I - Основы обработки и анализа данных

Аннотация

Курс задуман как введение в самые базовые методы работы с количественными данными, в том числе основы теории вероятностей и математической статистики. Он предназначен для широкой публики и не требует специальных знаний, однако, программа составлена таким образом, чтобы даже подготовленный слушатель смог не только отточить уже приобретенные навыки, но и почерпнуть что-то новое.

В рамках курса предполагается изучить базовые и продвинутые инструменты исследования связей между признаками, овладеть навыками сравнения групп и построения прогнозов на основе регрессионных моделей. Отдельное внимание посвящено проведению предварительной обработки данных, знакомству с типичными проблемами, связанными с особенностями формата датасета, а также интерпретации, визуализации и представления результатов статистического анализа.

Ключевые слова

Статистический анализ данных, корреляция, зависимости, проверка гипотез, разведочный анализ данных, визуализация.

Приобретаемые навыки

Знание основных понятий теории вероятностей: случайные исходы, случайные события, условная вероятность и независимые случайные события.

Применение основных вероятностных формул: формулу полной вероятности, формулу Байеса, а также теоремы математической статистики: закон больших чисел и центральную предельную теорему.

Понимание понятий нулевой гипотезы и альтернативной гипотезы, ошибок типа 1 и типа 2.

Умение проводить статистические тесты и строить доверительные интервалы для среднего значения популяции, медианы и доли в случае одного / двух выборочного отбора из независимых и зависимых популяций.

Знакомство с основами эконометрического анализа для пространственных и временных выборок, владение принципами сравнения моделей, интерпретацией коэффициентов при разной спецификации.

Практика очистки и преобразования датасета, понимание возможных последствий нарушения предпосылок моделирования или особенностей данных.

Знакомство с библиотеками для статистического анализа и визуализации данных.

План

1. Вероятность, виды и свойства. Формула Байеса. Априорные и апостериорные вероятности. Пример для 2-х и 3-х последовательных событий с множественными исходами. Случайные величины. Дискретные и непрерывные распределения, их свойства. Примеры распределений и их важность в анализе данных: равномерное, биномиальное, пуассоновское, нормальное, экспоненциальное. Характеристики

распределений: среднее, медиана, ковариация, дисперсия, квантили. Тяжелые хвосты. Связанные и независимые выборки. Матрица ковариаций. Свойства математического ожидания, ковариации и дисперсии нескольких величин.

2. Графики и интерпретация плотности и функции распределения, расчет вероятности. Центральная предельная теорема. Закон больших чисел. Оценки параметров распределений и их свойства: несмещенность, эффективность, состоятельность. Доверительные интервалы. Метод максимального правдоподобия. Оценка среднего, доли, медианы. Оценка разницы средних, долей и медиан. Бутстрап. Необходимый объем выборки для построения доверительного интервала заданной ширины.

3. Проверка простых статистических гипотез. Сравнение 2-х групп. Параметрические и непараметрические критерии. Статистическая значимость. Сложные гипотезы и дисперсионный анализ ANOVA. Z и T-критерии, критические значения и таблицы распредений, P-value, доверительные интервалы, уровень значимости (alpha), мощность критерия (1 – beta), U-критерий Манна - Уитни (Mann–Whitney U-test), Ошибка I и II рода.

4. Предварительная обработка данных. Типичные проблемы и способы борьбы с ними. Аномальные значения и выбросы. Винзоризация. Шум. Пропущенные переменные. Фиктивная подстановка. Аппроксимации интерполяцией и определение наиболее вероятного значения. Категориальные признаки. Дублированные значения. Противоречивые данные. Ошибки в данных, их логическая и физическая неадекватность, погрешность измерения. Мультиколлинеарность. Несбалансированность. Стандартизация. Нормализация. Преобразование Бокса-Кокса.

5. Зависимости и закономерности. Оценка взаимосвязи между различными факторами. Корреляция. Коэффициенты Пирсона, Кернела и Спирмена. Основы эконометрического анализа. Парная и множественные регрессия на пространственной выборке. Постановка задачи. Метод наименьших квадратов. Виды: line-line, line-log, log-line, log-log. Дамми переменные сдвига и наклона. Интерпретация коэффициентов. Метрики качества R2, SSE и информационные критерии AIC, BIC, HQIC. Проблемы неправильной спецификации. Omitted Variable Bias. Доверительные интервалы для оценок коэффициентов. Дельта метод для вершины параболы.

6. Временные ряды. Модели тренда и сезонности. Статистические модели временных рядов. Авторегрессия. Адаптивные и модели. EWMA. Корреляции между временными рядами. Ложная корреляция. Отсутствие причинно-следственных связей. Двусторонних эффект. Стационарность. Взвешенное скользящее среднее. Прочие модели: GARCH, ...

7. Разведочный анализ данных. Графическое представление датасета. Визуализация. Обзор основных типов графиков: line chart, scatter-plot, histogram, матрица парных корреляций… Библиотеки matplotlib, seaborn, plotly.

