Теория вероятностей и статистика в Машинном Обучении

В результате освоения дисциплины будут получены знания об основных понятиях теории вероятностей, математической статистики, методах статистического анализа данных в прикладных задачах;

базовые навыки «прочтения» и содержательной интерпретации статистических данных, специфика применения вероятностно-статистического подхода.

В данном курсе:

- основные понятия теории вероятностей, в том числе исчисление случайных событий и случайных величин

- основные понятия математической статистики, включая статистическую теорию выборочного метода, оценивания параметров, проверки параметрических и непараметрических гипотез

Теория вероятностей и статистика в Машинном Обучении

Разбираем примеры:

- на практических задачах, в которых использования методов теории вероятностей, прикладной статистики и обработки данных позволяет создавать новую ценность, основанную на описательной и предсказательной аналитике;

- решаем задачи регрессии, классификации и кластерного анализа и их применении в предсказательной аналитике;

Компетенции:

- формулирование задач сбора, анализа и обработки данных

- прогнозирование будущих состояний и выработки оптимальных решений, основанных на анализе данных;

- практическая интерпретация и визуализация результатов анализа данных.

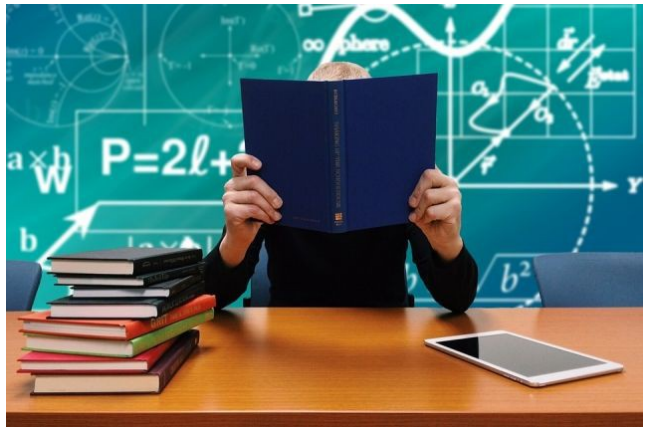
Содержание курса

– 10 лекционных занятий;

– 9 практических занятий;

– 3 лабораторные работы.

Экзамен



Рассматриваем практические задачи в области сферы образования, психометрики, экономики, транспорта и др.

ПО: **IBM SPSS Statistics, Python,** *Microsoft Excel*

Математика для анализа данных

Анализ данных - это процесс извлечения полезной информации из больших объемов данных.
Для успешного анализа данных необходимо не только использовать современные инструменты и технологии, но и обладать знаниями в экспертной области и глубокими знаниями в области математики.

При этом для самого машинного обучения знания предметной области не является обязательным

