Leonid Iosipoi
Research Fellow | Sber AI Lab | HSE University

Прикладная статистика

Программа «Прикладная статистика для машинного обучения», центр непрерывного образования, ВШЭ

Преподаватель: **Леонид Иосипой** (iosipoileonid@gmail.com).
Ассистент: **Константин Медведев**.

Сдача домашних работ по курсу организована в Google Classroom. Вам нужно зарегистрироваться в системе, нажать на + сверху и выбрать опцию “Присоединиться к курсу” (Join course). Код нашего курса: **nzbmdoy**. Все домашние задания будут появляться во вкладке “Задания” (Classwork).

У нашего курса есть папка в Dropbox со всеми материалами курса.

Организационная информация:
7 февраля — дедлайн по домашним работам, 11 февраля — объявление автоматов, 14 февраля — контрольная работа.

Я сделал небольшой опрос о курсе. Буду очень благодарен, если Вы оставите обратную связь после курса.
Очень интересно узнать, что Вам понравилось, а что — нет. Все, естественно, анонимно.

10.11.2020 **Введение в математическую статистику. Теория оценивания I.**

Оценивание параметров и сравнение оценок. Несмещенность и состоятельность. Метод моментов. Метод максимального правдоподобия.

Конспект: Презентация 1.

Полезные ссылки:

1. Материалы курса по теории вероятностей.
2. Визуализация некоторых идей теории вероятностей и статистики.
3. Probability and Statistics Cookbook – конспект с основными формулами.

Ссылки на литературу:
[1] М.Б. Лагутин. Наглядная математическая статистика;
[2] Н.И. Чернова. Математическая статистика. Учебное пособие.

17.11.2020 **Введение в математическую статистику. Теория оценивания II.**

Метод Монте-Карло. Тяжелые хвосты. Распределение Коши. Выборочное среднее, выборочная медиана и выборочная мода. Выборочная дисперсия. Среднеквадратическое/стандартное отклонение. Генерация случайных величин и решение задач в Python.

Конспект: Презентация 2. *Код:* Jupiter-ноутбуки 1.
Домашнее задание: Листок 1 (изменен: до 06.12.2020).

Дополнительное задание: прочитайте про «среднее» в одной из этих двух книг:

1. Д. Хафф. Как лгать при помощи статистики (Глава 2);
2. Ч. Уилан. Голая статистика (Глава 2).

Ссылки на литературу:
[1] М.Б. Лагутин. Наглядная математическая статистика;
[2] Н.И. Чернова. Математическая статистика. Учебное пособие.

Конспект: Презентация 3.

Дополнительное задание: посмотрите Видео про бутстрэп и прочитайте подробнее про доверительные интервалы в [2] (Глава 12).

Ссылки на литературу:
[1] Н.И. Чернова. Математическая статистика. Учебное пособие;
[2] М.Б. Лагутин. Наглядная математическая статистика.

01.12.2020 **Прикладаная статистика. Доверительные интервалы и Бутстрэп (практика). Проверка гипотез.**

Работа с распределениями в Python. Построение теоретических и бутстрэп доверительных интервалов в Python на примере нормального распределения. Доверительные интервалы для параметра «успеха» в модели Бернулли в Python.

Введение в проверку гипотез. Статистический критерий. Достигаемый уровень значимости (p-value).

Конспект: Презентация 4. *Код:* Jupiter-ноутбуки 2.
Домашнее задание: Листок 2, Данные к задачам (до 15.12.2020).

Дополнительное задание: прочитайте про парадокс «Неудачи» в [1] (стр. 24-25).

Ссылки на литературу:
[1] М.Б. Лагутин. Наглядная математиечская статистика;
[2] Н.И. Чернова. Математическая статистика. Учебное пособие.

08.12.2020 **Прикладаная статистика. Критерии согласия. Критерии однородности I.**

Критерии согласия. Критерий Колмогорова. Критерий Пирсона (хи-квадрат). Проверка равномерности. Проверка экспоненциальности (исключение неизвестного параметра, критерий Гини). Проверка нормальности (критерий Шапиро-Уилка, критерий Харке—Бера). Визуальный метод проверки гипотезы масштаба/сдвига – квантильный график (Q-Q Plot).

Введение в критерии однородности. Параметрические критерии: одновыборочный Z-критерий, одновыборочный t-критерий.

Конспект: Презентация 5. *Код:* Jupiter-ноутбуки 3.1.

Дополнительное задание: прочитайте в [1] про критерий согласия для нормального распределения, основанный на исключении неизвестных параметров, (стр. 167-168) и про критерии согласия для экспоненциального и нормального распределения, основанные на подстановке оценок параметров, (стр. 166 и стр. 168-169 соответственно).

Обратите внимание на новую книгу [3] в списке литературы. Это хорошая книга-справочник, в которой собрано огромное количество критериев и оценок. Кажется, что это самая полная книга по этим темам, доступная на русском языке.

Ссылки на литературу:
[1] М.Б. Лагутин. Наглядная математиечская статистика;
[2] Н.И. Чернова. Математическая статистика. Учебное пособие;
[3] А.И. Кобзарь. Прикладная математическая статистика. Для инженеров и научных работников.

