Phystech@DataScience

Вводная лекция

8 февраля 2025

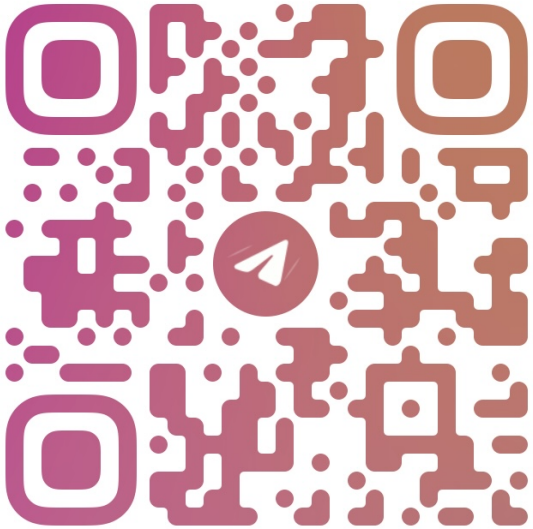
О курсе

ThetaHat

Проводимые нами учебные курсы

- Введение в анализ данных;
- DS-поток;
- Phystech@DataScience (в т.ч. статистика на ЛФИ);
- Статистика и машинное обучение на ФБМФ БТ;
- Мат. статистика на ФБМФ и ФЭФМ (семинары);
- Прикладная статистика на кафедрах в магистратуре;
- Машинное обучение в ШАД (частично);
- АБ-тестирование в ШАД.

Бот



@THETAHAT_PDS_BOT

Код регистрации **F1=0.874**

Сайт команды thetahat.ru

Почта thetahat@yandex.ru

План курса

Весенний семестр

- Вводный блок – 2 недели
  - AI-инструменты
  - Повторение теории вероятностей
  - Базовый анализ данных
- Линейные модели и нейронные сети – 3 недели
- Нелинейные модели – 2 недели
- Прикладная статистика – 5 недель

Осенний семестр

- Расширенный блок математической статистики
- Классическое машинное обучение (новые главы) и глубокое обучение

Формат занятий

Сдвоенная пара, теория перемешана с практикой.

Обязательное наличие ноутбука на парах.


ChatGPT


Gemini


YaGPT


deepseek


GIGA
CHAT


Claude

Не можешь предотвратить — возглавь!

Использование ИИ-инструментов в курсе

При выполнении *технической* работы в домашних заданиях **рекомендуется** использовать ИИ-инструменты!

Как? Узнаете на второй лекции!

Ограничения

1. ИИ может ошибаться. Его ошибка в вашей работе — ваша ошибка. Вам необходимо понимать и перепроверять ответы ИИ. *Аргументы "мне так сказал ИИ" не принимаются.*
2. Всю содержательную работу по задаче вам необходимо делать самостоятельно.
3. Злоупотребление ИИ приравнивается к списыванию.
4. Ваша цель — обучиться. Используйте ИИ для выполнения этой цели.

Система оценивания обязательной части

Активности:

- Л — доля выполнения легких заданий (их немного);
- С — доля выполнения сложных заданий (их много);
- ЛС — доля выполнения всех домашних заданий;
- В — доля правильных ответов на вопросы в боте на занятии;
- Т — доля выполнения тестов (на "удовл", в мае);

Списывания:

- Штраф -**2 балла за каждый случай всем участникам**;
- Объяснение "мы просто общались" не прокатит;
- Злоупотребление ИИ приравнивается к списыванию.

Система оценивания обязательной части

Правила:

Ставится максимальная оценка X, для которой A & (B | C) = True.