Список литературы и источников

Probability and Statistical Inference, 9th edition. Robert V. Hogg; Elliot Tanis; Dale Zimmerman
Stock JH, Watson MW Introduction to Econometrics

II - Введение в машинное обучение

Аннотация

Курс включает в себя рассмотрение всех основных этапов статистического анализа, начиная от изучения предметной области и правильного сбора и предобработки данных, выборе оптимизируемой метрики в зависимости от исходной проблемы и заканчивая оценкой адекватности построенных моделей, выборе лучшей и интерпретации результатов.

В курсе рассматриваются основные алгоритмы машинного обучения, их преимущества и недостатки, ограничения каждого алгоритма и допущения данных.

Ключевые слова

Машинное обучение, классификация, регрессия, кластеризация, понижение размерности, регуляризация, логические, метрические, линейные алгоритмы, валидация, метрики, подготовка датасета.

Приобретаемые навыки:

Знание основных задач анализа данных, решаемыми с помощью машинного обучения (классификация, регрессия, уменьшение размерности, кластеризация, коллаборативная фильтрация и ранжирование)

Знание основных алгоритмов решения поставленных задач, их преимущества и недостатки, какие виды алгоритмов более подходят для каких видов данных, критическое понимание предмета, выделяя ограничения каждого алгоритма, допущения данных, на которые опирается каждый алгоритм, его сильные и слабые стороны

Обучение пакетам прикладных программ python вместе с его основными библиотеками анализа данных-numpy, scipy, pandas, matplotlib и библиотекой машинного обучения scikit-learn

Практический опыт от применения изученных методов к реальным наборам данных

Построение всего пайплайна исследования и разработок методов машинного обучения

Предварительная обработка данных и преобразование их в более подходящий вид для алгоритмов машинного обучения

Интерпретация полученных результатов

План

1. **Основные термины и понятия**

Основные термины и понятия. Виды задач в ML: с учителем (классификация, регрессия, ранжирование), без учителя (кластеризация, понижение размерности, поиск аномалий), с частичным привлечением учителя, с подкреплением… Типичные примеры. Обзор метрик качества. Бизнес-метрики. Онлайн-метрики. Офлайн-метрики. Для классификации: матрица ошибок, accuracy, recall, precision, f-мера. Для регрессий: MSE, MAE, MAPE, AMAPE... Функция потерь. L2, L1, Huber Loss. Обобщающая способность. Переобучение, недообучение. Bias, Variance, Noise. Декомпозиция bias-variance.

2. **Обучение и оценивание алгоритма**

Метрики обучения. AUC, Gini, log-loss. Пороги отсечения вероятности. Параметры и гиперпараметры. Разбиение датасета для обучения и оценки. Подбор гиперпараметров: стохастический градиентный спуск, генетические алгоритмы, перебор по сетке, случайный перебор. Отложенная выборка. Кросс-валидация: k-Fold, LOOCV. Лик.

3. **Линейные методы**

Линейная регрессия. Логистическая регрессия. Интерпретация коэффициентов. «Штраф» за сложность. Регуляризация. Наивный байесовский классификатор.

4. **Метод опорных векторов**

Разделяющая гиперплоскость. Опорные вектора. Случаи линейно разделимой и неразделимых выборок. Ширина разделяющей полосы. Функционал ошибки и его регуляризация. Ядра и спрямляющие пространства.

5. **Метрические методы**

Алгоритм k (взвешенных) ближайших соседей. Проклятие размерности. Масштабирование: нормализация, стандартизация. Евклидова и другие метрики расстояния.

6. **Логические методы**

Решающие пни, деревья и леса. Методы расщепления. Критерии информативности для классификации: энтропия, Джини. Недостатки жадной стратегии и способы их устранения. Критерии остановки обучения. Отсечение ветвей. Ансамбли алгоритмов. Теорема Кондорсе о присяжных. Мудрость толпы. Бустинг. Оптимизация. Градиенты. Выбор величины шага. Выбивание из локальных минимумов. Ранняя остановка.

7. **Отбор признаков**

Определение «важности» признаков. Фильтрация на основе ранжирования по релевантности. Оберточные методы (forward / backward). Создание агрегированных признаков. Понижение размерности, метод главных компонент. Ridge, Lasso.

8. **Кластеризация**

Иерархические (восходящие, нисходящие) и плоские методы. Дендрограммы. Четкие и нечеткие. Способы объединения кластеров (одиночная, полная связь, взвешенные и центроидные методы). K-Means, на основе плотности – DBSCAN, на основе распределений – EM алгоритм, MeanShift. Визуализация и t-SNE. Метрики качества кластеризации.

Список литературы и источников

1. Hastie T., Tibshirani R, Friedman J. The Elements of Statistical Learning. Springer, 2009.
2. Конспекты курса по машинному обучению К.В. Воронцова
3. Семинары Е. Соколова