1 Параметрическая оценка

Генерируйте выборку объема $n = 100$ из равномерного распределения $R[-\theta, \theta]$ с параметром $\theta \in [1, 5]$. По этой выборке найдите две оценки $\hat{\theta}$ — методом моментов и методом максимального правдоподобия. Проведите несколько численных экспериментов, меняя параметр θ. Какая из оценок дает лучшее предсказание параметра? Увеличьте параметр n. Выполнена ли асимптотическая нормальность оценок $\sqrt{n}(\hat{\theta} - \theta) \to \mathcal{N}(0, 1/I(\theta))$ (проверьте визуально)?

Генерируйте выборку объема $n = 100$ из распределения $f_\theta(x) = \exp(\theta - x)I_{x>\theta}$ с параметром $\theta \in [0, 2]$. По этой выборке найдите две оценки $\hat{\theta}$ — методом моментов и методом максимального правдоподобия. Проведите несколько численных экспериментов, меняя параметр θ. Какая из оценок дает лучшее предсказание параметра? Увеличьте параметр n. Выполнена ли асимптотическая нормальность оценок $\sqrt{n}(\hat{\theta} - \theta) \to \mathcal{N}(0, 1/I(\theta))$ (проверьте визуально)?

2 Доверительные интервалы

В файле приведены данные об n детях. Постройте асимптотический и точный доверительный интервал для доли мальчиков. Сравните результаты.

В файле приведены данные о пробеге автомобиля между заправками полного бака. Предполагая нормальность распределения пробега постройте доверительный интервал для среднего: а) точный на основе распределения Стьюдента; б) асимптотический на основе нормального распределения; в) бутстрэпом с опцией normal. Сравните результаты.

3 Проверка вида распределения

В файле приведены данные о рекордах выпадения дождя в определенном месте. Можно ли утверждать, что они подчиняются экспоненциальному распределению? Используйте визуальный метод проверки и любые два из критериев на выбор.

В файле приведены данные о рекордах выпадения дождя в определенном месте. Можно ли утверждать, что они подчиняются нормальному распределению? Используйте визуальный метод проверки и любые два из критериев на выбор.

4 Однородность

В файле содержится информация об обуви детей из одного класса. Проверьте: а) можно ли утверждать, что мальчики и девочки одинаковы с точки зрения длины стопы б) ширины стопы в) левши и правши одинаковы с точки зрения ширины стопы г) чем больше возраст, тем меньше в целом стопа?

В файле приведены результаты о длительности жизни животных после принятия одного из нескольких ядов и одного из нескольких препаратов. Можем ли мы утверждать, что а) яды одинаково эффективны б) препараты одинаково эффективны в) каждый из препаратов одинаково работает против любого из ядов. Данные считайте близкими к нормальным из физических соображений.

5 Корреляционный анализ

Файл содержит данные об углеродном следе различных автомобилей на различных трассах. Какие параметры наиболее тесно связаны с углеродным следом?

Файл содержит информацию об измерениях нескольких параметров человека. Определите, какие из них наиболее сильно зависят друг от друга, учитывая исключенные (скорректированные) корреляции.

6 Регрессия

Файл содержит информацию об автомобилях. Используя разные предикторы (кроме модели), предскажите параметр «длина пробега на 1 литр топлива» с помощью линейной регрессии. Проведите анализ остатков.

Файл содержит информацию о нескольких физических характеристиках у мужчин. Можем ли мы предсказать массу по остальным характеристикам с помощью линейной регрессии? Проведите анализ остатков.

Файл содержит информация о преступлениях. Постройте линейную модель а) МНК б) Хубера в) RANSAC. Какая из моделей справилась лучше с точки зрения кросс-валидации?

7 Кластеризация

В файле содержится набор точек (вас интересуют только первые два столбца). Кластеризуйте его на а) 3 б) 6 в) 7 кластеров с помощью а) k-means б) DBScan в) OPTICS г) Spectral Clustering. Какой из алгоритмов дал лучший результат (сравните с третьим столбцом)?

В файле содержится набор точек. Кластеризуйте его с помощью иерархической кластеризации. Постройте дендрограмму и выберите оптимальное число кластеров. Разделите данные на такое число кластеров методом k-средних и сравните с оригиналом.

8 Классификация

Примените к данным из файла алгоритм случайного дерева и выявите наиболее важные переменные. Используйте случайный лес и выведите результаты его анализа. Используйте обучающую и тестирующую выборки.

Примените к данным из файла методы SVM и LDA. Являются ли данные линейно разделимыми? Используйте тестовую и обучающую выборки и исследуйте качество полученной модели.