Оценка X	Условие A	Условие B	Условие C
3	Т > 34%		
4	Т > 67%		
5	Л > 25%		
6	Л > 50%		
7	Л > 75%		
8	Л > 25% и С > 25%	ЛС > 50% и В > 50%	ЛС > 75%
9	Л > 25% и С > 25%	ЛС > 65% и В > 65%	ЛС > 85%
10	Л > 25% и С > 25%	ЛС > 80% и В > 80%	ЛС > 95%

Хочу зачесть этот курс

Указывать в заявлении код **РУП 31 001**

Факультатив

1. Пройти отбор*
2. Получить оценку за курс
3. **При желании** взять отрывной и записать оценку за курс в свой диплом

Индивидуальный план

1. Пройти отбор*
2. Договориться со своей физтех-школой и кафедрой о замене предмета
3. Подписать все документы

Правила комфорта

- Постарайтесь задавать вопросы на занятии в тот момент, когда это актуально, не перебивая на полуслове.
  Другой вопрос лучше задать в перерыве или после занятия.

- Цените труд проверяющих :)
  В каком из случаев проверяющему больше захочется пойти навстречу автору вопроса?
  - "*Объясните вашу претензию, почему вы мне сняли баллы, я же все сделал, я не согласен*"
  - "*Добрый день! По такой-то задаче вы написали ..., но я считаю ..., потому что ..., и у меня в работе написано ...*"

Особенности проверки домашних заданий

Иногда к нам приходят отзывы:
"*Проверяющие проверяют рандомно,*
за одну и ту же ошибку разным студентам
сняли разное количество баллов."

Пусть на курсе 300 человек,
каждый проверяющий проверяет 30 работ.
Сколько результатов проверок нужно ему посмотреть для сравнения?
Сколько всего таких сравнений?
Оцените количество времени, которое нужно затратить.

Особенности проверки домашних заданий

Как мы решаем проблему?

- Общие критерии для всех проверяющих в табличке,
  с общим текстом и количеством баллов.
  Проверяющему достаточно поставить галочку.
- Сравнение частоты применения критерия между проверяющими
  с помощью статистического t-test'а.
- Сравнение среднего балла между проверяющими
  с помощью статистического t-test'а.
- Студенты 4 курса DS-потока разрабатывают ML-модель,
  которая ищет похожие комментарии проверяющих.

Мы стараемся, но мы не волшебники :)

Если вы заметили несправедливость проверки, пожалуйста,
напишите нам. Мы посмотрим и при необходимости поправим.
И учтем это для совершенствования наших методов проверки.

ThetaGrader

ThetaGrader — система автоматической проверки домашних заданий с помощью технологий искусственного интеллекта, разрабатываемая командой ThetaHat.

Она будет помогать проверять ваши домашки!

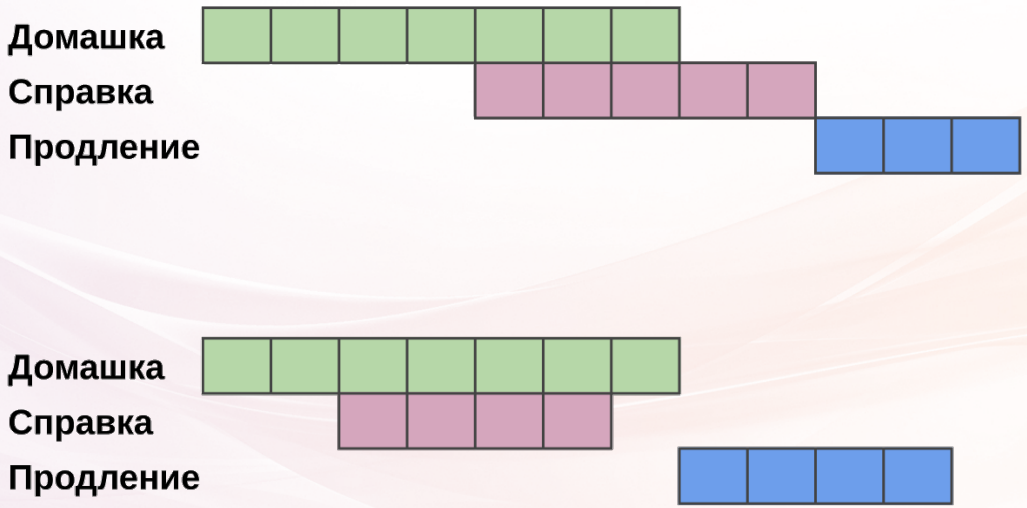
Переносы дедлайнов по уважительным причинам

Уважительные причины

- Медицинская справка с подписью и печатью.
- Приказ по институту об освобождении.

