Статистический анализ для решения задач медицинской диагностики

Гундырев Иван Анатольевич

ML Meetup 19 июля 2017

Вводная информация

Что за задача ?

В медицине задача ранней диагностики заболеваний является одной из самых актуальных.

Врач и пациент.

Человек приходит к врачу, проходит опрос, диалог, назначаются обследования и по итогу врач ставит диагноз и выдает предписания.

Вводная информация

Что за задача ?

В медицине задача ранней диагностики заболеваний является одной из самых актуальных.

Врач и пациент.

Человек приходит к врачу, проходит опрос, диалог, назначаются обследования и по итогу врач ставит диагноз и выдает предписания.

Диагноз:

1. Здоров (на самом деле недообследован).
2. Болен.

Чем мы занимались?

Исследование

- Объект исследования — Биохимические параметры слюны.
- Цель исследования — Создание классификатора для решения задачи медицинской диагностики.

Чем мы занимались?

Исследование

- Объект исследования — Биохимические параметры слюны.
- Цель исследования — Создание классификатора для решения задачи медицинской диагностики.

Отношение со стороны врачей к идее:

- Ничего не выйдет, потому что слюна очень динамичная среда.
- Разработки методов неинвазивной диагностики на основе биохимических показателей слюны очень перспективны.

Что за данные?

Исследуются заболевания определенного органа. Выборка состоит из двух групп:

- Люди у которых уже диагностировано заболевание (Больные).
- Люди у которых нет исследуемых заболеваний (Здоровые).

Что за данные?

Исследуются заболевания определенного органа. Выборка состоит из двух групп:

- Люди у которых уже диагностировано заболевание (Больные).
- Люди у которых нет исследуемых заболеваний (Здоровые).

Недостатки выборки

- Вполне может быть, что заболевания различных органов дают похожие особенности в биохимических параметрах слюны.
- Здоровые - могут быть не вполне "здоровы".

Постановка задачи

Просто классика — Бинарная классификация!
Дано: таблица с 35 биохимическими параметрами слюны (все непрерывные) и принадлежность классу (ZDOR или NO) для каждого объекта.
Требуется: создать классификатор, который на основе биохимических параметров позволяет прогнозировать принадлежность классу.

Постановка задачи

Просто классика — Бинарная классификация!
Дано: таблица с 35 биохимическими параметрами слюны (все непрерывные) и принадлежность классу (ZDOR или NO) для каждого объекта.
Требуется: создать классификатор, который на основе биохимических параметров позволяет прогнозировать принадлежность классу.

Замечание: Пол и возраст

Биохимические параметры слюны неинформативны (?) для прогнозирования пола и возраста.

Средства решения

Язык R и его библиотеки.

«R разрабатывают очень умные люди и от этого много зла!»

В. Л. Аббакумов.

Средства решения

Язык R и его библиотеки.

«R разрабатывают очень умные люди и от этого много зла!»

В. Л. Аббакумов.

Методика решения

Всё как обычно:

- Визуализация (ggplot2 благословенный!)
- Отбор признаков (FSelector, Boruta, RFE).
- Построение классификаторов (caret).
- Финализация.
- Размышления.

Трудности сбора данных

Необходимо, чтобы у людей был подтвержденный диагноз или подтверждено его отсутствие. Плюс к этому выполнение ряда дополнительных условий для возможности корректного сбора диагностической информации.

Где взять таких людей ? — Только сотрудничество с медицинскими учреждениями.

Колокол ?

Если на графике увидишь колокол — не верь глазам своим!

Колокол ?

Если на графике увидишь колокол — не верь глазам своим!

Что делать?

Будьте внимательны и осторожны при использовании статистических критериев!
Всегда необходимо обоснование.

Суточные пробы - Проведен

Сдача слюны каждые три часа в течении суток.
Эксперимент показал высокую вариативность биохимических параметров слюны, что конечно же не новость и не открытие.
Как все знают сдавать анализы необходимо с утра и натощак.

Суточные пробы - Проведен

Сдача слюны каждые три часа в течении суток.
Эксперимент показал высокую вариативность биохимических параметров слюны, что конечно же не новость и не открытие.
Как все знают сдавать анализы необходимо с утра и натощак.

Утренние пробы каждый день или через день - Не проведен

Получить несколько образцов проб с одного человека - каждое утро в течении недели или через день — допустим утро понедельник, среда, пятница, воскресенье в течении двух недель. При этом желательно, чтобы человек сохранял свой обычный образ жизни (пробы сдаются утром и натощак, но простудные заболевания, прием новых медикаментов и другие факторы могут оказать влияние на биохимию слюны)

Учет инструментальных и методических неточностей - In Progress

Вопрос: Насколько существенные искажения могут внести в результат классификации погрешности сбора первичной диагностической информации ?
Биохимические анализы проводятся методами, допускающими погрешность измерений, поэтому для каждого базового образа (результата анализа конкретного пациента) целесообразно рассматривать образы с малыми отклонениями в каждом параметре как единый объект.

Теорема Байеса

$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)}, \tag{4.1}$$

где

1. $P(A)$ — априорная вероятность гипотезы A;
2. $P(A \mid B)$ — вероятность гипотезы A при наступлении события B (апостериорная вероятность);
3. $P(B \mid A)$ — вероятность наступления события B при истинности гипотезы A;
4. $P(B)$ — полная вероятность наступления события B.

Yandex Events - Data & Science: здоровье

- Как наука о данных помогает развитию медицины
  https://habrahabr.ru/company/yandex/blog/330152/
- Остальные доклады этой конференции
  https://events.yandex.ru/events/ds/15-apr-2017/

YAC - 2017

Секция Digital Health
https://events.yandex.ru/events/yac/30-may-2017/