Фундаментальные математические дисциплины для машинного обучения

- Линейная алгебра и

  Аналитическая геометрия

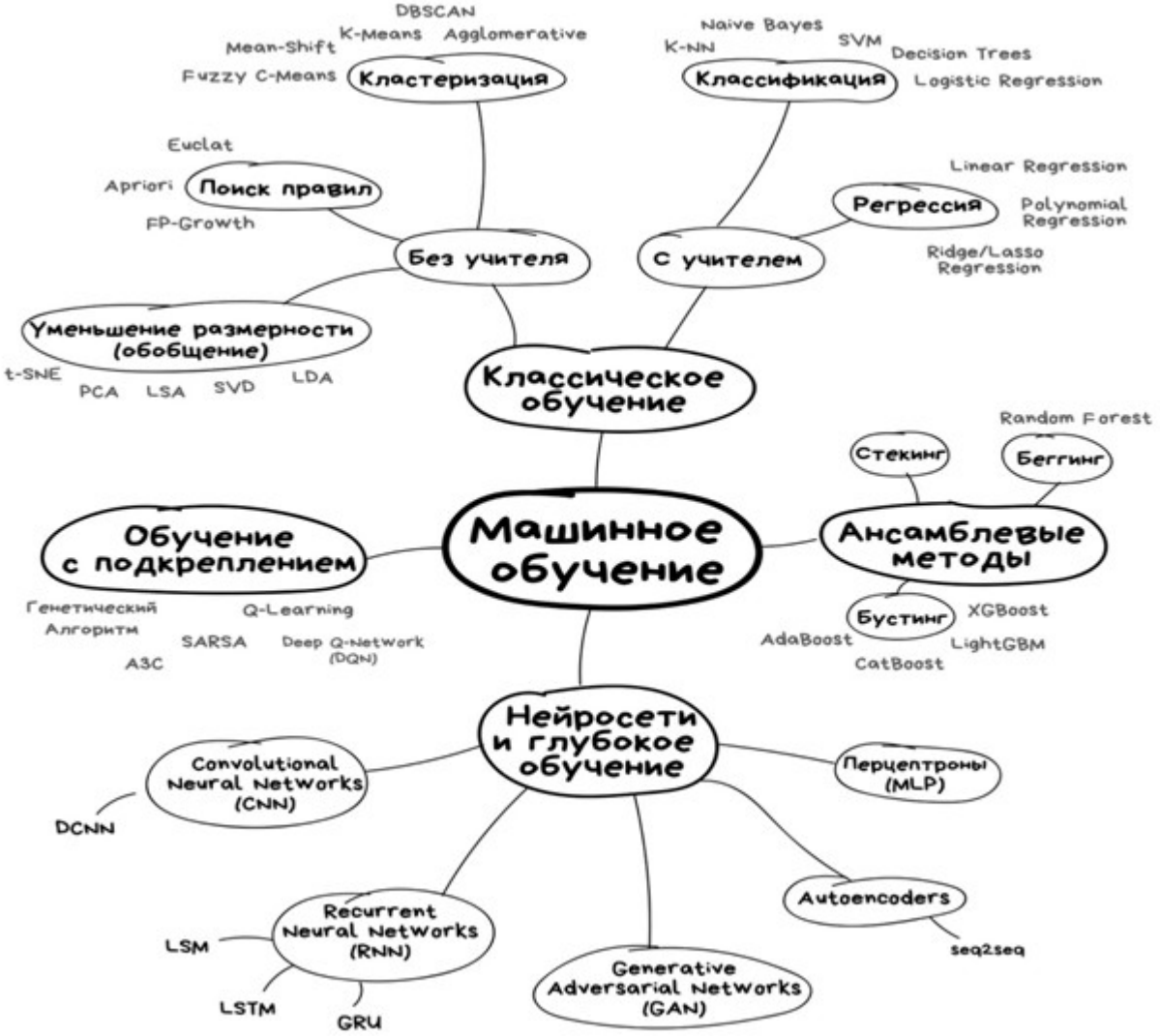
- Математический анализ

- Теория вероятностей и статистика

Необходимые в машинном обучении математические знания распределяются примерно следующим образом:
35% – линейная алгебра;
25% – теория вероятности и математическая статистика;
15% – математический анализ;
15% – алгоритмы;
10% – подготовка данных.

Виды машинного обучения

Область ML с применением теории вероятностей и статистики



Примеры практических задач

Сфера образования

Кластеризация регионов для проверки ЕГЭ

Подтверждения гипотезы о нормальности распределения (подтягивание на граничных баллах)

Завышение отметок регионами, муниципалитетами, ОО (необъективность)

