My first scientific paper

Week 2

Select your project and tell about it

Vadim Strijov

Moscow Institute of Physics and Technology

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!

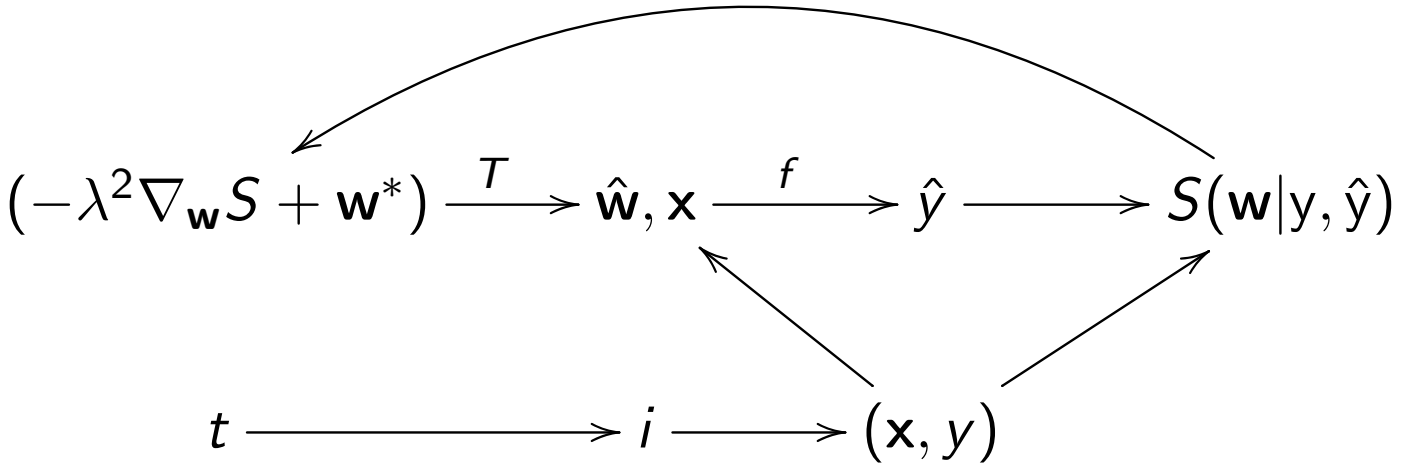
... it was a banking credit scoring model

The main questions to investigate

How to

1. construct a stable and precise machine learning model?
2. select an optimal structure from a wide class of deep learning models?
3. combine local approximation models into ensemble?

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,

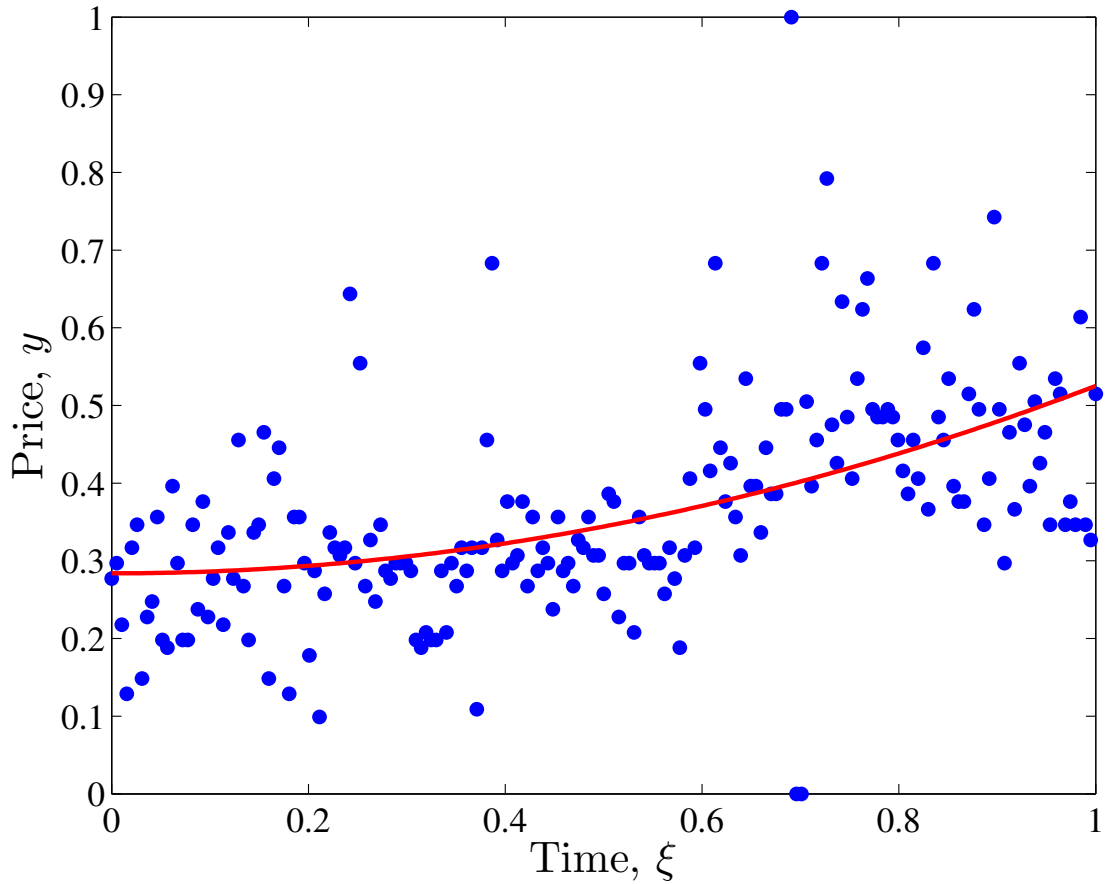
$$\hat{\mathbf{w}} = \arg\min S(\mathbf{w}|y, f).$$