На сколько можно перенести

На количество дней пересечения интервала выполнения задания и датам по справке от даты дедлайна или окончания справки.



Тех. ассистент

Анна Бурханова

tg: @Anches_here

- Организация проверки домашних заданий
- Перенос дедлайнов по уважительным причинам
- Разные технические вопросы

Вопросы?

Анализ данных это

процесс поиска закономерностей в данных

при помощи

- средств визуализации данных,
- математических методов,
- программных алгоритмов.

Отличительная особенность:

нет четко зафиксированного ответа на каждый входящий объект.

Что можно почитать:

Анализ данных — основы и терминология
https://habr.com/ru/post/352812/

Всё, что вам нужно знать об ИИ — за несколько минут
https://habr.com/ru/post/416889/

Сравним задачи

Алгоритмы и структуры данных

Задача: дан массив x, нужно его отсортировать.

Ровно один правильный ответ, можно получить с помощью четких алгоритмов.

Комбинаторика

Задача: Сколько имеется способов раздать 11 разных цветков, трём девушкам: какой-то – 5, а остальным – по 3 цветка?

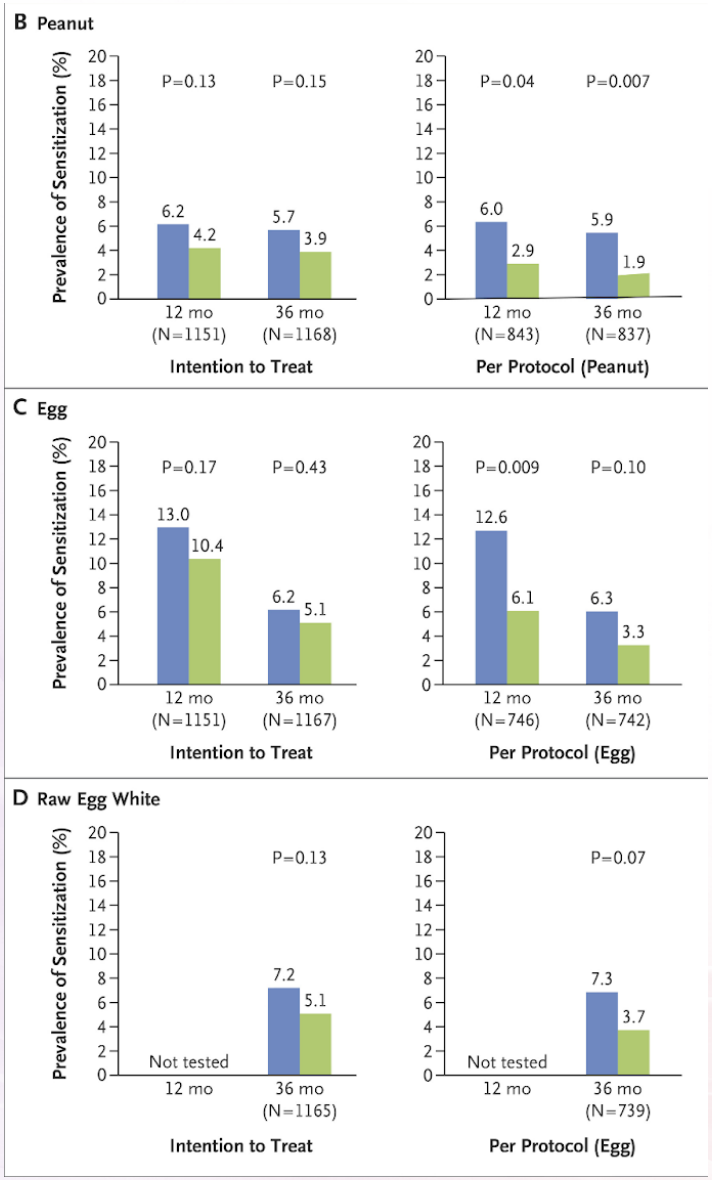
Ровно один правильный ответ.