Корреляции школьных и экзаменационных отметок

Аномальные расхождения в проверках экспертов

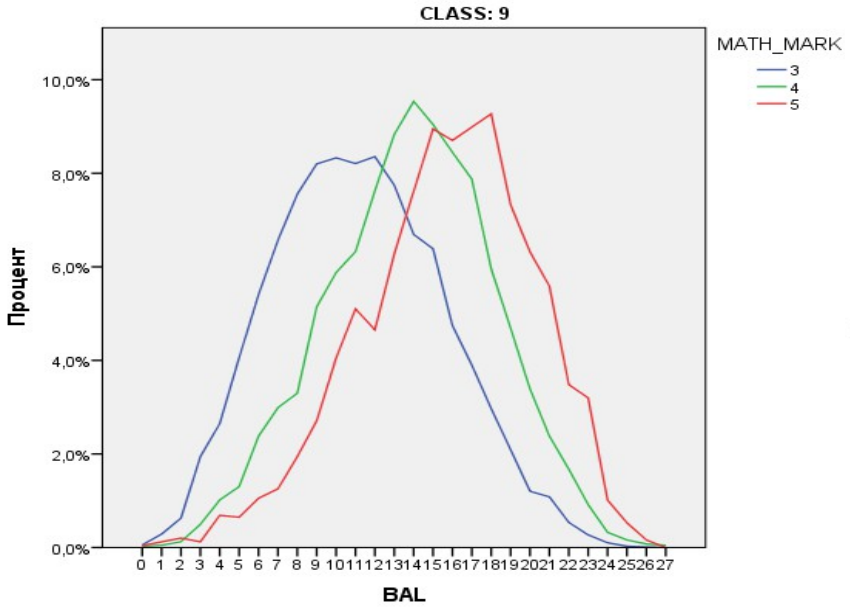
Корреляционный анализ в НИКО (отметки с соц.индексом)

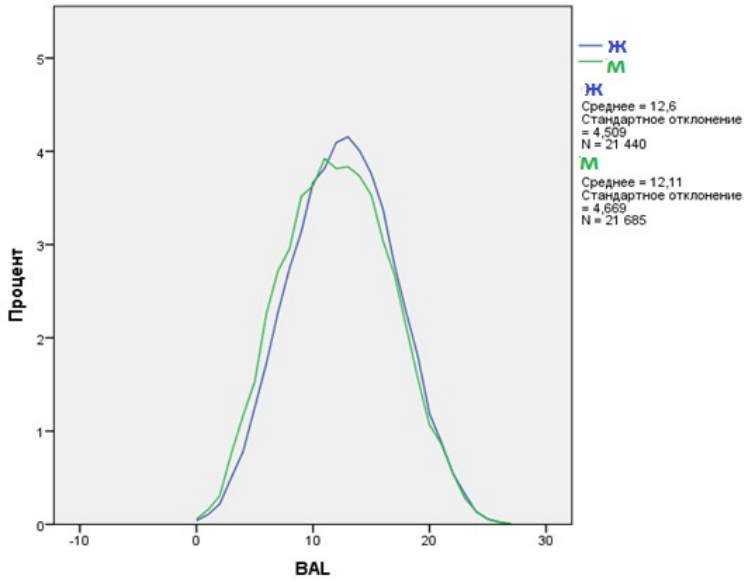
Создание геокластеров, как характеристики ОО

Шкалирование баллов (IRT)

Прогнозирование успеваемости

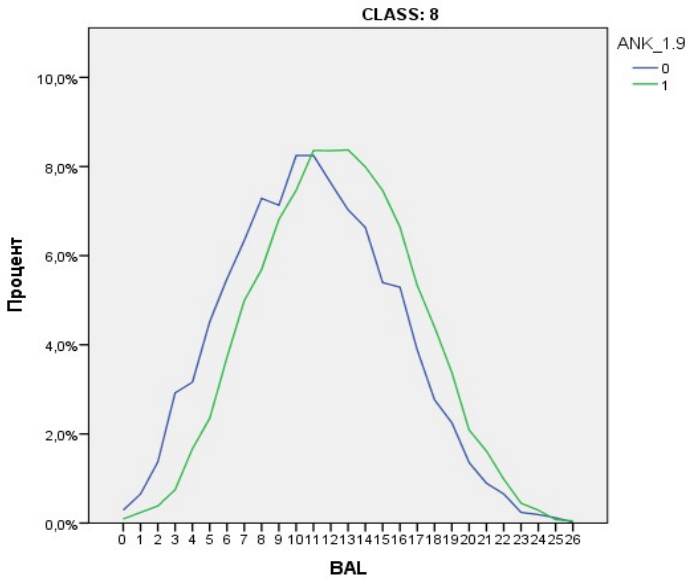
Примеры распределений, корреляций, проверки гипотез в сфере образования



