Статистический анализ данных / Python. Задание №6. Регрессия

Условия

- Дедлайн:
  - **Очное:** 4 января 23:59
  - **Дистанционное:** 7 января 23:59
- **ВНИМАНИЕ** Выполненную работу нужно отправить на **psad.homework@gmail.com**, указав тему письма "[MADE19] Фамилия Имя - Задание 6". Квадратные скобки обязательны
- Максимальный балл: 10 баллов. В случае, если в решении есть грубые ошибки, задание отправляется на доработку. За отправку на доработку начисляется штраф **-2 балла** ПЛЮС **-1 балл** за каждый день, пока задание находится на доработке.
- Дедлайн – мягкий. За один день просрочки вы получаете **-1 балл**.

Задание

Требуется подобрать и применить наилучший статистический метод, позволяющий ответить на вопрос прикладной задачи; обосновать выбор метода, его применимость и оптимальность. Не нужно использовать все возможные методы, нужно выбрать наиболее подходящий. Помимо выводов, касающихся математических особенностей решения, необходимо в терминах предметной области сформулировать выводы, которые могли бы быть понятны гипотетическому заказчику-нематематику.

Необходимо сдать: ipython notebook и сгенерированный по нему pdf-файл с подробным отчётом по проведённому исследованию, содержащий визуализацию исходных данных, описания и выводы каждого этапа анализа — используемые методы, обоснование их применимости, графики.

По каждой разновидности регрессии в слайдах есть список требований в отчёту.

Данные для всех задач: `https://yadi.sk/d/ilC2z4s2yuoRL`

Вес детей при рождении

Имеется выборка из 1009 детей, родившихся в Северной Каролине в 2004 году; известны пол ребёнка, вес при рождении, период вынашивания, возрастная группа матери, а также курила ли мать во время беременности и употребляла ли алкоголь (birthweight.csv).

A: как вес ребёнка зависит от курения и употребления алкоголя (после поправки на остальные признаки)?

Токсичность рыб

Полихлорированные дифенилы — органические соединения, активно использовавшиеся в промышленности до 1970 годов, когда была показана их токсичность. Накопление ПХБ в организме приводит к подавлению иммунитета, провоцирует развитие рака, поражений печени, почек, нервной системы, кожи, способствуют развитию детской патологии. Из-за накопления ПХБ в озёрах США некоторые виды рыб в некоторых областях запрещены

к употреблению в пищу. Для своевременного обновления таких запретов необходимо периодически проводить мониторинг ПХБ. К сожалению, существует 209 различных разновидностей ПХБ, концентрация каждой из которых измеряется отдельным тестом. Для 69 видов рыбы известны концентрации семи соединений ПХБ (в миллионных долях), а также суммарная концентрация всех разновидностей ПХБ, их токсическая эквивалентность (TEQ) и суммарная токсическая эквивалентность образца, определяемая также вкладом диоксинов и фуранов (pcb.txt).

Б: насколько точно токсичность рыбы можно предсказывать по концентрации только нескольких ПХБ? Концентрации какого минимального количества соединений ПХБ нужно измерить, чтобы достаточно точно предсказать суммарную токсичность, или хотя бы токсичность только совокупности ПХБ?

Биоразлагаемость молекул

1055 химических молекул описаны с помощью 41 признака (число атомов кислорода, нитратных групп, донорных связей с водородом, потенциал ионизации и т.д.); 355 из них биоразложимы (biodeg.xlsx).

В: какие свойства молекул влияют на их биоразлагаемость?

Психическое здоровье в IT

1433 респондента поучаствовали в опросе о состоянии психического здоровья. На вопрос "Do you currently have a mental health disorder?"575 ответили положительно, 531 отрицательно, а ещё 327 — "Maybe"(IT_mental_health.csv).

Г: по данным о респондентах, давших однозначный ответ на вопрос о наличии психических расстройств, построить модель и предсказать с её помощью вероятность наличия расстройства у респондентов, не уверенных в своём психическом здоровье. Дать интерпретацию модели.