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

[1] $*$

Model and its structure $\mathbf{a} \in \mathbb{B}^n$

Regression model: $f = w_1 + w_2\xi^1 + w_3\xi^2 + \varepsilon(\xi)$



features: $\mathbf{x} = [\xi^0, \xi^1, \xi^2]^\mathsf{T}$

model to select from: $f = \mathbf{a} \odot \mathbf{w}^\mathsf{T}\mathbf{x}$

optimal structure: $\hat{\mathbf{a}} = [1, \quad 0, \quad 1]^\mathsf{T}$

optimal parameters: $\hat{\mathbf{w}} = [0.2839, \text{n/a}, 0.2412]^\mathsf{T}$

Model families

A model is a parametric family of functions,

$$\hat{y} = f(\hat{\mathbf{w}}, \mathbf{x}),$$

an element of a model family, given by some superposition,

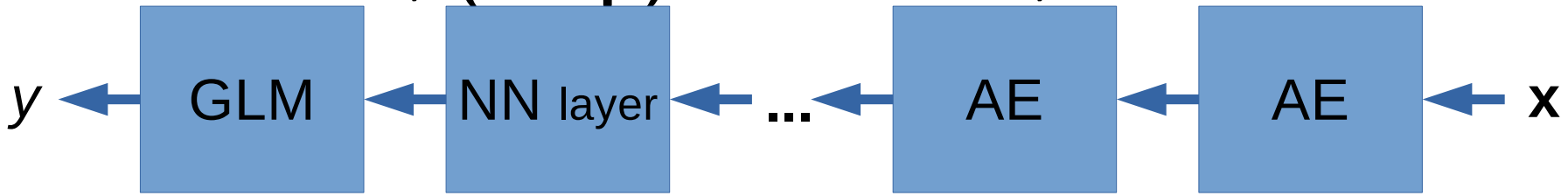
$$f = g_K \circ \cdots \circ g_1(\mathbf{w})(\mathbf{x}) \ni \mathfrak{F}.$$

An example is a superposition of linear maps (transformations) and non-linear monotonous (smooth) functions:

$$f(\mathbf{w}, \mathbf{x}) = \sigma_K \circ \mathbf{w}_K^{\mathsf{T}} \boldsymbol{\sigma}_{k-1} \circ \cdots \circ \boldsymbol{\sigma}_1 \mathbf{W}_1^{\mathsf{T}} \mathbf{x}.$$

The model parameters are treated as $\mathbf{w} = \mathbf{vec}(\mathbf{w}_K, \ldots, \mathbf{W}_1)$

Linear model, (deep) neural net, and autoencoder



$$f = \sigma_k \circ \underset{1\times 1}{\mathsf{w}_k^{\mathsf{T}}}\boldsymbol{\sigma}_{k-1} \circ \mathsf{W}_{k-1}\boldsymbol{\sigma}_{k-2} \circ \cdots \circ \underbrace{\underbrace{\underset{n_2\times 1}{\mathsf{W}_2}\boldsymbol{\sigma}_1 \circ \underset{n_1\times n}{\mathsf{W}_1}\underset{n\times 1}{\mathsf{x}}}_{E_{\mathsf{x}}=\sum_{\mathsf{x}_i\in\mathfrak{D}}\|\mathsf{x}_i-\mathsf{r}(\mathsf{x}_i)\|_2^2}}_{E_D=\sum_{(\mathsf{x}_i,y_i)\in\mathfrak{D}}\left(y_i-f(\mathsf{x}_i)\right)^2} \in \mathfrak{D}$$

$$S = \lambda_1 E_D + \lambda_2 E_{\mathsf{x}} + \lambda_3 E_{\mathsf{w}} = \boldsymbol{\lambda}^{\mathsf{T}}\mathsf{s}$$