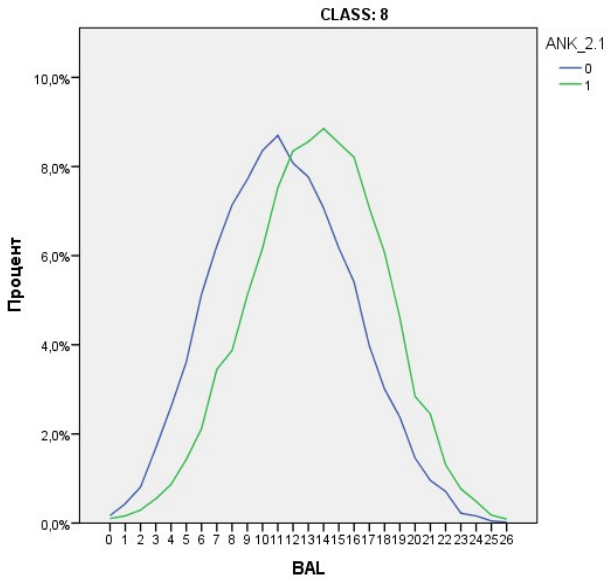


Центральная предельная теорема - сумма достаточно большого количества слабо зависимых случайных величин, имеющих примерно одинаковые масштабы, имеет распределение, близкое к нормальному.

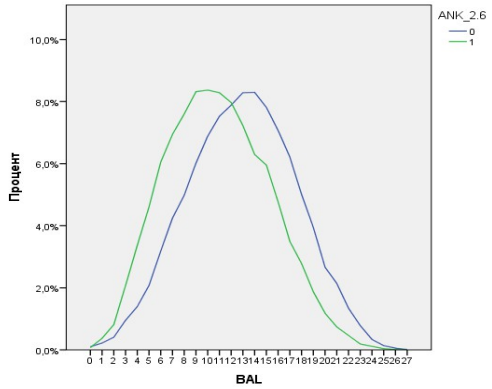
Примеры распределений, корреляций, проверки гипотез в сфере образования



Мне нравится слушать музыку
(0 – нет, 1 - да)



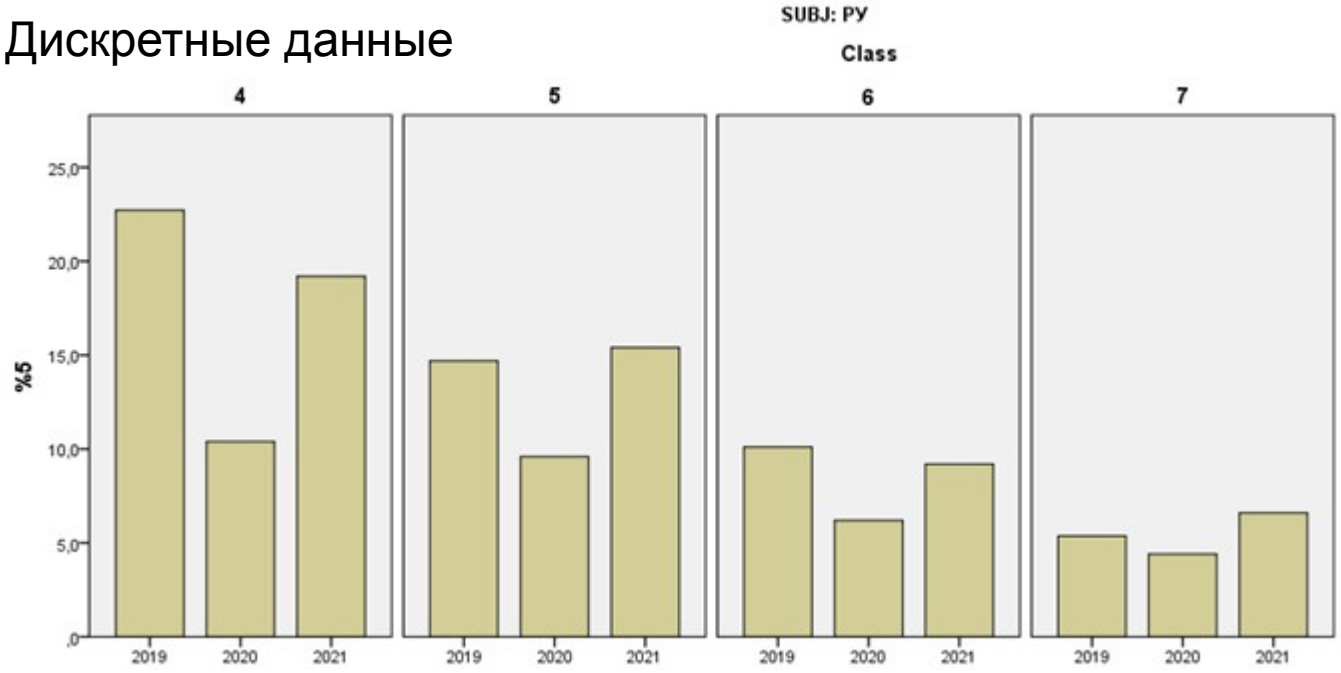
Я участвую в олимпиадах по математике
(0 – нет, 1 - да)