Открытие депозита

Имеются результаты обзвона 4119 клиентов португальского банка, которым предлагалось завести депозит. Известны социально-демографические характеристики клиентов, история предыдущих коммуникаций, социально-экономические показатели на момент совершения звонка (deposit.xlsx).

Д: какие признаки определяют готовность клиента открыть депозит по результатам обзвона?

Клетки опухолей груди

357 испытуемым с доброкачественными и 212 со злокачественными опухолями груди была сделана тонкоигольная аспирационная пункция с гистологическим исследованием пунктата. По полученным изображениям определялись следующие признаки опухолевых клеток: радиус, однородность текстуры, периметр, площадь, гладкость, компактность, степень вогнутости, доля вогнутых участков контура, симметричность, фрактальная размерность. Для каждого изображения были рассчитаны среднее значение каждого из этих признаков, стандартное отклонение и среднее по трём клеткам с максимальным значением признака (breast cancer.xls).

Е-Ё: оценить вероятность того, что опухоль злокачественная, по набору рассчитанных по изображению признаков. Построить функции, дающие точечную оценку и границы 95% доверительного интервала.

Данные антропометрии

Для 247 мужчин и 260 женщин измерены две группы антропометрических показателей — легко измеримые характеристики скелета и обхваты, всего 21 признак. Указаны возраст, пол, вес и рост (body.xlsx).

Ж: построить функцию, оценивающую по наименьшему набору признаков вероятность того, что испытуемый — женщина, и доверительный интервал для этой вероятности.

З: построить функцию, оценивающую возраст по имеющимся признакам; сравнить точность оценки возраста при отсутствии информации по обхватам и отсутствии информации по характеристикам скелета.

Цены домов

Известна продажная стоимость 1198 домов, проданных в 2006-2010, и значения 71 признака, описывающего эти дома (houses.xls).

И-Й: построить модель оценки продажной стоимости дома по его признаковому описанию; дать интерпретацию воздействия наиболее сильных факторов.

Эффективность тромболитической терапии

Собраны данные по 206 пациентам второго кардиологического отделения московской городской клинической больницы №25. Имеются результаты 14 анализов, а также 8 дополнительных признаков, описывающих пациента (пол, возраст, курение, наличие диабета и т.д.) (cardio.xls).

К: построить функцию, оценивающую вероятность выздоровления пациента в результате тромболитической терапии по приведённым признакам.

Сейсмическая опасность в шахтах

Собраны данные мониторинга сейсмической активности в польских угольных шахтах столбовой системы разработки. При сейсмической опасности существует серьёзный риск обрушения; в этом случае необходимо отозвать рабочих или использовать направленные взрывы для нейтрализации напряжения породы. Для каждого измерения известен бинарный индикатор сейсмической опасности — наличия в следующую восьмичасовую смену сейсмических толчков с энергией выше 10^4 джоулей (seismic.xlsx).

Л: построить модель сейсмической опасности, дать интерпретацию вклада показателей сейсмической активности.

Диагностика заболеваний позвоночника

Для 310 испытуемых измерены: наклон и смещение таза, угол изгиба поясницы, наклон плоскости тазовой поверхности крестца, радиус таза, степень смещения позвонков. Каждый из испытуемых либо здоров, либо болен спондилолистезом или межпозвонковой грыжей (spine.csv).

М: построить и интерпретировать модель, предсказывающую вероятность наличия заболевания позвоночника.

Успеваемость и потребление алкоголя старшеклассниками

Для 649 учеников старших классов двух португальских школ известны ряд демографических показателей и показателей успеваемости; для каждого студента известны также

уровень потребления алкоголя по выходным и будним дням в пятибалльной шкале от очень низкого до очень высокого и финальная оценка по португальскому языку (student-por.xlsx).

Н: смоделировать финальную оценку как функцию от всех показателей, кроме итоговых оценок по промежуточным семестрам; оценить влияние уровня потребления алкоголя на неё.

Вкус португальского вина

Для 1599 образцов красного и 4898 белого португальского вина Vinho Verde известны оценки (от 0 до 10), выставленные дегустаторами при слепом тестировании, а также значения одиннадцати биохимических показателей, полученных при лабораторном анализе (wine.xlsx).

О: построить модель, оценивающую содержания алкоголя по остальным характеристикам вина.