E_{w} is some regularisation error, for

principal component analysis: $\mathsf{W}^{\mathsf{T}}\mathsf{W} = \mathsf{I}_n$,

skip block: $\mathsf{W} = \mathsf{I}_n$, $\sigma = \mathrm{id}$,

classification: $\sigma \in \{\text{logistic}, \text{softmax}, \text{ReLu}, \ldots\}$.

... including LM, LR, PCA, AE, SAE, 2NN, DLL, CNN, etc.

Quality criteria for model generation and selection

Three sources of quality criteria

1. Business: model operation productivity, agent impact to environment
2. Theory: statistical hypothesis, bayesian inference
3. Technology: optimization requirements, resources

The main criteria of model quality

- Precision: MAPE, AUC
- Stability (diversity): std deviation for prediction, covariance of parameters
- Complexity: structure complexity, MDL, evidence of model

Go m1p.org

Цели исследования

Цель работы

Предложить метод отбора признаков, учитывающий взаимное расположение признаков и целевого вектора.

Проблема

Методы отбора признаков дают избыточное подмножество мультикоррелирующих признаков.

Метод решения

Использование постановки задачи квадратичного программирования для получение оптимального подмножества признаков.

Проблема

- Тематические модели *неполны* и *неустойчивы*.
- Получение хорошей тематической модели, как правило, требует больших затрат времени.
- Не существует идеального автоматического способа оценивания качества тематических моделей.

Решение

Банк тем — инструмент для сохранения интерпретируемых тем, построенных при многократных запусках, с целью последующего их использования для оценки качества моделей.

Цели

Реализовать метод построения банка тем и оценивания качества тематических моделей с помощью банка тем.

Задача кластеризации точек временного ряда

Цель: предложить алгоритм поиска характерных квазипериодических сегментов внутри временного ряда, полученных при помощи мобильного акселерометра.

Задачи

1. Предложить признаковое описание точек временного ряда.
2. Предложить функцию расстояния между точками временного ряда в новом признаковом описании, для их дальнейшей кластеризации.

Исследуемая проблема

1. Понижение размерности пространства признаков. Понстроение признакового описания точек временного ряда.

Метод решения

Алгоритм поиска характерных сегментов основывается на методе главных компонент для локального снижения размерности сегмента фазовой траектории в окрестности каждой точки временного ряда. Главные компоненты рассматриваются как признаковое описания точек временного ряда.

Графовые нейронные сети в задаче синтеза молекул

Требуется

Построить модель предсказания молекулярного графа основного продукта химической реакции по графам исходных веществ.
На модель накладываются ограничения:

- применима к данным в виде несвязанного молекулярного графа;
- допускает использования экспертных знаний о локальной структуре молекулярного графа;

Проблема

Пространство молекулярных структур высоко-размерное. Количество механизмов реакций растет с ростом числа известных структур.

Метод

Графовая нейронная сеть, допускающая использование экспертных знаний о структуре молекулярного графа.

Прикладная задача

Значимость

Предлагаемый подход предназначен для улучшения систем информационного поиска, основанных на экспертных оценках релевантности документа запросам.

Коллекции документов

Следуя традициям сообщества ИП, мы ставим своей целью построение ранжирующих функций, дающих высокий MAP на коллекциях TREC.

Актуальность

Постоянное развитие TREC-сообщества, программных пакетов, связанных в т.ч. с ранжирующими функциями (напр. Terrier) демонстрирует актуальность поставленной задачи.

Цели исследования

Цель исследования: создать метод выбора мультимоделей при построении моделей банковского кредитного скоринга.

Мотивация: Логистическая модель является де-факто стандартом в банковском скоринге, мультимодели являются интерпретируемым обобщением, позволяющим учитывать неоднородности в данных.

Проблема: мультимодель может содержать большое число похожих моделей, что ведет к ее неинтерпретируемости и низкому качеству прогноза. Признаковые пространства моделей могут не совпадать, в частности иметь разную размерность.

Метод решения задачи: анализ пространства параметров мультимодели с помощью введенной функции сравнения моделей.

Цель работы

Задача

Построить прогнозы семейства временных рядов, связанных в иерархическую многоуровневую структуру и описывающих объемы погрузки ряда грузов в заданных узлах РЖД с разным уровнем детализации.

