Машинное обучение

НИЯУ МИФИ, Кафедра финансового мониторинга.

Лабораторный практикум.

В.Ю. Радыгин, Д.Ю. Куприянов

Семестр 1. Лабораторная работа 3

Лабораторная работа 3 рассчитана на два занятия и работу дома. Её целью является изучение основ анализа данных, заданных в номинальной шкале.

Вариант 1

Зарегистрируйтесь на сайте https://www.kaggle.com/datasets и загрузите с него набор статистических данных, посвящённый опросам людей https://www.kaggle.com/freecodecamp/2016-new-coder-survey-/version/1

1. На основе загруженного CSV-файла создайте Pandas DataFrame, подобрав правильные типы данных столбцов.
2. Создайте новый Pandas DataFrame, выбрав только переменные EmploymentField, EmploymentStatus, Gender, JobPref, JobWherePref, MaritalStatus, Income.
3. Удалите все наблюдения, содержащие либо значения поля пол (Gender), отличные от male или female, либо значения NA (нет ответа) в каких-либо из полей.
4. Исследуйте связи между парами переменных (используйте только наблюдения, где эти поля заполнены):
    a. Gender, JobPref;
    b. Gender, JobWherePref;
    c. JobWherePref, MaritalStatus;
    d. EmploymentField, JobWherePref;
    e. EmploymentStatus, JobWherePref.

    Выполняя исследование, не используйте процедуру ANOVA. Для каждой пары постройте таблицу сопряжённости, таблицу ожидаемых значений. Обоснованно выберите один из методов: хи-квадрат Пирсона, хи-квадрат Пирсона с поправкой Йейтса, точный критерий Фишера (обычный или на основе приближения Монте-Карло), точный критерий Фримана-Холтона (обычный или на основе приближения Монте-Карло).
5. Для каждой пары интерпретируйте результаты.
6. Замените переменную Income на три уровня дохода: низкий, средний, высокий.
7. Исследуйте связи между парой переменных Gender, Income (в новом формате) аналогично заданию 4. Интерпретируйте результаты.
8. С помощью процедуры ANOVA исследуйте, как доход зависит от остальных переменных. При этом проверьте:
    a. Нормальность распределения дохода (методы Жака (Харке)-Бера, Шапиро-Уилка, Андерсона-Дарлинга, Колмогорова-Смирнова):
        i. Если нормальность не выполняется, выполните лог-трансформацию дохода и проверьте заново.
        ii. Если нормальность не выполняется, ограничьте выборку 100 первыми записями.
    b. Отсутствие автокорреляции (тест Дарбина — Уотсона);
    c. Гомоскедастичность (Omnibus Test).
9. Дайте интерпретацию результатам.

Вариант 2

Зарегистрируйтесь на сайте https://www.kaggle.com/datasets и загрузите с него набор статистических данных, посвящённый опросам людей https://www.kaggle.com/freecodecamp/2016-new-coder-survey-/version/1

1. На основе загруженного CSV-файла создайте Pandas DataFrame, подобрав правильные типы данных столбцов.
2. Создайте новый Pandas DataFrame, выбрав только переменные EmploymentField, EmploymentStatus, Gender, LanguageAtHome, JobWherePref, SchoolDegree, Income.
3. Удалите все наблюдения, содержащие либо значения поля пол (Gender), отличные от male или female, либо значения NA (нет ответа) в каких-либо из полей.
4. Исследуйте связи между парами переменных (используйте только наблюдения, где эти поля заполнены):
    a. Gender, SchoolDegree;
    b. Gender, JobWherePref;
    c. JobWherePref, LanguageAtHome;
    d. EmploymentField, LanguageAtHome;
    e. EmploymentStatus, LanguageAtHome.

