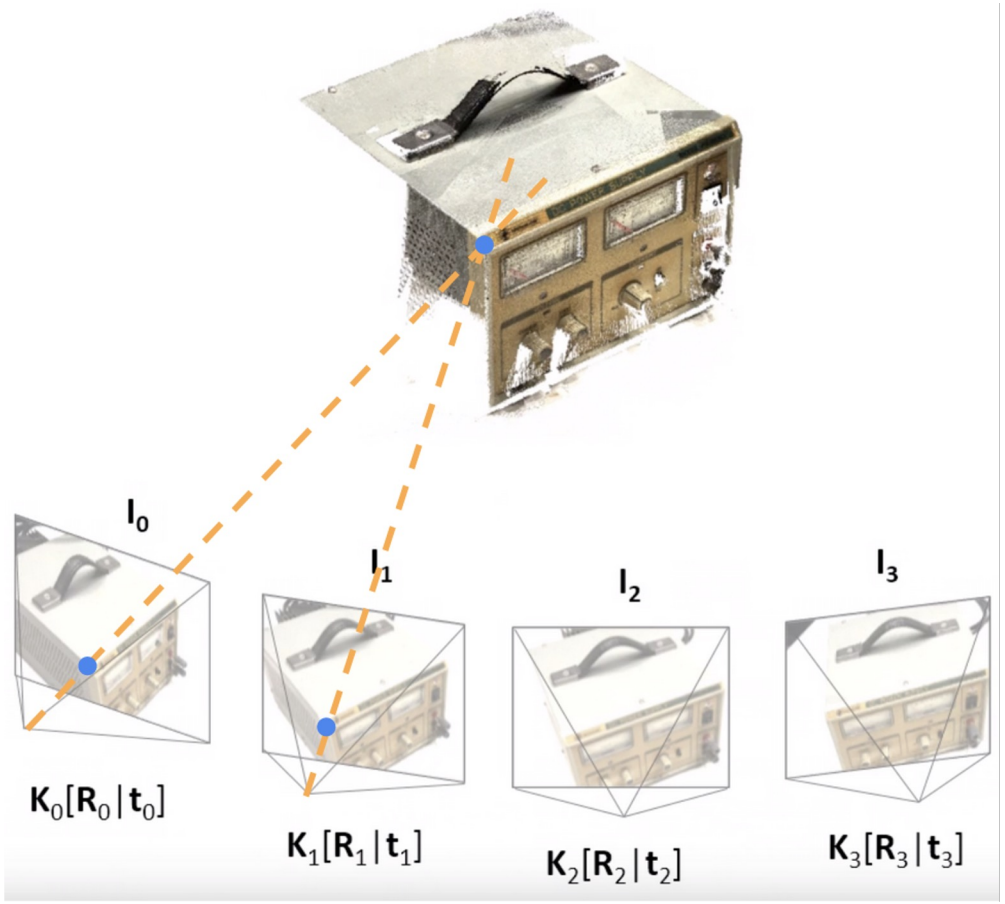
Gleb Bobrovskikh, @BobrG

Problem: Solve a *3D computer vision problem* as good as possible only using the methods of machine learning from this course.

Tasks:

1. Find and review N=(2 * team size) recent works that apply classical machine learning to the problem. (30%)
2. Test N=(team size) methods on some data, e.g the data provided by the authors. (30%)
3. Compare all these methods on the same data. (20%)
4. Propose modifications of these methods aimed to eliminate their drawbacks. (20%)

3D reconstruction from images



Quantization Aware Matrix Factorization for model compression

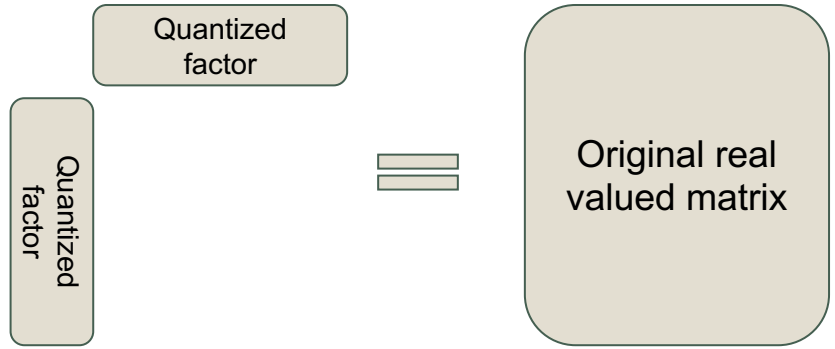
Egor Shvetsov @daLime

Problem: In this project we want to study various possibilities to combine matrix decomposition weight quantization to further enhance model compression approaches.

Tasks:

1. Conduct a literature overview on quantization aware tensor decomposition
2. Implement algorithm described in [1] for a pre-trained models which uses linear layers only. Use at least 3 datasets. It is possible to use some small toy datasests like MNIST
3. Implement at least one algorithm from the literature overview. (we will propose one algorithm if you do not find any)

[1] Quantization Aware Factorization for Deep Neural Network Compression



An example of possible quantization aware factorization where we can choose decomposition rank - **r** and quantization level - **b** to obtain the highest compression rates.

Hard and Rare samples estimation with NN based models

Egor Shvetsov @daLime

Problem: Develop a method which characterizes samples through training procedure without relying on labels. For example, track how representations (embeddings) change over training steps.

Tasks:

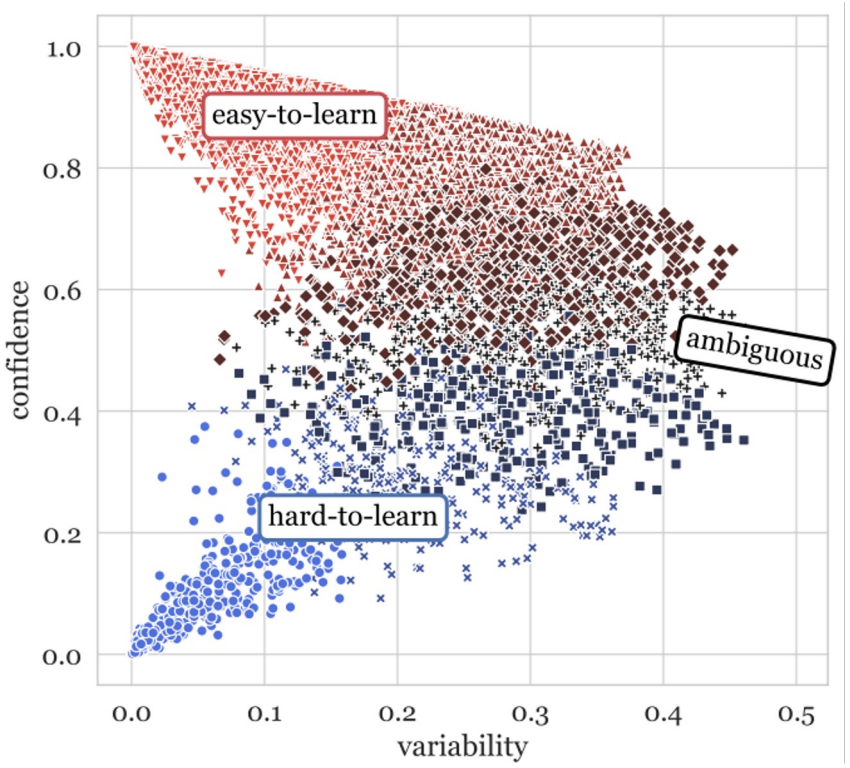
1. Propose a set of metrics to track changes in latent representations during training epochs. One possible approach is to examine the evolution of the mean squared difference between latent representations at each epoch. Do not use labels to obtain data representations.