П: построить модель экспертной оценки по характеристикам вина, оценить влияние содержания алкоголя на экспертную оценку.

Хроническая болезнь почек

Госпиталь города Карайкуди, Тамилнад, Индия, собрал данные анализов 250 пациентов с хронической болезнью почек и 150 пациентов без неё (chronic_kidney_disease.xlsx).

Р: построить диагностическую модель хронической болезни почек, оценить вклад факторов.

Просрочка платежей по кредитам

Для 30000 клиентов тайваньского банка известны сумма кредита, демографические показатели и история платежей по кредитам за последние пять месяцев (факт просрочки, сумма неоходимой выплаты, сумма платежа) (default.xls).

С: построить модель, предсказывающую вероятность просрочки следующего платежа, оценить вклад факторов.

Пожертвования на благотворительность

Благотворительная организация разослала 4268 писем с предложением сделать пожертвование и получила отклик с пожертвованиями от 1707 адресатов. Для каждого адресата известны: индикатор ответа на предыдущее письмо, число недель, прошедших с момента предыдущего пожертвования, размеры текущего, предыдущего и среднего по всем предыдущим пожертвованиям в голландских гульденах, число писем, отправляемых адресату в год, доля писем, в ответ на которые приходят пожертвования (charity.xlsx).

Т: построить функцию, оценивающую вероятный размер пожертвования от адресата по историческим данным.

У: построить функцию, оценивающую вероятность получения пожертвования от адресата по историческим данным.

Ценообразование бриллиантов

Имеются данные о цене и свойствах 53940 бриллиантов. Известны: линейные размеры и признаки, построенные на их комбинациях, вес в каратах, цвет (закодирован буквами латинского алфавита: наиболее чистый цвет — буквой D, менее чистые — буквами E, F, G

и т.д., чем ближе к концу алфавита, тем "грязнее"), группа чистоты (отсутствие дефектов, профессиональная оценка, выдаваемая специалистами при исследовании бриллианта в лупу десятикратного увеличения; бриллианты без трещин и включений получают оценку IF ("internally flawless"), далее в порядке убывания чистоты следуют группы VVS1 и VVS2 ("very very slightly imperfect"), VS1 и VS2 ("very slightly imperfect")), стоимость бриллианта в долларах США (diamonds.xlsx).

Ф: существует общепринятая система классификации бриллиантов на мелкие — до 0.29 карата, средние — от 0.30 до 0.99 карата и крупные — свыше 1 карата. Достаточно ли для предсказания цены знать о весе бриллианта только к какому классу он относится, или предсказания с использованием знаний о точном весе значимо лучше?

Х: построить модель ценообразования бриллиантов, учитывая все особенности имеющихся данных

Внешний вид и привлекательность самок мечехвостов

Изучалось влияние внешних характеристик самок морских ракообразных на их привлекательность для самцов. Выборка состоит из данных о наблюдениях над 173 особями и содержит закодированные данные о размере самок, их весе, цвете, состоянии панциря, а также о количестве спутников (horseshoe crab.txt).

Ц-Ч: построить функцию, по внешним параметрам самки предсказывающую, будет ли у неё хотя бы один спутник. Оценить значимость каждого фактора.

Диабетическая ретинопатия

Имеются результаты обработки 1147 изображений сетчаток. По изображениям рассчитаны значения 17 признаков; записаны также результаты предварительного скрининга на наличие диабетической ретинопатии и окончательный диагноз (retinopathy.xlsx).

Ш-Щ: построить модель, оценивающую вероятность наличия диабетической ретинопатии, дать интерпретацию коэффициентов.

Массовая доля жира в организме

Массовая доля жира, важная характеристика здоровья, рассчитывается через плотность тела, измеряемую при помощи взвешивания в воде. Для 252 мужчин проведены такие расчёты. Имеются также данные антропометрии (возраст, рост, вес, обхват грудной клетки и т.д.) (fat.xls).

Э-Ю: построить функцию, оценивающую индекс ожирения без использования данных взвешивания.

Я: построить функцию, оценивающую массовую долю жира по легко измеряемым антропометрическим признакам.