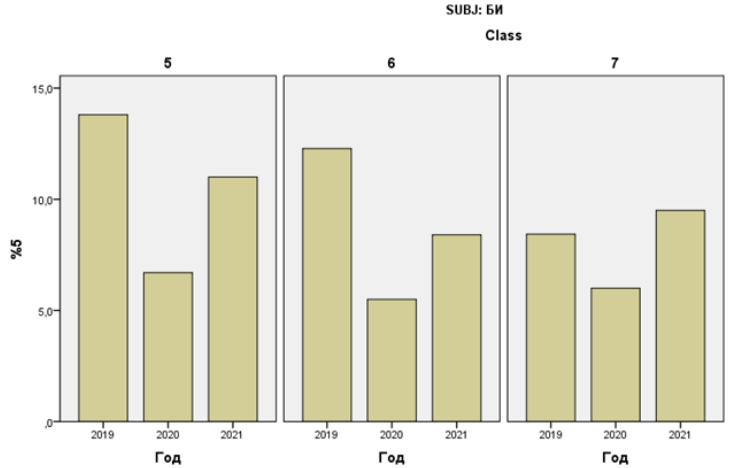


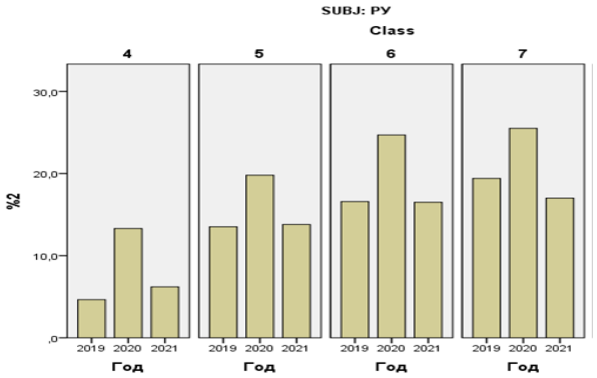
Я не участвую в олимпиадах (0 – нет, 1 - да)