Анализ данных

Задача: Имеются данные $(x_1, y_1), \ldots, (x_n, y_n)$.
Восстановите по ним функцию $f : x \mapsto y$.

Особенности: нет четкого ответа, требуется только приближение, но есть критерии качества.

Клинические испытания по аллергии у детей

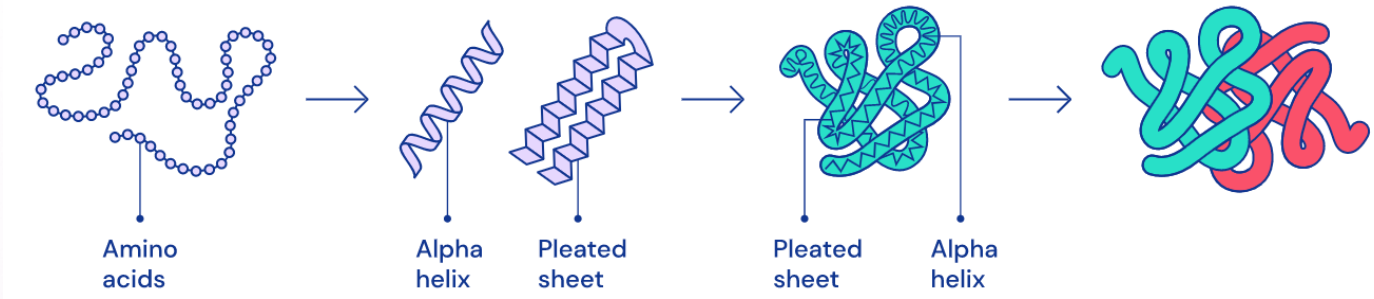


Цель: оценить, защитит ли раннее введение аллергенных продуктов в рацион детей, находящихся на грудном вскармливании, от развития пищевой аллергии.

Статистически значимые результаты получились при более раннем введении в рацион **арахиса и яиц**.

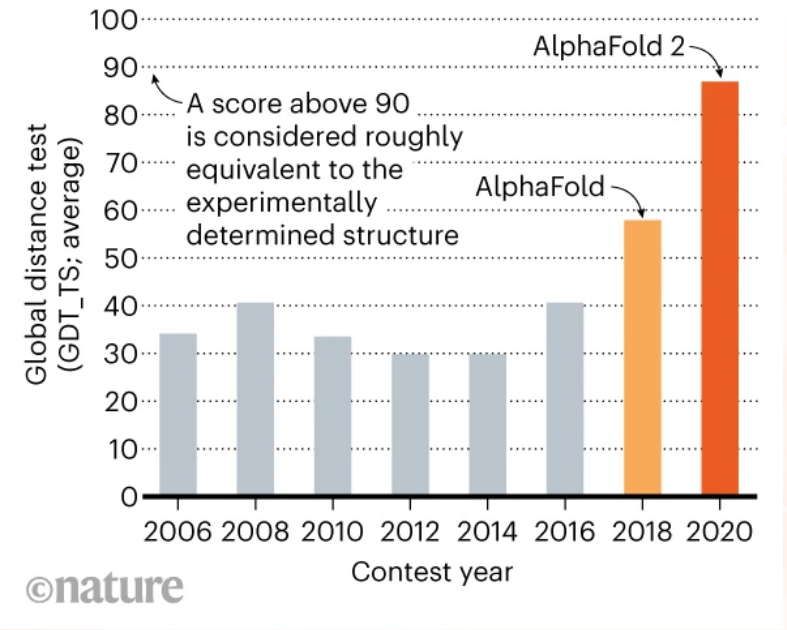
Методы: логистическая регрессия, хи-квадрат, точный тест Фишера