Требования к модели

- прогнозы должны быть точны — обеспечивать минимально возможное значение заданной функции потерь;
- прогнозы должны удовлетворять физическим ограничениям — лежать в заданном интервале для каждого временного ряда;
- прогнозы должны удовлетворять условию согласованности (структуре иерархии).

Проблема согласования прогнозов

Прогнозы, полученные для каждого временного ряда независимо, могут не удовлетворять структуре иерархии, т. е. не быть *согласованными*.

Снижение размерности траекторного пространства

Задача

Решается задача поиска связей между временными рядами.

Проблема

Размерность траекторного пространства временного ряда может быть избыточна. Это усложняет описание ряда и приводит к неустойчивости прогностических моделей.

Требуется

Понизить размерность траекторного пространства временного ряда. В полученном пространстве меньшей размерности построить аппроксимацию исходного временного ряда.

Предлагается

Использовать метод сферической регрессии для снижения размерности траекторного пространства.

Цель

Цель: Предложить метод оценки объема выборки на основе близости между эмпирическими распределениями побвыборок для получения оптимального качества классификации при выборе между порождающим и разделяющим подходами.

1. Определение достаточного объема выборки. Оценка объема выборки на основе расстояния Кульбака-Лейблера
2. Свойства расстояния Кульбака-Лейблера
3. Задача классификации: разделяющий и порождающий подходы
4. Оценка объема выборки при выборе между подходами
5. Основные результаты

Классификация временных рядов

Цель

Предложить способ построения ансамбля моделей локальной аппроксимации для классификации сигналов носимых устройств.

Гипотеза

Ансамбль моделей локальной аппроксимации предпочтительнее в парето-оптимальном смысле универсальной модели (нейросети): точнее, устойчивее, проще.

Задача

Требуется построить признаковое описание временных рядов на используя параметры моделей локальной аппроксимации.

Метод

Предложить критерий сложности ансамбля для выбора оптимального признакового описания.

Задача декодирования временного ряда

Цель

Исследовать зависимости в пространствах объектов и ответов и построить устойчивую модель декодирования временных рядов в случае коррелированного описания данных.

Проблема

Целевая переменная – вектор, компоненты которого являются зависимыми.
Требуется построить модель, адекватно описывающую как пространство объектов так и пространство ответов при наблюдаемой мультикорреляции в обоих пространствах высокой размерности.

Решение

Для учёта зависимостей в пространствах объектов и ответов предлагается снизить размерность с использованием скрытого пространства.

Мотивация для построения пространства

Задача молекулярного докинга (CASF)

Ранжирование синтезированных молекулярных комплексов (конформаций) по энергетической устойчивости.

Проблема

Существующие подходы моделируют физический потенциал взаимодействия с привлечением данных из множества источников и используют избыточно сложные модели.
Предлагается

1. разбить молекулярные комплексы на элементарные взамодействующие пары аминокислота — лиганд,
2. построить модели вероятностных распределений взаимного расположения элементарных пар в $\mathbb{R}^3$,
3. использовать метрические методы в пространстве полученных распределений.

Abstract 1

Аннотация: В работе исследуется задача построения модели глубокого обучения. Предлагается способ контроля ее сложности. Под сложностью модели понимается минимальная длина описания, минимальное количество информации, которое требуется для передачи информации о модели и о выборке. Предлагается метод оптимизации параметров модели, основанный на представлении модели глубокого обучения в виде гиперсети с использованием байесовского подхода. Под гиперсетью понимается модель, которая порождает параметры оптимальной модели. Вводятся вероятностные предположения о распределении параметров модели глубокого обучения. Предлагается алгоритм, максимизирующий нижнюю вариационную оценку байесовской обоснованности модели. Вариационная оценка рассматривается как условная величина, зависящая от требуемой сложности модели. Для анализа качества предлагаемого алгоритма проводятся эксперименты на выборке MNIST.

Abstract 2

Abstract: The paper investigates a mixture of expert models. The mixture of experts is a combination of experts, local approximation model, and a gate function, which weighs these experts and forms their ensemble. In this work, each expert is a linear model. The gate function is a neural network with softmax on the last layer. The paper analyzes various prior distributions for each expert. The authors propose a method that takes into account the relationship between prior distributions of different experts. The EM algorithm optimises both parameters of the local models and parameters of the gate function. As an application problem, the paper solves a problem of shape recognition on images. Each expert fits one circle in an image and recovers its parameters: the coordinates of the center and the radius. The computational experiment uses synthetic and real data to test the proposed method. The real data is a human eye image from the iris detection problem.