2. Investigate which of these metrics correlate with labeled data approaches [1,2,3,4]. Train a model using labeled data, apply your proposed metrics, and compare them with label-based metrics from existing papers to determine if there is a correlation.

3. Train a model to generate latent representations in the absence of labels and examine if the hard samples identified using approaches with labels remain the same.

4. Evaluate if these metrics correlate with uncertainty estimation. Explore different approaches to estimate uncertainty and assess if there is a correlation between the proposed metrics and uncertainty estimation.

Regarding datasets, we suggest using MNIST and CIFAR10, but you are not limited to these options.



LPC based voice anonymisation

Egor Shvetsov @daLime

Problem: Implement and study LPC based voice anonymization algorithm

Tasks:

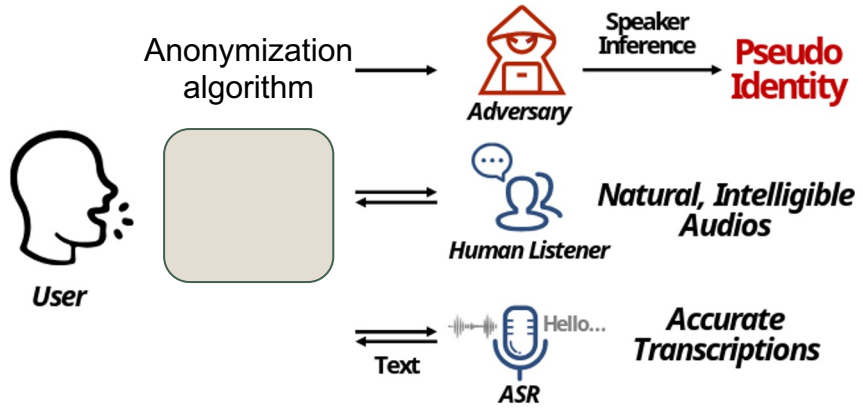
1. Conduct a literature review on voice anonymization techniques, privacy preservation methods, and related research.

2. Reproduce a method mentioned in [1] based on LPC coding.

3. Reproduce any other method from the literature overview.

4. Use any pre-trained Text to Speech models and WER (word to edit distance) as your metric to assess content of anonymized speech.

5. Use cosine similarity between speaker embeddings to verify that the voice is anonymized, to obtain embeddings is possible to use [2] or any other model.



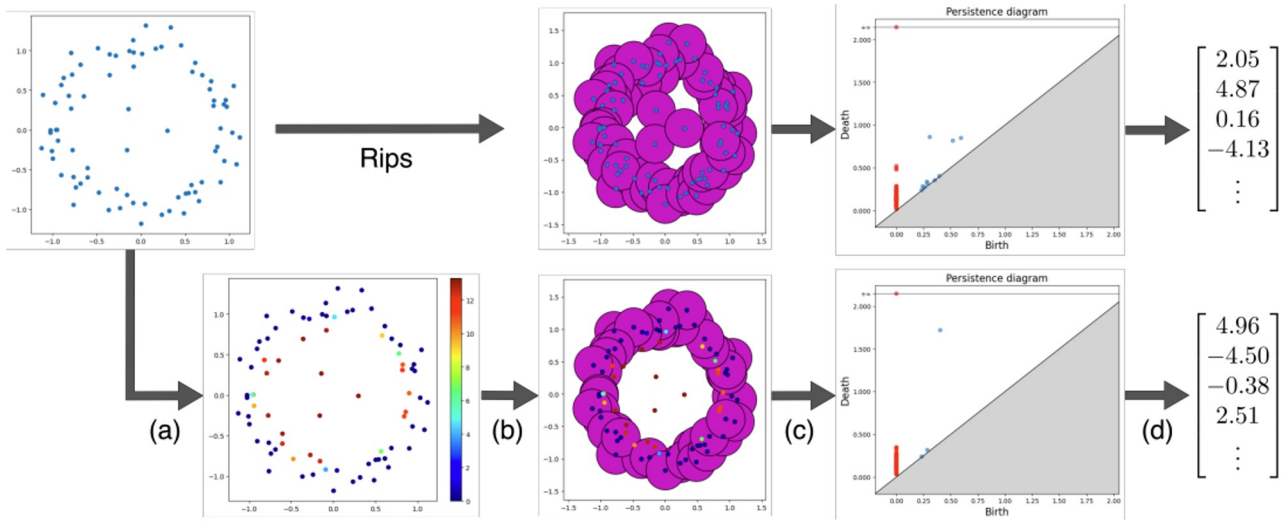
Adaptive Topological Features via Persistent Homology: Filtration Learning for Point Clouds

Anton Dmitriev @touhoutoha

Problem: Test a novel method of learning adaptive filtration and compare it with the classical method of persistent barcode computation

Tasks:

1. Implement method and run it on synthetic datasets from paper
2. Run at least one non-synthetic experiment from paper
3. Run some experiments on the text domain (for example try to integrate it with some transformer-based classifier)



Dataset:

3D-CAD dataset

Text classification datasets

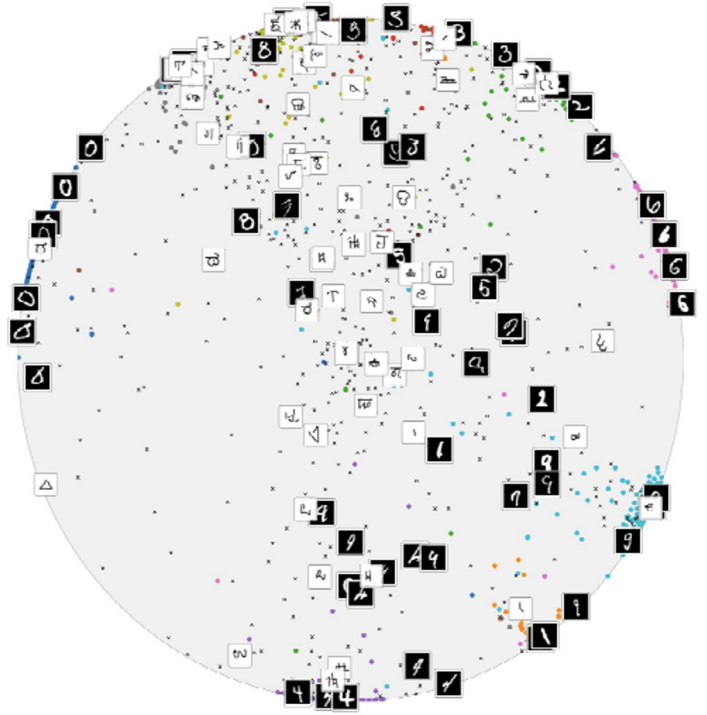
Hyperbolic Image Embeddings

Anton Dmitriev, @touhoutoha

Problem: Most of the method of image processing rely on linear hyperplanes for image classification. Maybe there is something better?

