Требования к диссертации бакалавра

Вадим Стрижов

Кафедра интеллектуальных систем
Московский физико-технический институт

m1p.org/is
youtube.com/c/MachineLearningPhystech

2022

Положение о государственной итоговой аттестации студентов ВШЭ

Приложение к приказу НИУ ВШЭ от 26.01.2021
№ 6.18.1-01/2601-05

4.1. Формы, требования, критерии оценивания, порядок выбора темы, сроки и особенности этапов подготовки ВКР в НИУ ВШЭ определяются в Положении о КР/ВКР и в Правилах подготовки ВКР.

“8. Общие требования к выпускным квалификационным работам”

Академический формат — исследование, осуществляемое в целях получения новых знаний о структуре, свойствах и закономерностях изучаемого объекта (явления).

Проектно-исследовательский формат — разработка прикладной проблемы, результатом которой является создание некоторого продукта (проектного решения).

Факультет компьютерных наук НИУ ВШЭ Образовательная программа магистратуры «Системная и программная инженерия» Правила подготовки, оценивания, защиты и публикации курсовых и выпускных квалификационных работ студентов https://cs.hse.ru/data/2017/07/21/1173845874/Правила ВКР КР ОП СПИ.pdf

“В работах по физ.-мат. наукам нет внедрения, а в технических нет теорем;)”

Приложение 8 — исследовательский проект, Приложение 9 — программный проект

1. Полнота достижения поставленных целей и задач работы
2. Полнота освещения состояния предметной области и использования источников информации
3. Сложность и (8) полнота проведенного исследования (9) объёмность программной реализации или предложенных технологических решений
4. Качество (8) оформления отчёта о ВКР, в т.ч. списка использованных источников. Ясность и четкость изложения (9) итогового продукта, в т.ч. полнота верификации и тестирования, и т.д.
5. (9) Качество оформления работы, в т.ч. отчёта и программного кода. Ясность и четкость изложения в отчёте

5. (8), 6. (9) Четкость выдерживания запланированного графика работы, своевременность прохождения основных этапов выполнения ВКР, взаимодействие с руководителем ВКР

Избежим противоречия

Деканату важно защитить себя от нештатных ситуаций и он выпускает об этом документ

Студенту важно защитить качественный диплом, который останется с ним на всю жизнь, и он должен иметь гарантии его качества

Риски научного исследования студента

1. Я хочу заниматься темой $\mathfrak{X}$.
2. Вложу усилия, а у меня ничего не выйдет.
3. Как повысить шансы на успех?

Курс “Моя первая научная статья” m1p.org читается с 2012 года.

Самая интересная научная работа для меня — разобраться в новой научной статье и запрограммировать ее.

Антон П., студент третьего курса

Покупая птицу, проверьте, есть ли у нее зубы

Является ли ваша деятельность научным исследованием?

Проект — это либо наука, либо коммерция, либо самообразование. Дипломная работа является квалификационной, она должна показывать квалификацию студента.

Положение о выпускной квалификационной работе студентов МФТИ

Содержательная часть рецензии содержит заключение об

1. актуальности (и новизне) исследования,
2. достоверности результатов исследования,
3. теоретической и практической значимости полученных результатов,
4. основных результатах исследования (личный вклад),
5. положительных сторонах и недостатках исследования (единство исследования).

от 21.01.2019

Формула требований

Научность, личный вклад, единство, новизна, достоверность

Критерий

Максимум значимости при минимуме усилий.

Смотри как построены такие работы.

Например, Neur ODE 2018 и работы Понтрягина 1962.

Чек-лист руководителя при планировании

Ясно видеть

1. результат,
2. реакцию научного сообщества на него.

Прочее — в Положении о ВКР.

За какую задачу браться?

1. Масштабность проблемы: решение проблемы должно касаться большого числа людей, специалистов, лиц принимающих решения.
2. Заброшенность (популярность) проблемы. Общая ошибка: решать популярные проблемы.
3. Решаемость проблемы. Выбор просто и элегантно решаемых проблем.
4. Наша готовность к решению проблемы, квалификация: похожими проектами мы уже занимались.*

Пункт 4. Иначе риском является студент, а не проект.

Синдром FoMO при выборе темы исследований

«Все, кроме меня»

1. Программируют на Питоне
2. Используют нейросети
3. Знают, что такое NeurODE
4. ...

Плюс — актуальность (даже при дальнейшей тупиковости)
Минус — нет новизны (подготовка и публикация занимает 1–3 года)
Полезно следить за призовыми публикациями и их судьбой

FoMO is fear of missing out: «АрХив заменяет мне Инстаграм»

Исследование как спорт

Цель — показать на новой модели качество (precision, F1, AUC, . . .) выше, чем у альтернатив.

Насколько быстро и каким способом будет побит твой рекорд?

Достаточное условие достижения критерия

Публикция в **рецензируемом** журнале

Выбрать нужное в списке ArXiv, ВАК, Скопус/Конференции, WoS, Core, опубликовать, наклеить обложку, — диплом готов!

До момента выдачи студенту темы диплома

1. Понимание объекта исследования: что именно развиваем, метод или код, какой именно
2. Точное целеполагание: каких именно новых результатов ждем
3. Формальная постановка задачи вместе с критериями качества ее решения: обоснованность предполагает анализ ошибки
4. Что выносится на защиту: что студент объявит как личный значимый вклад

Это требует квалификации научного руководителя. Квалификации студента для этого недостаточно.