Примеры распределений, корреляций, проверки гипотез в сфере образования

Дискретные данные



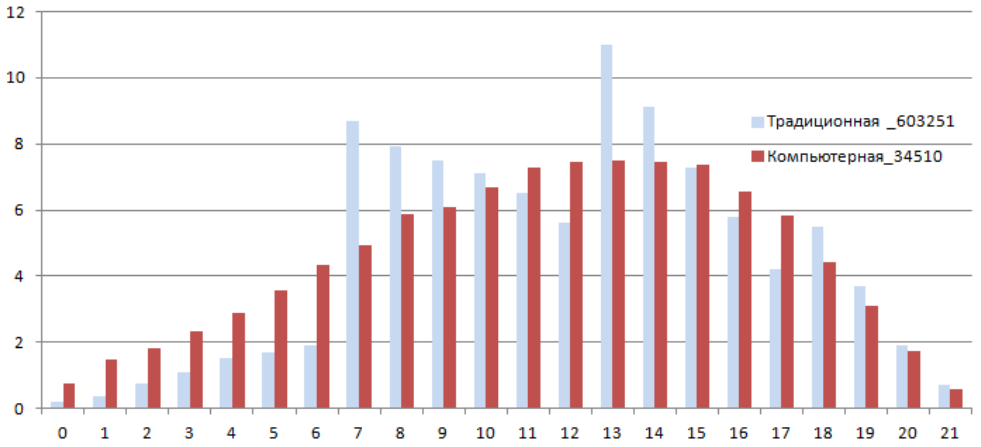




Примеры распределений, корреляций, проверки гипотез в сфере образования

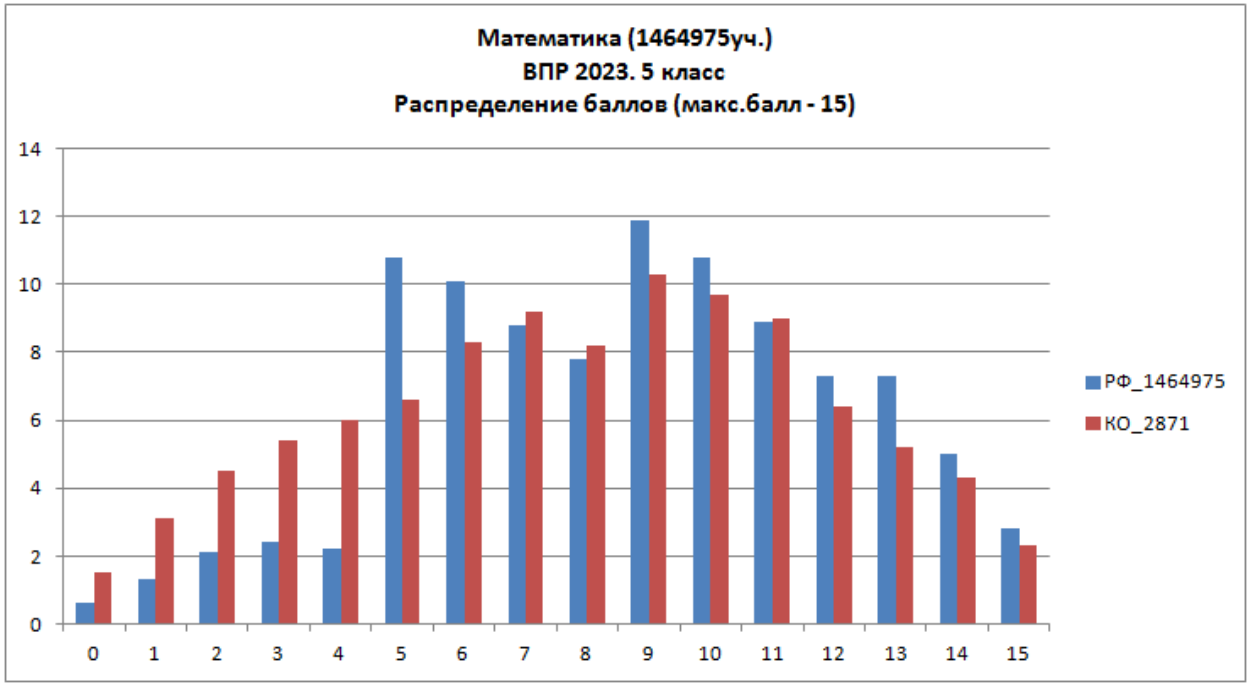
Распределение баллов – бумажная форма экзамена