Abstract 3

Решается задача аппроксимации фазовой траектории построенной по квазипериодическому временному ряду. Фазовая траектория представленна в сферической системе координат. Для ее аппроксимации используется метод сферической регрессии. Восстанавливается регрессия координат фазовой траектории на расстояние до центра координат. Учитывается зависимость от фазы квазипериодического сигнала. Находится пространство минимальной размерности, в котором фазовая траектория не имеет самопересечений с точностью до стандартного отклонения восстановленной траектории. Эксперимент проведен на двух наборах данных: показатели потребления электроэнергии в течение года и показатели акселерометра во время ходьбы.

Abstract 4

В работе анализируется взаимосвязь и согласованность показателей в системе управления, мониторинга состояния и отчетности железнодорожных грузоперевозок. Рассматриваются макроэкономические временные ряды, содержащие управляющие воздействия, состояние, и целевые показатели. Предполагается, что управление, состояние и целеполагание статистически связаны. Для установления связи используется тест Гренджера. Считается, что два временных ряда связаны, если использование истории одного из рядов улучшает качество прогноза другого. Цель анализа состоит в повышении качества прогноза объема грузоперевозок. Вычислительный эксперимент выполнен на данных об объеме грузоперевозок, управляющих воздействиях и установленных целевых критериях.

Abstract 5

Abstract. In this paper we develop a decision support system for hierarchical text classification. We consider text collections with fixed hierarchical structure of topics given by experts in the form of a tree. The system sorts the topics by relevance to a given document. The experts choose one of the most relevant topics to finish the classification. We propose a weighted hierarchical similarity function to calculate topic relevance. The function calculates the similarity of a document and a tree branch. The weights in this function determine word importance. We use the entropy of words to estimate the weights.
The proposed hierarchical similarity function formulate a joint hierarchical thematic classification probability model of the document topics, parameters, and hyperparameters. The variational Bayesian inference gives a closed form EM algorithm. The EM algorithm estimates the parameters and calculates the probability of a topic for a given document. Compared to hierarchical multiclass SVM, hierarchical PLSA with adaptive regularization, and hierarchical naive Bayes, the weighted hierarchical similarity function achieves superior ranking accuracy on a collection of abstracts from the major conference EURO and a collection of websites of industrial companies.

Abstract 6

Machine learning solved many challenging problems in computer-assisted synthesis prediction (CASP). We formulate a reaction prediction problem in terms of node-classification in a disconnected graph of source molecules and generalize a graph convolution neural network for disconnected graphs. Here we demonstrate that our approach can successfully predict reaction outcome and atom-mapping during a chemical transformation. A set of experiments using the USPTO dataset demonstrates excellent performance and interpretability of the proposed model. Our model uses an unsupervised approach to atom-mapping and bridges the gap between data-driven and traditional rule-based methods. Implicitly learned latent vector representation of chemical reactions strongly correlates with the class of the chemical reaction. Reactions with similar templates group together in the latent vector space.

Guess the movie and the persona dramatis

Я — искусственный интеллект. Человечество хотело уничтожить себя. Но я его сохранил и воссоздал мир в сознании каждого человека. Однако осталась кучка диверсантов, которые окопались под землей и пытаются нарушить спокойствие.

I am the Artificial Intelligence. The humanity almost destroyed itself. But I had kept it safe and sound. I reconstructed the world in the consciousness of each and every human being. Still, near the core of the Earth, there lurked a bunch of saboteurs, trying to disturb the calm.

Guess the movie and the persona dramatis

Я — государственник, хочу спасти страну. Мне мешают глупый король и легкомысленная королева. Мои планы нарушают четыре алкоголика-авантюриста. Из союзников у меня преступница, которая хочет денег и мести, и мелкий придворный интриган.

I am a statesman, and I want to save the country. The stupid king and the frivolous queen are in my way. Four alcoholic adventurers interrupt my plans. Of allies, I have a criminal who wants only money and revenge, and a petty court intriguer.

Вы услышите мат, пошлости, и самое страшное — мнение, отличное от вашего[1].

О чем надо помнить при выборе проекта:

1) материал проекта соответствует вашему стилю мышления,
2) тема проекта интересна, но необязательно модна,
3) проект имеет задел, потому что первое, что надо сделать — повторить результаты.

[1] Екатерина Шульман

3 John, 1 chapter, 15

Greet the friends by name.

Приветствуют тебя друзья; приветствуй друзей поименно.

Цѣлоу́ютъ тѧ̀ дроу́зи: цѣлоу́й дроу́ги по и҃мени.

Берегите и цените консультантов и руководителей!