Tasks:

1. Run the method from the paper on MNIST or another toy dataset
2. Run the method on the more complex data (like AFHQ) and compare it with a couple of euclidean models with the same size
3. Try to attach riemannian training for hyperbolic neural network (extra)



Text clusters changing via time

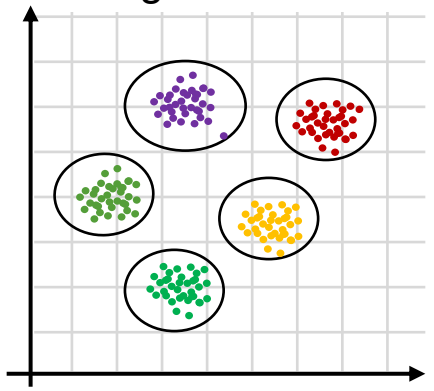
Petr Sokerin, TA, @psokerin

Problem: business clusters texts and after gets new data objects that are needed to be associated with some cluster and find new clusters

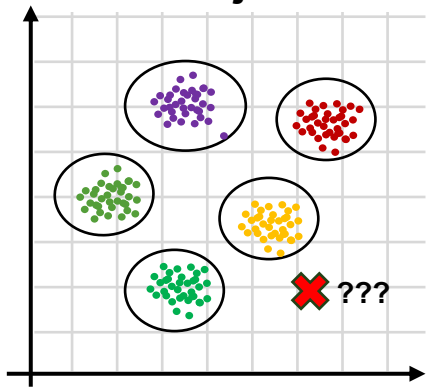
Tasks:

1. propose a heuristic for prediction new objects for a cluster
2. implement and test realization of the prediction method
3. propose a heuristic for new cluster of new objects estimation
4. implement and test realization of the prediction new cluster estimation

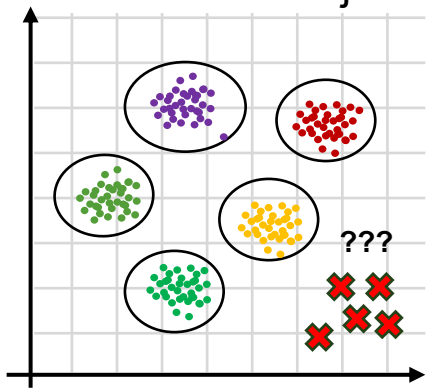
1. **Clustering** original data



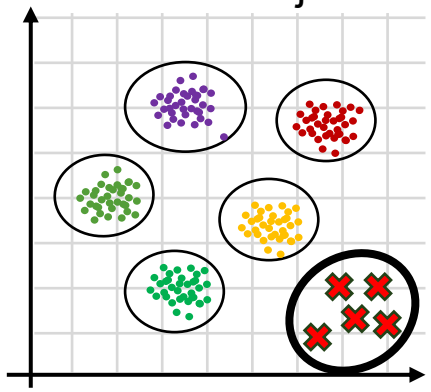
2. Predict cluster for **new object**



3. **Implement method predict** for new objects



4. **Detect new cluster** form new objects



Robustness of time-series models to adversarial attacks

Petr Sokerin, TA, @psokerin

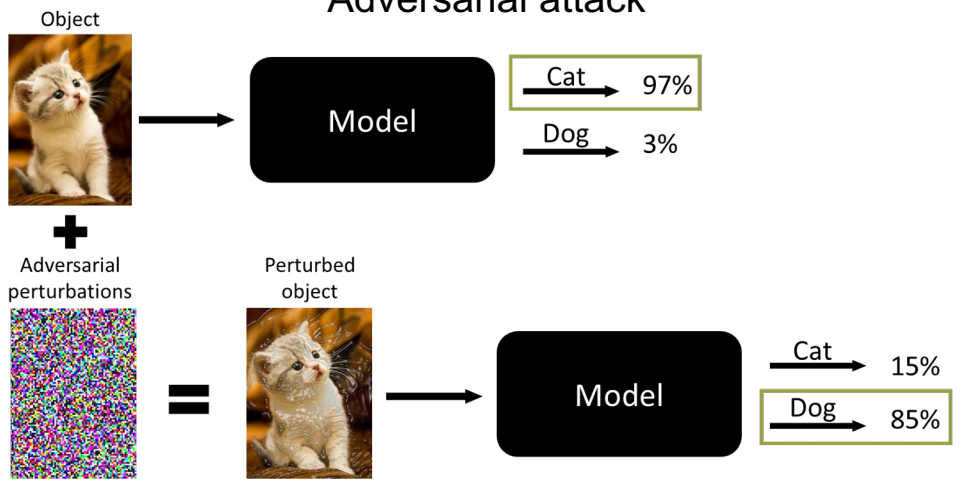
Problem: Different time series neural networks have different robustness to adversarial attacks

Adversarial attacks are methods for decreasing machine learning models' quality of prediction.

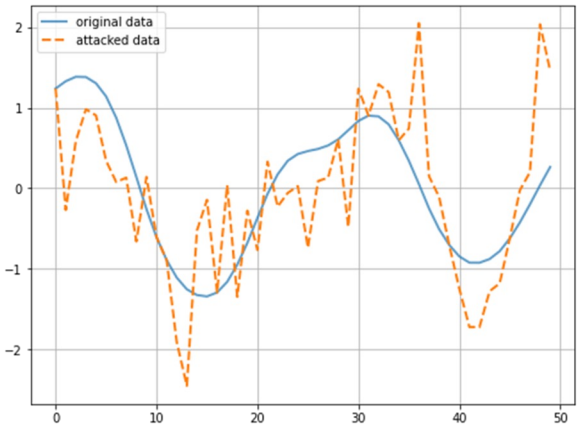
Tasks:

1. train 2-3 time series neural networks with different architectures: LSTM, CNN, Transformer (optional) FordA dataset classification
2. implement adversarial attacks IFGSM, DeepFool, SimBA
3. run experiments with adversarial attacks and compare the level of attacked model metrics decrease