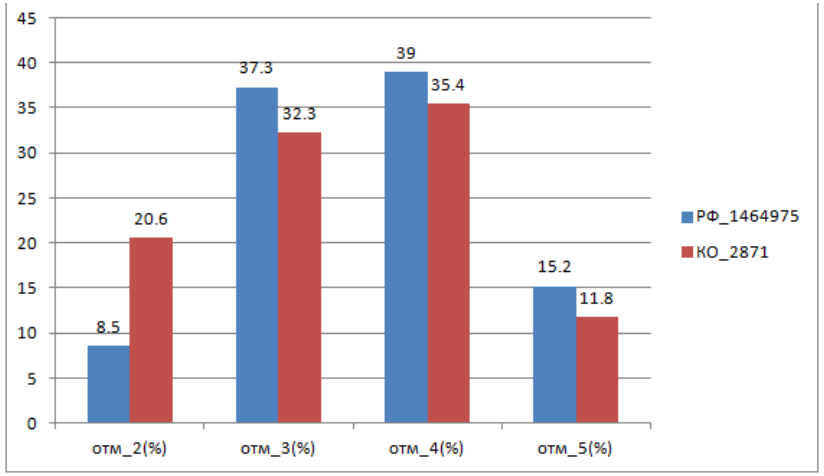
Распределение баллов – компьютерная форма экзамена



Гипотеза о нормальности распределения - ?

Создание контрольной выборки





Контроль объективности – выборка 0,2 %

Закон больших чисел - среднее арифметическое значение некоррелированных случайных величин при возрастании их числа стабилизируется, стремится к неслучайному числу — среднему арифметическому математических ожиданий этих случайных величин.

Темы теории вероятностей и математической статистики для ML

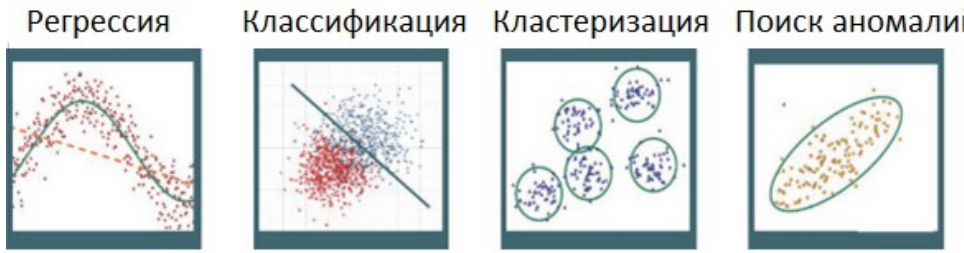
Теория вероятностей:
- комбинаторика;
- события и их вероятности;
- теоремы сложения и умножения вероятностей;
- формулы Байеса, Пуассона и Бернулли;
- дискретные случайные величины;
- дискретные распределения (геометрическое, биномиальное, Пуассона);
- непрерывные случайные величины;
- непрерывные распределения (равномерное, показательное, нормальное).

Статистика:
- генеральная совокупность и выборка;
- вариационные ряды (дискретные и интервальные);
- основные показатели статистики (мода, медиана, среднее и т.д.);
- графическое представление данных;
- оценки параметров генеральной совокупности;
- статистические гипотезы и методы их оценки;
- корреляция;
- регрессия (линейная, логистическая);
- кластеризация.

Заключение

Содержание курса

- *Введение в анализ данных*
- *Разведывательный анализ данных, очистка данных* (Л.р. 1)
- Основы теории вероятностей
- *Основные теоремы теории вероятности и математической статистики*
- Оценивание параметров (Л.р. 2)
- Проверка статистических гипотез
- Основные алгоритмы ML: (Л.р.3)



Литература:

- Ширяев А. Н. **Вероятность**. 3 - е изд., перераб. и доп. — М.: МЦНМО, 2004
- Тюрин Ю. Н., Макаров А. А., Симонова Г. И. **Теория вероятностей. Учебник для экономических и гуманитарных специальностей**. — М.: МЦНМО, 2009.
- **Анализ данных на компьютере**: учеб. пособие, Тюрин, Ю. Н., Макаров А.А — М.: ИНФРА–М, 2002.
- Соловьев В.И. **«Анализ данных в экономике: Теория вероятностей, прикладная статистика, обработка и анализ данных в Microsoft Excel.»** — М.: КноРус, 2021
- **Mathematics for machine learning**, M. P. Deisenroth, A. A. Faisal, C. Ong – Cambridge University Press, 2021