О, ужас!

- доклад посвящен обзору известных решений, в конце невнятный эксперимент
- целеполагание неясно, пересказ модной работы
- работа является сборной солянкой из проектов студента
- терминология не проработана, термины не определены
- теория и формулы не связаны с кодом, так как студент не понимает значения формул, а код для него черный ящик
- личный вклад неясен, невозможно отделить то, чем пользовался студент от того, что он сделал
- цитируются схемы, значение которых студенту не ясно
- нет анализа ошибки, неясно, верить ли таблицам
- практика противоречит теоретическим предположениям по неосведомленности

Низкое качество работы более всего заметно в докладе.

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.

Торжественный комплект — универсальная инструкция по изготовлению исследований

Планируем исследование по мере возрастания риска

1. повторить чужую статью (анализ свойств алгоритма),
2. старую задачу решить старым способом (пара "задача–способ решения" не была найдена в литературе),
3. прочие три варианта предыдущего пункта,
4. ...
5. применить новые эвристики в оптимизации ("рационализаторское предложение")
6. наконец-то выписать все предположения о порождении выборки и вывести вероятностную модель
7. применить методы абстрактной математики и теоретической физики к методам машинного обучения

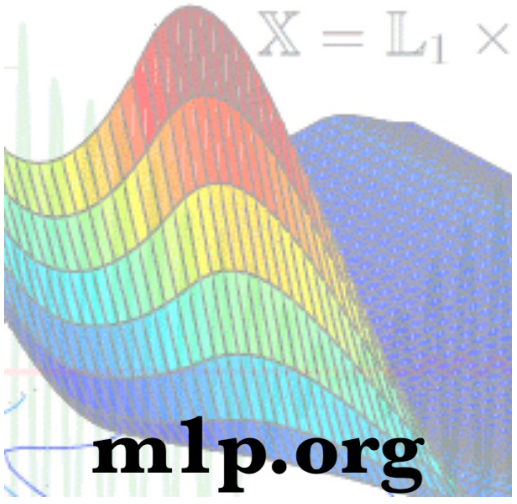
По мотивам Ильи Ильфа и Евгения Петрова «Золотой телёнок».

NB! Все эти способы призваны повышать достоверность научных результатов.

The course produces student research papers

Machine learning projects: how to

- state the problem,
- make the project feasible,
- present results of the experiment



Science requires community:

- Student is a project driver, who wants to plunge into scientific research activities.
- Consultant, a graduated student, conducts the research and helps the student.
- Expert, a professor, states the problem and enlightens the road to the goal.

Four talks to convey your message to the audience

Week 3	Introductory pitch
6	The message
9	Computational experiment
12	Conference talk

Roadmap

1. Set the toolbox
2. Select your project
3. Read papers
4. Write introduction
5. State the problem
6. Set your experiment
7. Develop your theory
8. Make error analysis
9. Paper draft
10. Share your results
11. Finalize your paper
12. Present your talk

Deliveries are scheduled

- LinkReview with references and the literature review
- GitHub with the code and computational experiment
- Paper is ready for submission
- Slides for the presentation
- Video of the conference talk

До начала планирования исследования аналитик и (эксперт) обсуждают ключевые вопросы

1. Цель проекта. (Ожидаемый результат разработки.) **Ожидаемая цель исследования.**
2. Прикладная задача, решаемая в проекте. (Как результат будет использован?) **чем результат будет проиллюстрирован?**
3. Описание исторических измеряемых данных. (Форматы и тайминг.) **Алгебраическая структура данных.**
4. Критерии качества. (Как измеряется качество полученного результата, что будет в отчете?) **Функция ошибки, что будем оптимизировать.**
5. Выполнимость проекта. (Как показать, что проект выполним, список возможных рисков.) **План анализа ошибки.**
6. Условия, необходимые для успешного выполнения проекта. (Организация работ.) **Требования к выборке.**
7. Методы решения. (Библиотеки процедур.) **Поставленные гипотезы, оптимальные вероятностные модели.**

Problem statement for machine learning

Formal problem statement, an analyst has to set

1) an algebraic structure for the dataset from measurements
2) a data generation hypothesis from 1)
3) a model, or a mixture from 2)
4) an error function (quality criteria with restrictions) from 2)
5) an optimization algorithm from 3) and 4)

The result of the model construction is a Cartesian product

$$\{\textbf{models} \times \textbf{data sets} \times \textbf{quality critea}\}.$$

Def: *Big data rejects the i.i.d. (independent and identically distributed random variables) data generation hypothesis from 2). It requests a mixture model.*

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

Analyst creates an optimal model for expert to put it to operation

Quality criteria

- Accuracy: MAPE, AUC, F1 score
- Stability: forecasting variance, failure rate, parameter variance
- Complexity: number of parameters, Kolmogorov complexity

Origins of quality criteria

1. Theory: statistical hypotheses of data generation, algebraic structures of data, models of measurement
2. Computations: a criterion is useful to an optimisation procedure
3. Deployment: revenue, loss, failure rate

Get info and ask your questions

Telegram

t.me/m1p_org

Info

m1p.org

YouTube

Machine Learning

Russian: Вадим Викторович Стрижов, МФТИ

Автоматизация научных исследований

mlalgorithms@gmail.com

`http://www.machinelearning.ru/wiki/index.php?title=m1`