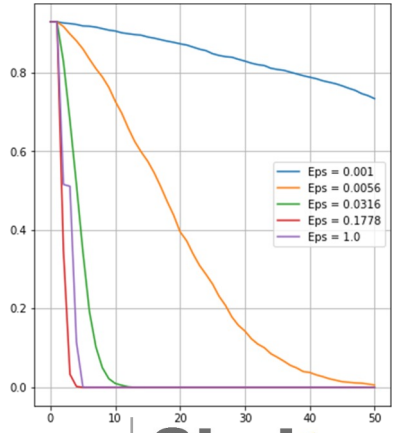
Adversarial attack



Adversarial attack data on time series domain



Examples of decreasing quality of TS2Vec model



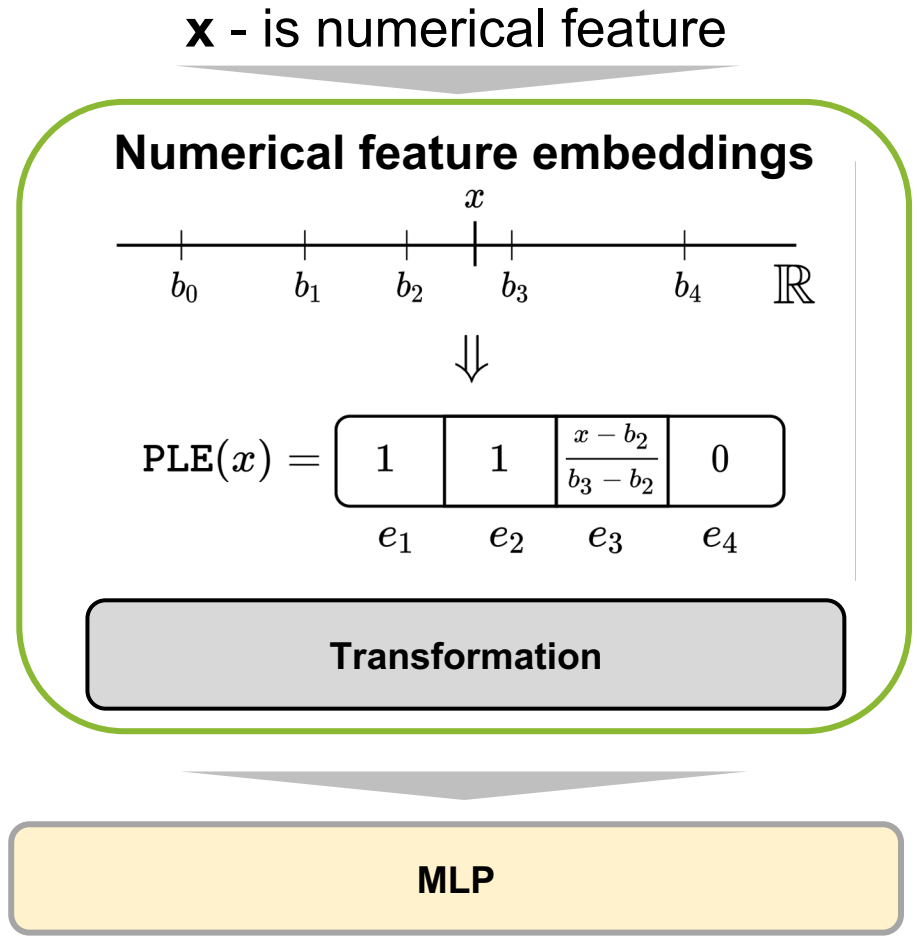
On Embeddings for Numerical Features in Tabular Deep Learning

Petr Sokerin, TA, @psokerin

Problem: Compare neural networks for tabular data with classic machine learning methods

Tasks:

1. train a neural network for numerical embeddings on datasets from paper (at least 4) and estimate results
2. train classic machine learning models on original datasets: CatBoost, XGBoost, LGBM, Random Forest, kernel linear regression.
3. train classic machine learning models on numerical embeddings from paper: CatBoost, XGBoost, LGBM, Random Forest, kernel linear regression.
4. Tune parameters and estimate results



Boosting approaches with multi-label imbalanced data problem

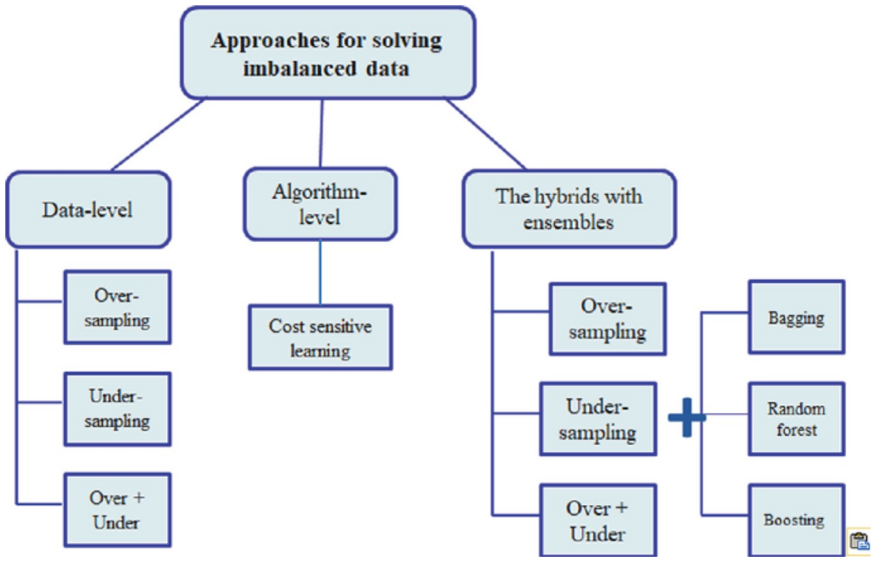
Diana Koldasbayeva, @huanpops

Problem: Enhance the boosting algorithm for more effective performance in multi-label classification tasks

Main Tasks:

3. Try to improve boosting algorithms:
    3.1. explore the impact of another base learners under the boosting procedure;
    3.2. study how preprocessing operators affect the performance of the boosting algorithms;
    3.3. use combined ensemble techniques with using data-level approach resampling methods (for example TomekLinks, Adasyn, Smote etc)

- Dataset $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$, where $\mathbf{x}_i \in \mathbb{R}^d$, $y_i \in \{-1, +1\}$
- $C_{+1}(S) = \{(\mathbf{x}_i, y_i) \in S \mid y_i = +1\}$ is a major class,
- $C_{-1}(S) = \{(\mathbf{x}_i, y_i) \in S \mid y_i = -1\}$ is a minor class, i.e. $|C_{+1}(S)| > |C_{-1}(S)|$
- Imbalance ratio $IR(S) = \frac{|C_{-1}(S)|}{|C_{+1}(S)|}$, $IR(S) \leq 1$