AlphaFold



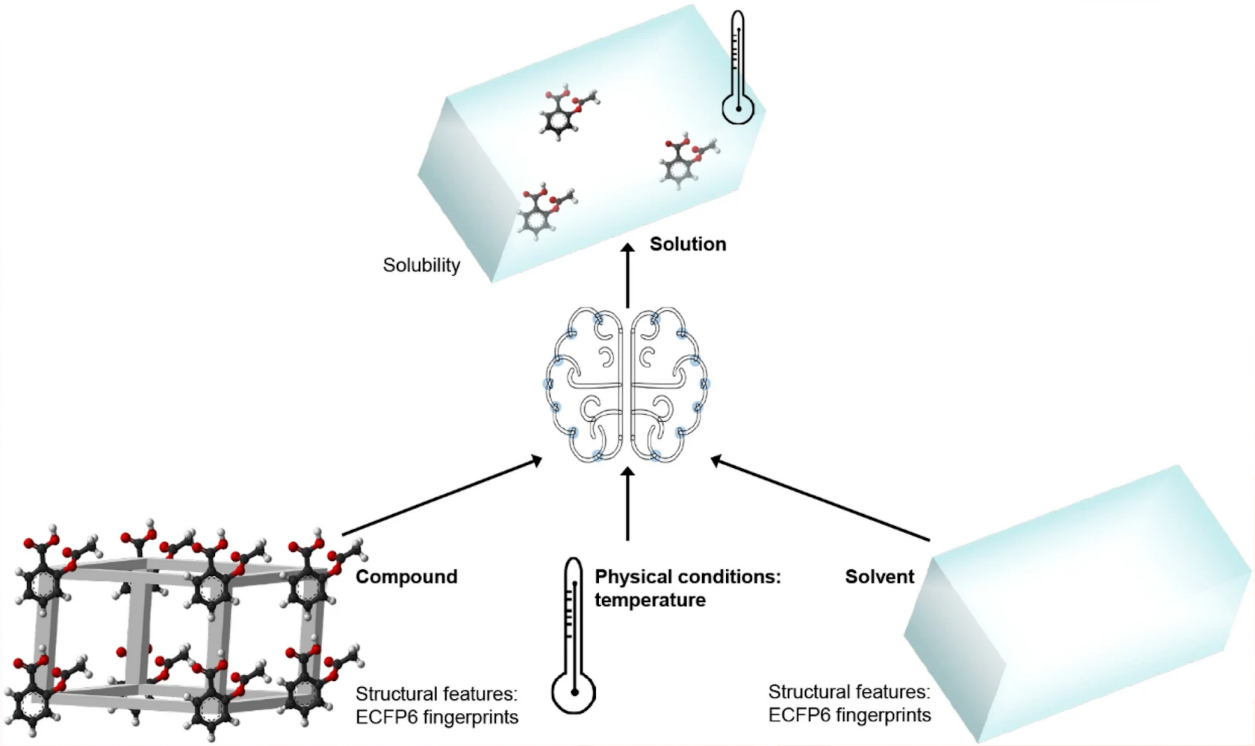
Алгоритм решил проблему, над которой бились ученые посление 50 лет: предсказание структуры белка по его аминокислотной последовательности.

Метод: нейронные сети



Предсказание растворимости вещества по его строению

Измерения растворимости зависят от трудоемких и дорогостоящих экспериментов.
Алгоритмы машинного обучения упростили эту задачу с высокой точностью ($R_2 = 0.91$ для lightGBM)



Методы: бустинг lightGBM, кластеризация (kNN, SVM), нейронные сети (GNN, RNN)