15.12.2020 **Прикладаная статистика. Критерии однородности II.**

Параметрические критерии однородности: двухвыборочный Z-критерий и двухвыборочный t-критерий (независимые и зависимые выборки). Непараметрические критерии однородности для независимых выборок: критерий Колмогорова-Смирнова, критерий хи-квадрат, критерий Манна-Уитни. Непараметрические критерии однородности для зависимых выборок: критерий знаков, критерий знаковых рангов Уилкоксона. Оценка параметра сдвига. Критика критериев Стьюдента.

Конспект: Презентация 6. *Код:* Jupiter-ноутбуки 3.2.
Домашнее задание: Листок 3, Данные к задачам (до 30.12.2020).

Дополнительное задание: прочитайте про критерии однородности на случай нескольких (зависимых и независимых) выборок в [1] (стр. 237-248 и стр. 259-265).

Ссылки на литературу:
[1] М.Б. Лагутин. Наглядная математиечская статистика;
[2] Н.И. Чернова. Математическая статистика. Учебное пособие;
[3] А.И. Кобзарь. Прикладная математическая статистика. Для инженеров и научных работников.

Ковариация и корреляция. Коэффициенты корреляции Пирсона, Спирмена, Кендалла. Критерий Пирсона. Критерий Кендалла. Причинно-следственная связь и корреляция.

Задача регрессионного анализа. Формализация линейной регрессии. Метод наименьших квадратов (МНК). TSS, ESS, RSS. Коэффициент детерминации. Стандартные предположения в линейной регрессии и некоторые следствия из них.

Конспект: Презентация 7. *Код:* Jupiter-ноутбуки 3.3, Jupiter-ноутбуки 4.1.

Дополнительное задание: прочитайте про множественную и частную корреляцию в [1] (стр. 347-350).

Ссылки на литературу:
[1] М.Б. Лагутин. Наглядная математиечская статистика;
[2] Н.И. Чернова. Математическая статистика. Учебное пособие;
[3] А.И. Кобзарь. Прикладная математическая статистика. Для инженеров и научных работников.

12.01.2021 Прикладаная статистика. Регрессия II. Временные ряды I.

Статистические свойства оценок метода наименьших квадратов. Значимость значений регрессионных коэффициентов: критерий Стьюдента и Фишера. Парадоксы и ошибки в регрессии. Реализация линейной регрессии в Python. Удаление, добавление и преобразование признаков. Оценка влияния признаков на отклик.

Временной ряд. Тренд, сезонность, цикл. Автокорреляционная функция. Частная автокорреляционная функция.

Конспект: Презентация 8. *Код:* Jupiter-ноутбуки 4.2.
Домашнее задание: Листок 4, Данные к задачам (до 26.01.2021).

Дополнительное задание: прочитайте про критерий Фишера и общую линейную гипотезу [1] (стр. 368-372).

Ссылки на литературу:
[1] М.Б. Лагутин. Наглядная математиечская статистика.

21.01.2021 Прикладаная статистика. Временные ряды II.

Стационарность временного ряда. Критерий Дики-Фуллера. Модель авторегрессии AR(p). Модель скользящего среднего MA(q). Модели ARMA(p,q), SARMA(p,q)x(P,Q), ARIMA(p,d,q), SARIMA(p,d,q)x(P,D,Q).

Подгонка модели SARIMA(p,d,q)x(P,D,Q) к временному ряду. Стабилизация дисперсии. Преобразование Бокса-Кокса. Дифференцирование. Сезонное дифференцирование. Выбор параметров модели SARIMA. Информационный критерий Акаике (AIC). Анализ шума модели. Q-критерий Льюнга-Бокса.

Конспект: Презентация 9. *Код:* Jupiter-ноутбуки 5.
Домашнее задание: Листок 5, Данные к задачам (до 07.02.2021).

Дополнительное задание: прочитайте про методы прогнозирования временных рядов, отличные от SARIMA, здесь.

Ссылки на литературу:
[1] P. Cowpertwait, A. Metcalfe. Introductory Time Series with R;
[2] J. Cryer, K.-S. Chan. Time Series Analysis with Applications in R.

Дополнительные материалы для интересующихся

Кроме базовых книг, ссылки на которые есть выше, могут быть интересны:

[1] Г. Ивченко, Ю. Медведев «Введение в математическую статистику» – хороший классический учебник, если Вам нравится четкий академический стиль изложения.
[2] М. Кельберт, Ю. Сухов. «Вероятность и статистика в примерах и задачах» – в первой части этого трехтомника, на который мы часто ссылались в курсе по теории вероятностей, есть материал про теорию оценивания и проверку гипотез.
[3] Г. Джеймс, Д. Уиттон, Т. Хасти, Р. Тибширани «Введение в статистическое обучение с примерами на языке R» – перевод базовой версии очень популярной книги Хасти-Тибширани.
[4] Р. Кабаков «R в действии. Анализ и визуализация данных на языке R» – хорошо написанный прикладной учебник с большим количеством примеров и кода на R.