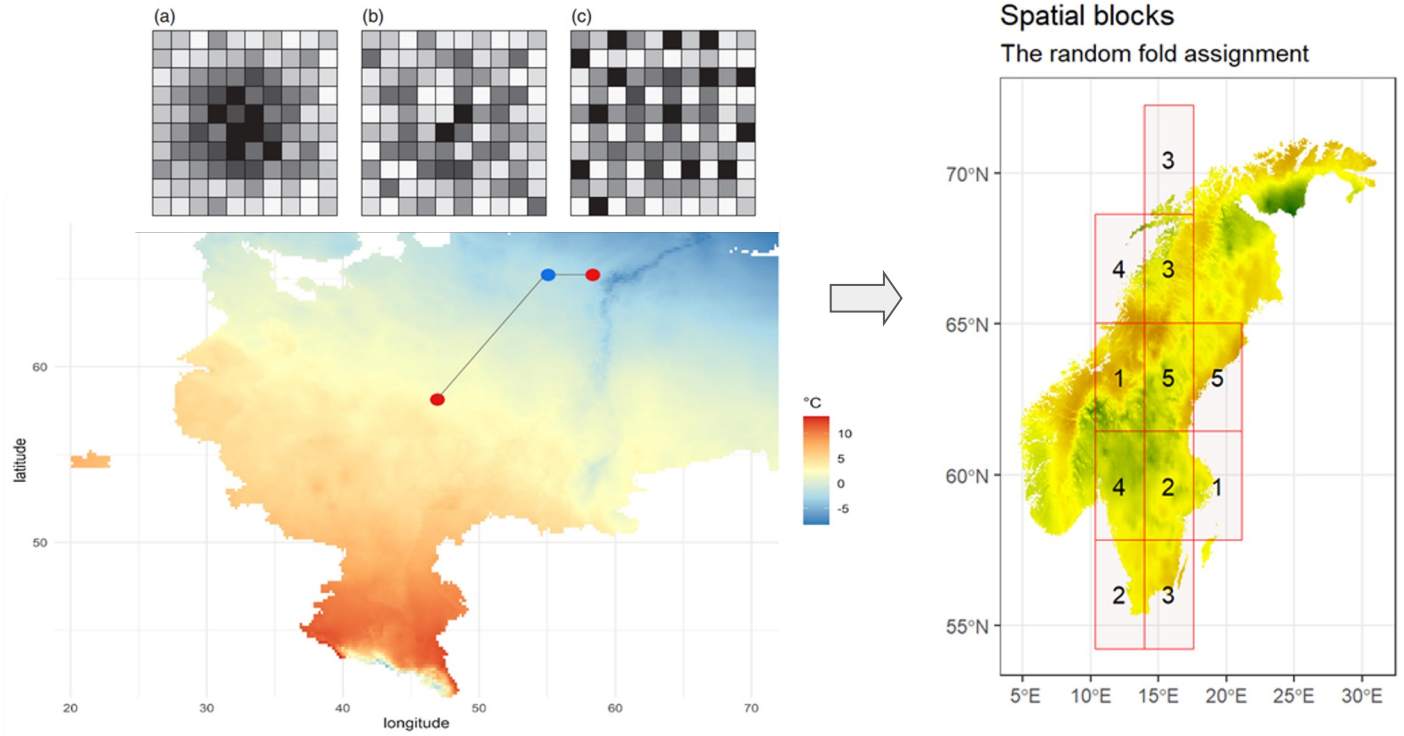
Application of spatial cross-validation to mitigate Spatial Autocorrelation (SAC) problem

Diana Koldasbayeva, @huanpops

Problem: The goal is to estimate spatial models prediction performance due to spatial cross validation (CV) methods accounted for SAC

Tasks:

1. Study the problem SAC and methods of CV which are able to solve it;
2. Perform spatial, environmental and buffered-LOO types of CV;
3. Use other 1-2 methods to solve SAC problem;
4. Implement your own method to solve this problem



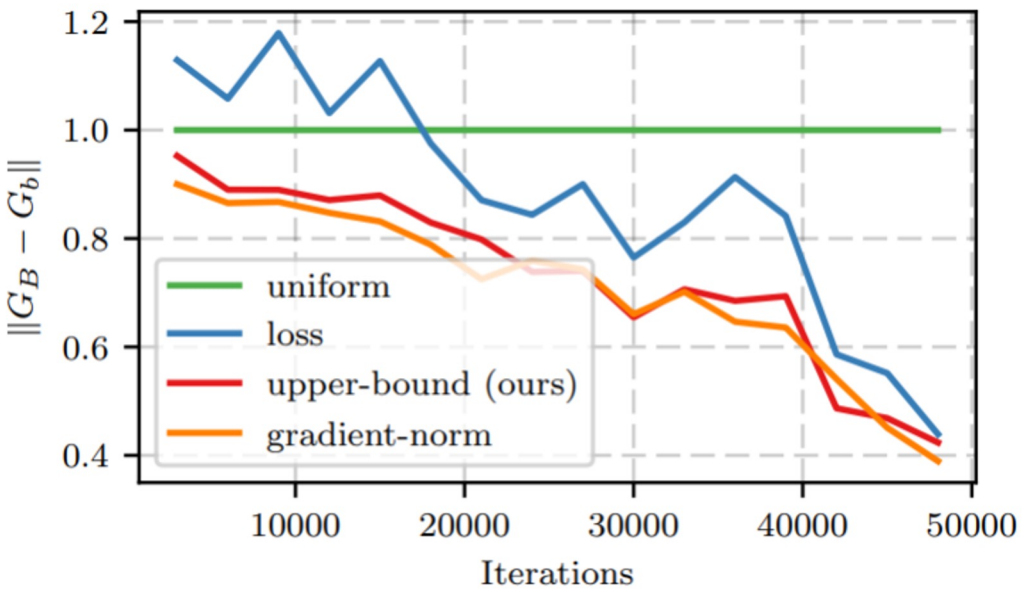
Importance Sampling strategies for accelerating Deep Neural Network training

Diana Koldasbayeva, @huanpops

Problem: Apply an Importance Sampling (IS) scheme designed to expedite the training process of various neural network architectures

Tasks:

2. investigate how the proposed importance sampling scheme can be integrated into a standard SGD procedure: utilize a specific technique of IS and minimize the time or iterations required for convergence during training:
2.1 use upper-bound method and loss-based method from paper,
2.2 compare them with other possible methods of IS



Application of Graph Neural Network to spatial time series forecasting problem

Polina Pilyugina, @Fjosp

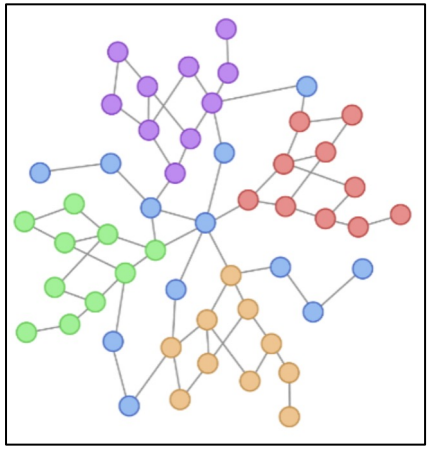
Problem: Investigate applicability of Graph Neural Networks for spatial time series forecasting to the problem of ATM transactions prediction