    Выполняя исследование, не используйте процедуру ANOVA. Для каждой пары постройте таблицу сопряжённости, таблицу ожидаемых значений. Обоснованно выберите один из методов: хи-квадрат Пирсона, хи-квадрат Пирсона с поправкой Йейтса, точный критерий Фишера (обычный или на основе приближения Монте-Карло), точный критерий Фримана-Холтона (обычный или на основе приближения Монте-Карло).
5. Для каждой пары интерпретируйте результаты.
6. Замените переменную Income на три уровня дохода: низкий, средний, высокий.
7. Исследуйте связи между парой переменных SchoolDegree, Income (в новом формате) аналогично заданию 4. Интерпретируйте результаты.
8. С помощью процедуры ANOVA исследуйте, как доход зависит от остальных переменных. При этом проверьте:
    a. Нормальность распределения дохода (методы Жака (Харке)-Бера, Шапиро-Уилка, Андерсона-Дарлинга, Колмогорова-Смирнова):
        i. Если нормальность не выполняется, выполните лог-трансформацию дохода и проверьте заново.
        ii. Если нормальность не выполняется, ограничьте выборку 100 первыми записями.
    b. Отсутствие автокорреляции (тест Дарбина — Уотсона);
    c. Гомоскедастичность (Omnibus Test).
9. Дайте интерпретацию результатам.

Вариант для сильной подгруппы (по желанию и по силам)

Вариант 1

Зарегистрируйтесь на сайте https://www.kaggle.com/datasets и загрузите с него набор статистических данных, посвящённый опросам людей https://www.kaggle.com/freecodecamp/2016-new-coder-survey-/version/1

1. На основе загруженного CSV-файла создайте Pandas DataFrame, подобрав правильные типы данных столбцов.
2. Создайте новый Pandas DataFrame, выбрав только переменные EmploymentField, EmploymentStatus, Gender, LanguageAtHome, JobWherePref, SchoolDegree, Income.
3. Удалите все наблюдения, содержащие либо значения поля пол (Gender), отличные от male или female, либо значения NA (нет ответа) в каких-либо из полей.
4. С помощью веб-сервера Flask реализуйте веб-приложение, позволяюще добавлять новые наблюдения.
5. Реализуйте веб-страницы, которые при обращении проводят исследование связи между парами переменных (используйте только наблюдения, где эти поля заполнены):
    a. Gender, SchoolDegree;
    b. Gender, JobWherePref;
    c. JobWherePref, LanguageAtHome;
    d. EmploymentField, LanguageAtHome;
    e. EmploymentStatus, LanguageAtHome.

    Выполняя исследование, не используйте процедуру ANOVA. Для каждой пары постройте таблицу сопряжённости, таблицу ожидаемых значений. Обоснованно выберите один из методов: хи-квадрат Пирсона, хи-квадрат Пирсона с поправкой Йейтса, точный критерий Фишера (обычный или на основе приближения Монте-Карло), точный критерий Фримана-Холтона (обычный или на основе приближения Монте-Карло). Выбор метода реализуйте алгоритмически.
6. Для каждой пары интерпретируйте результаты.

Вариант 2

1. Соберите данные о ценах на зерно, нефть, бензин, курсе доллара, ставке рефинансирования и уровне инфляции в определённой стране за последние 36 месяцев (1 значение на каждый месяц). Все цены должны быть выраженных в одинаковых единицах измерения.
2. Переведите данные о ценах в порядковую шкалу измерения, заменив числовые значения на значения ‘very low’, ‘low’, ‘medium’, ‘high’, ‘very high’ пропорционально диапазону исследуемых цен.
3. Исследуйте связи между любыми парами переменных (используйте только наблюдения, где эти поля заполнены):
    Выполняя исследование, не используйте процедуру ANOVA. Для каждой пары постройте таблицу сопряжённости, таблицу ожидаемых значений. Обоснованно выберите один из методов: хи-квадрат Пирсона, хи-квадрат Пирсона с поправкой Йейтса, точный критерий Фишера (обычный или на основе приближения Монте-Карло), точный критерий Фримана-Холтона (обычный или на основе приближения Монте-Карло).
4. С помощью процедуры ANOVA исследуйте, как уровень инфляции зависит от остальных переменных. При этом проверьте:

   a. Нормальность распределения дохода (методы Жака (Харке)-Бера, Шапиро-Уилка, Андерсона-Дарлинга, Колмогорова-Смирнова):
      i. Если нормальность не выполняется, выполните лог-трансформацию дохода и проверьте заново.
      ii. Если нормальность не выполняется, ограничьте выборку 100 первыми записями.
   b. Отсутствие автокорреляции (тест Дарбина — Уотсона);
   c. Гомоскедастичность (Omnibus Test).
5. Дайте интерпретацию результатам.

Варианты

1. Казахстан и Украина
2. Беларусь и Армения
3. Россия и Азербайджан