Tasks:

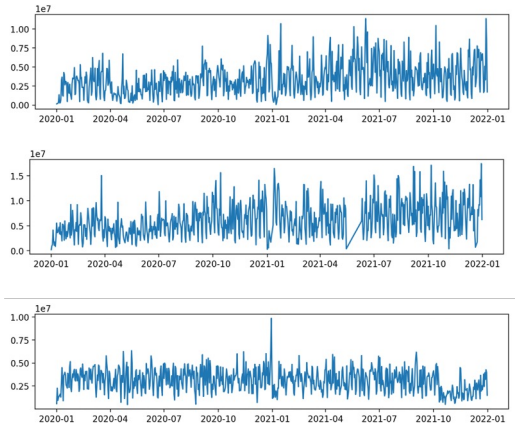
1. Select models to use (3-4) and preprocess the dataset to fit the models. (20%)
2. Run models on the dataset with and without tuning hyperparameters. (40%)
3. Run baseline models to compare. (10%)
4. Analyse the results and propose next research steps (30%)

Dataset:

2.5M Danish ATM Transactions from 2017



Graph of connections between ATMs



Time series of targets on ATMs

Meta-learning for time series forecasting in application to ATM load time series

Polina Pilyugina, @Fjosp

Problem: Build the meta-learning approach which will outperform the brute-force baseline and allow to improve the computational time without loss of prediction quality

Tasks:

1. List of features to be extracted from time series (0 points will be given for only using tsfresh features) and algorithm to extract those. (30%)
2. Meta learning algorithm including full validation set up to test the algorithm. (30%)
3. Analysis of the features contribution/ importance for meta-learning algorithm. (20%)
4. Algorithm applied to the dataset with results showing improvement over the baseline. (20%)

RMSSE	MAPE	SMAPE	naming_orig	model_name	dataset_name	horizon	split
0.7920865617973475	inf	16.386863589286804	danish_atm_daily_5	CatBoostAutoRegressivePipelineEtna_3lags_gl	danish_atm_daily	30	test
0.9180722190102938	41.1573350429534	17.157183587551117	danish_atm_daily_84	CatBoostAutoRegressivePipelineEtna_3lags_gl	danish_atm_daily	30	validation
0.7577647505260843	39.02260363101959	13.038493692874908	danish_atm_daily_32	CatBoostAutoRegressivePipelineEtna_3lags_gl	danish_atm_daily	30	test
0.7929251887194474	57.07847476005554	18.962322175502777	danish_atm_daily_25	CatBoostAutoRegressivePipelineEtna_3lags_gl	danish_atm_daily	30	validation
0.5976039011395582	21.996028721332555	10.027625411748886	danish_atm_daily_6	CatBoostAutoRegressivePipelineEtna_3lags_gl	danish_atm_daily	30	validation
0.8174417324237805	37.2047483921051	15.833663940429688	danish_atm_daily_10	CatBoostAutoRegressivePipelineEtna_3lags_gl	danish_atm_daily	30	validation
1.26363560651282	inf	23.24129045009613	danish_atm_daily_102	CatBoostAutoRegressivePipelineEtna_3lags_gl	danish_atm_daily	30	test
2.136854356103441	inf	100.0	danish_atm_daily_87	CatBoostAutoRegressivePipelineEtna_3lags_gl	danish_atm_daily	30	validation
0.8733682609771917	inf	22.032369673252106	danish_atm_daily_16	CatBoostAutoRegressivePipelineEtna_3lags_gl	danish_atm_daily	30	validation
0.8960758739205139	inf	12.28131577372551	danish_atm_daily_73	CatBoostAutoRegressivePipelineEtna_3lags_gl	danish_atm_daily	30	validation
0.7503695733536034	28.2947838306427	12.60123997926712	danish_atm_daily_43	CatBoostAutoRegressivePipelineEtna_3lags_gl	danish_atm_daily	30	validation
1.003732508636627	inf	28.42548191547393	danish_atm_daily_39	CatBoostAutoRegressivePipelineEtna_3lags_gl	danish_atm_daily	30	test
0.9633038499845402	64.20406699180603	20.97110897302628	danish_atm_daily_83	CatBoostAutoRegressivePipelineEtna_3lags_gl	danish_atm_daily	30	validation
0.9356655393429324	35.556620359420776	13.525134325027466	danish_atm_daily_18	CatBoostAutoRegressivePipelineEtna_3lags_gl	danish_atm_daily	30	validation
1.1677613474881527	inf	48.93269836902618	danish_atm_daily_99	CatBoostAutoRegressivePipelineEtna_3lags_gl	danish_atm_daily	30	test
2.026379373757171	inf	100.0	danish_atm_daily_80	CatBoostAutoRegressivePipelineEtna_3lags_gl	danish_atm_daily	30	validation
1.0741981406134198	inf	46.91608548164368	danish_atm_daily_103	CatBoostAutoRegressivePipelineEtna_3lags_gl	danish_atm_daily	30	test
1.0488583337931074	84.15878415107727	21.52030318975449	danish_atm_daily_38	CatBoostAutoRegressivePipelineEtna_3lags_gl	danish_atm_daily	30	test
0.8421458883396552	24.4303897023201	11.759278178215029	danish_atm_daily_28	CatBoostAutoRegressivePipelineEtna_3lags_gl	danish_atm_daily	30	validation
0.6967982666297485	67.88281798362732	23.152069747447968	danish_atm_daily_71	CatBoostAutoRegressivePipelineEtna_3lags_gl	danish_atm_daily	30	validation
0.9115530642864844	40.92070162296295	13.408422470092772	danish_atm_daily_59	CatBoostAutoRegressivePipelineEtna_3lags_gl	danish_atm_daily	30	test
1.0532447431625092	inf	33.002135157585144	danish_atm_daily_35	CatBoostAutoRegressivePipelineEtna_3lags_gl	danish_atm_daily	30	test
0.9622246646331531	70.49328684806824	18.47943812608719	danish_atm_daily_43	CatBoostAutoRegressivePipelineEtna_3lags_gl	danish_atm_daily	30	test
0.7810168504572181	25.06163418293	11.441988497972488	danish_atm_daily_64	CatBoostAutoRegressivePipelineEtna_3lags_gl	danish_atm_daily	30	validation

Results of time series forecasting on multiple datasets by various models

Meta-Learning algorithm based on time series characteristics

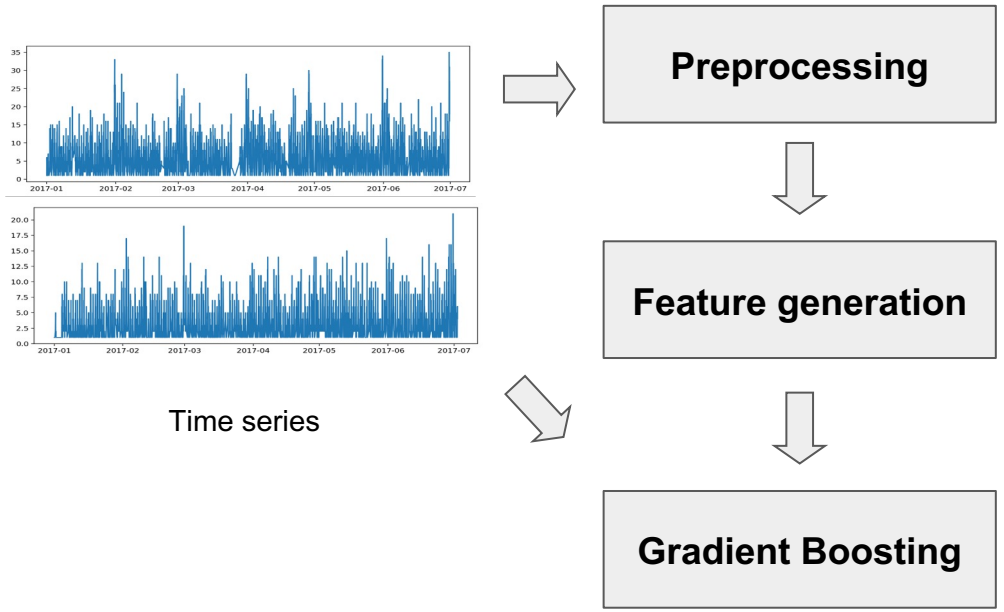
Feature improvement of Gradient Boosting application for time series forecasting

Polina Pilyugina, @Fjosp

Problem: Come up with feature generation and time series preprocessing algorithms to improve quality of Gradient Boosting on the task of time series forecasting on a given ATM transactions dataset

Tasks:

1. Feature generation pipeline is proposed. (20%)
2. Preprocessing steps are proposed. (20%)
3. Clear connection between Gradient Boosting parameters and preprocessing steps is established and explained. (30%)
4. Proposed pipeline is outperforming the "baseline' defined Boosting run on the given time series dataset (30%)



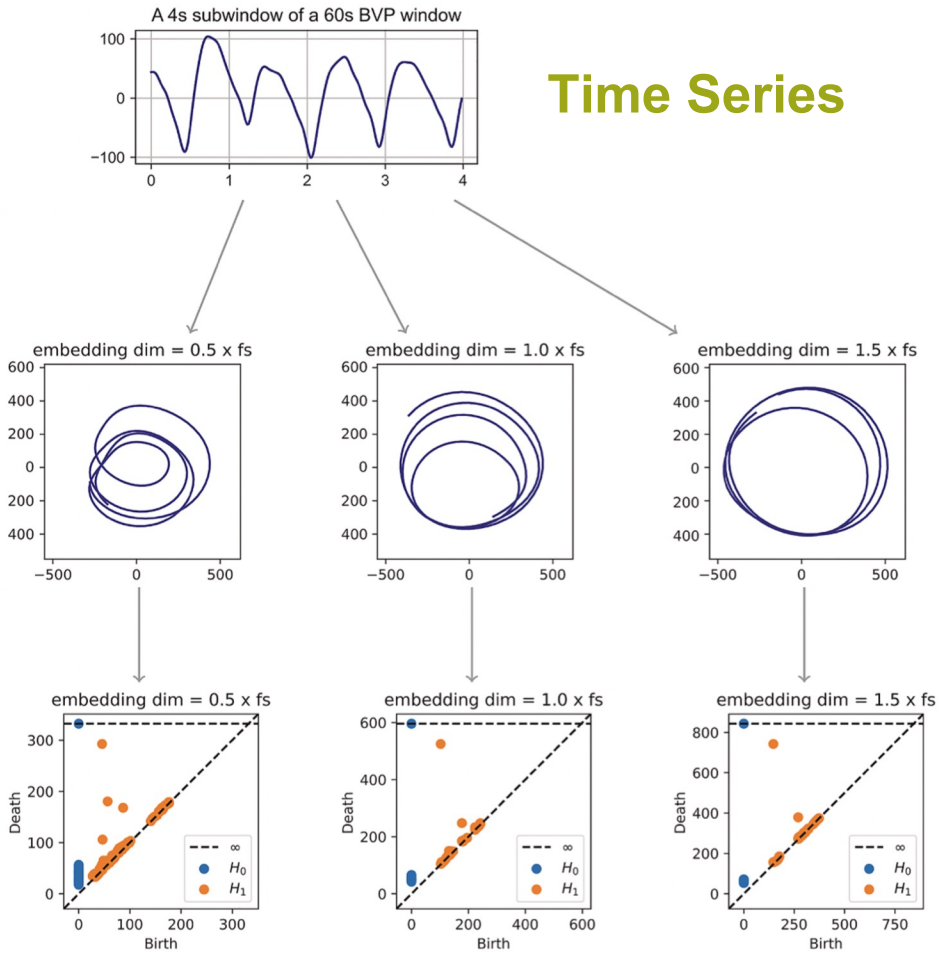
Topological Data Analysis for Univariate Time Series Classification

Kristian Kuznetsov, @pyashy

Problem: Analyze the applicability of features from persistence barcodes (diagrams) to the selected domain of univariate time series

Tasks:

1. Choose the domain of univariate time series for classification
2. Extract topological embedding using TDA
3. Try to solve the task using these features and different ML algorithms
4. Assess the feature importance
5. Try different preprocessing steps for time-series and different settings to time-delayed embeddings



Analysing Rashomon Importance Distribution

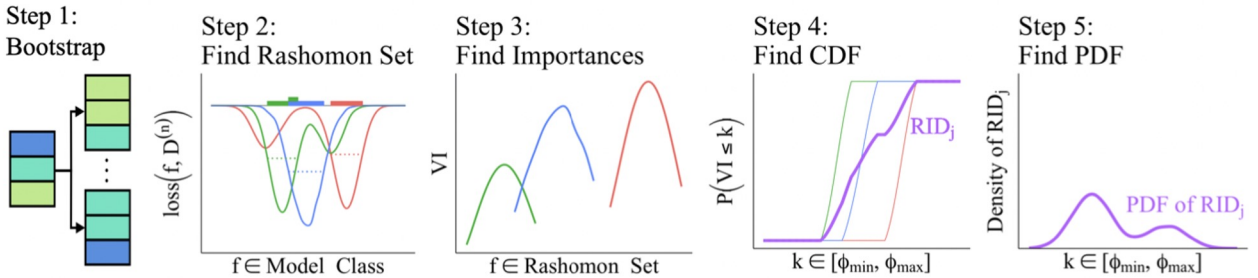
Kristian Kuznetsov, @pyashy

Problem: Replicate the results in [1] and provide the feature importance analysis by baseline and RID methods on new dataset

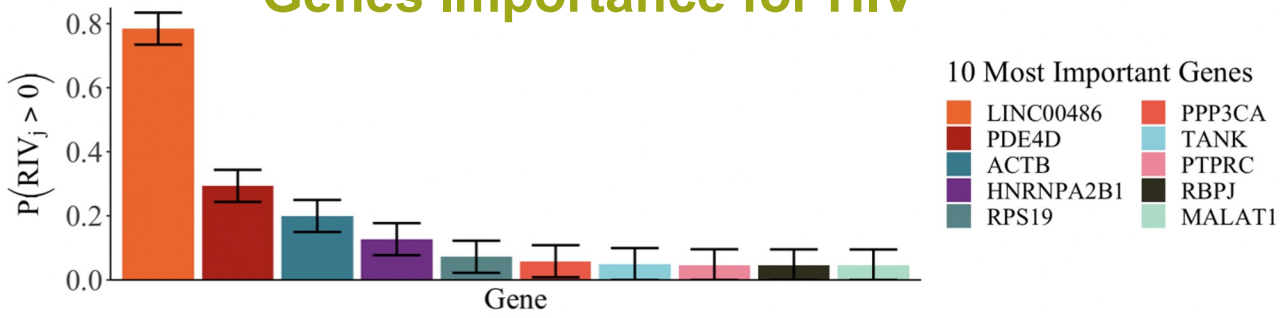
Tasks:

1. Investigate RID approach by [1]
2. Replicate results on synthetic tasks using their code
3. Choose the real-world dataset(s) to analyse variable importance by RID and provide baseline methods and RID analysis
4. Analyse whether the RID approach allows to improve the end task quality

RID framework



Genes Importance for HIV



[1] Donnelly, Jon, et al. "The Rashomon Importance Distribution: Getting RID of Unstable, Single Model-based Variable Importance." *Thirty-seventh Conference on Neural Information Processing Systems* (2023)

Analysing the Importance of LLMs Embeddings' Components on Probing Linguistic Tasks

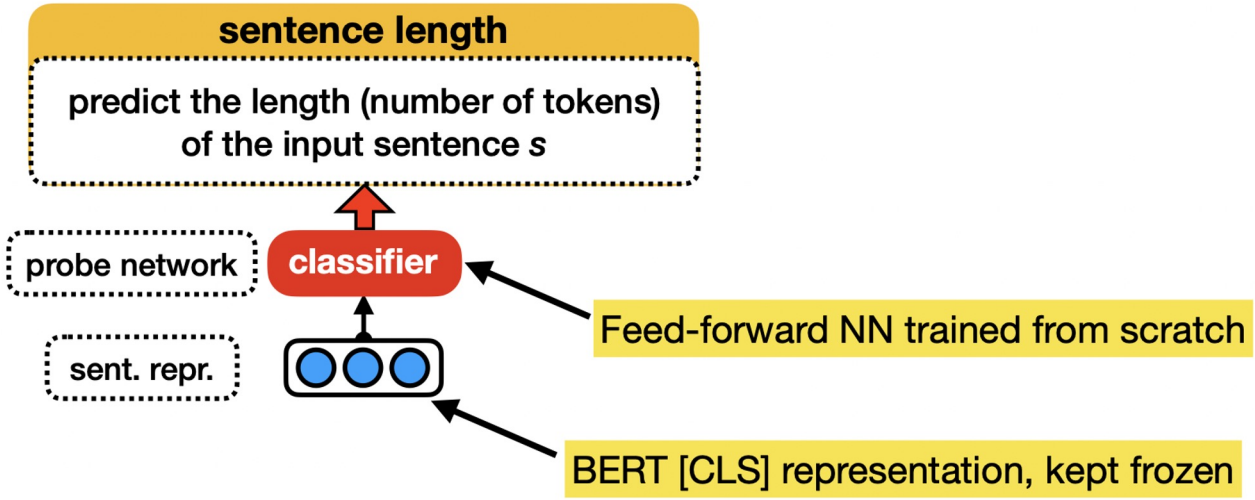
Kristian Kuznetsov, @pyashy

Problem: perform probing of embeddings from LLM on some set of probing tasks and analyse the feature importance

Tasks:

1. For given LLM get the pooled embeddings for probing tasks and achieve the baseline scores.
2. Try different classical ml models on embeddings (logreg, boosting, etc)
3. Perform the feature importance analysis by different methods and compare them to outlier dimensions [1]

Example of probing task



[1] Kovaleva, Olga, et al. "BERT busters: Outlier dimensions that disrupt transformers." *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021* (2021)

Topological Data Analysis (TDA) for Travelling Salesman Problem (TSP)

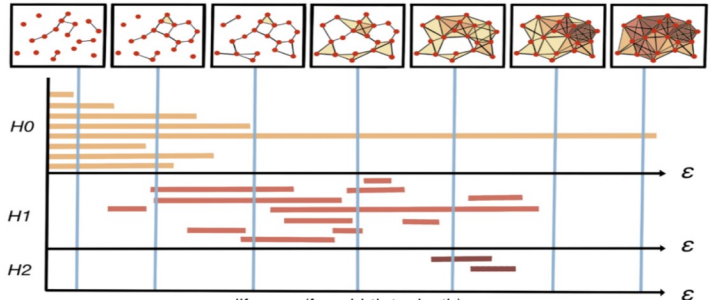
Serguei Barannikov, @serguei_barannikov

Problem: Topological Data Analysis (TDA) for Travelling Salesman Problem (TSP)

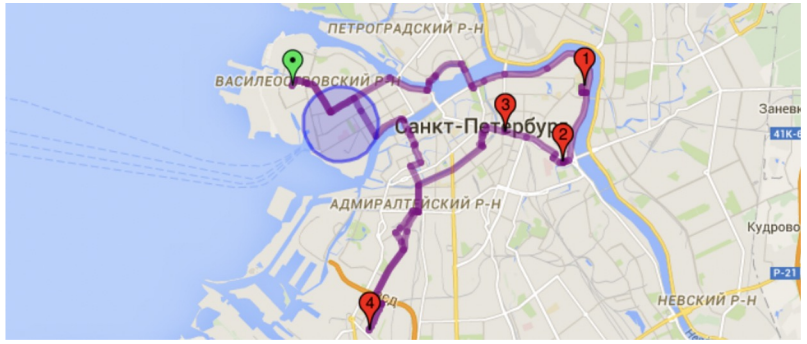
Tasks:

1. Literature Review
2. Data Preparation
3. Develop a framework for applying TDA to TSP, focusing on extracting topological features.
1. Hypothesize how topological features might correlate with the effectiveness of TSP solutions.
2. Algorithm Development: Based on insights gained, develop or modify existing TSP algorithms or heuristics to leverage identified topological features.

Topological Data Analysis (TDA)



Travelling Salesman Problem



[1] Attention, Learn to Solve Routing Problems arXiv:1803.08475
[2] The Transformer Network for the Traveling Salesman Problem arXiv:2103.03012
[3] Neural TSP Solver with Progressive Distillation