Прикладная статистика и анализ данных

Задание 3

Правила:

- Дедлайн **02 марта 16:30**. После дедлайна работы не принимаются кроме случаев наличия уважительной причины.
- Выполненную работу нужно отправить на почту `mipt.stats@yandex.ru` , указав тему письма `"[asda] Фамилия Имя - задание 3"` . Квадратные скобки обязательны. Если письмо дошло, придет ответ от автоответчика.
- Прислать нужно ноутбук и его pdf-версию (без архивов). Названия файлов должны быть такими: `3.N.ipynb` и `3.N.pdf` , где `N` - ваш номер из таблицы с оценками.
- Решения, размещенные на каких-либо интернет-ресурсах не принимаются. Кроме того, публикация решения в открытом доступе может быть приравнена к предоставлении возможности списать.
- Для выполнения задания используйте этот ноутбук в качестве основы, ничего не удаляя из него.
- Никакой код из данного задания при проверке запускаться не будет.
- В каждой задаче не забывайте делать **пояснения и выводы**.

Баллы за задание:

- Задача 1 - 4 балла
- Задача 2 - 3 балла
- Задача 3 - 10 баллов
- Задача 4 - 3 балла

Задача 1

Пусть $\mathscr{X} = \mathbb{R}^2$ --- пространство признаков, $\mathscr{Y} = \{0, 1\}$ --- множество классов. Рассматривается квадратичный дискриминантный анализ. Условное распределение X при условии $Y = k$ равно $\mathcal{N}(a_k, \Sigma_k)$. Приведите примеры таких параметров a_k, Σ_k и вероятностей $\mathsf{P}(Y = k)$, при которых разделяющая поверхность является

- гиперболой;
- параболой;
- двумя параллельными прямыми;
- двумя пересекающимися прямыми.

Задача 2

Загрузите датасет `galapagos.csv` (https://github.com/txm676/sars/tree/master/R), в котором содержатся данные об островах на Галапагосском архипелаге:

- `island` -- наименование острова;
- `species` -- количество наблюдаемых видов растений на острове;
- `endemics` -- количество уникальных видов [*в задаче не рассматриваем*];
- `area` -- площадь (км^2);
- `elevation` -- высота (м);

- `nearest` -- расстояние до ближайшего острова (км);
- `scruz` -- расстояние до самого крупного острова архипелага Санта-Крус (км);
- `adjacent` -- площадь соседнего острова (км^2).

In [ ]:

```
1
```

Рассмотрим пуассоновскую (колличественную) регрессию. Какая ожидается зависимость отклика от признаков?

Подсказка: чему равно математическое ожидание отклика?

<...>

Постройте графики зависимости отклика `species` от всех признаков. Значения каких признаков лучше прологарифмировать?

In [ ]:

```
1
```

Обучите пуассоновскую регрессию по всем признакам, предварительно прологарифмировав некоторых из них. В R регрессию можно сделать функцией `glm` , указав конкретный тип обобщенной модели как `family = poisson()` . Напечатайте `summary` модели.

In [ ]:

```
1
```

Оставьте только значимые признаки, обучите модель еще раз и проинтерпретируйте полученные результаты. Какой смысл имеют коэффициенты модели?

In [ ]:

```
1
```

При использовании статистических свойств необходимо выполнить проверку предположений модели. Аналогом гомоскедастичности для гауссовской линейной модели в случае пуассоновской регрессии является равенство математического ожидания и дисперсии (это свойство пуассоновского распределения).

Аналогично гауссовской линейной модели можно определить устойчивые оценки дисперсии. Посчитайте и напечатайте ковариационную матрицу оценок коэффициентов модели с помощью функции `vcovHC` , рассмотрев тип оценки `HC3` .

In [ ]:

```
1
```

Напечатайте таблицу статистических свойств оценок коэффициентов и доверительные интервалы. Для этого можно использовать те же функции, что и для гауссовской модели.

In [ ]:

```
1
```

Придумайте на Галапагосском архипелаге еще один остров, задайте ему некоторые характеристики и назовите своим именем. Оцените, в каком интервале лежит ожидаемое количество видов растений на этом острове.

Указания. Установите в функции `predict` опцию `se.fit = TRUE` . Из полученных значений сформируйте доверительный интервал линеаризованного отклика. Доверительный интервал для ожидаемого отклика получите с помощью функции `<имя обученной модели>$family$linkinv` .

In [ ]:

```
1
```

Наконец, получите предсказательный интервал для количества растений на вашем острове. Для этого посчитайте предсказательные интервалы для пуассоновских случайных величин, параметры которых соответствуют границам доверительного интервала, построенного на предыдущем шаге. Объедините эти два интервала.

Задача 3

Кардиотокография (КТГ) — непрерывная одновременная регистрация частоты сердечных сокращений плода и тонуса матки с графическим изображением физиологических сигналов на калибровочной ленте. В настоящее время КТГ является ведущим методом наблюдения за характером сердечной деятельности, который из-за своей простоты в проведении, информативности и стабильности получаемой информации практически полностью вытеснил из клинической практики фоно- и электрокардиографию плода.

Для облегчения задачи диагностики, результаты кардиотокографии некоторых эмбрионов были классифицированы специалистами на нормальные и патологические. По показаниям приборов было сгенерировано некоторое количество признаков.

1. Скачайте данные по ссылке: https://archive.ics.uci.edu/ml/datasets/cardiotocography (https://archive.ics.uci.edu/ml/datasets/cardiotocography). Файл `xls` в `R` можно загрузить функцией `read_xls` пакета `readxl` . Перед загрузкой данных в `R` для удобства можно выполнить преобразования над самим файлом `xls` . Данные расположены в листе `Data` . Рассматриваемые нами признаки указаны в листе Description в колонке Features. Целевой меткой является столбец `NSP` . Для анализа мы не будем рассматривать данные, целевой меткой которой является `Suspect` .

In [ ]:

```
1
```

2. Проведите визуальный анализ данных. Например, можете построить оценки плотности по каждому признаку отдельно для каждого класса. Какие признаки лучше всего разделяют классы?

In [ ]:

```
1
```

3. Разделите выборку на обучающую и тестовую часть случайным образом в соотношении 4:1.

In [ ]:

```
1
```

4. Обучите логистическую регрессию по всем признакам. В R это можно сделать функцией `glm`, указав конкретный тип обобщенной модели как `family = binomial()`. Напечатайте `summary` модели. В чем причины такого поведения модели? Какие выводы можно сделать? Что нужно сделать, чтобы это исправить?

In [ ]:

```
1
```

5. Проверьте модель на линейность по значимым признакам, используя сглаженные диаграммы рассеивания. Для избежания влияния выбросов стройте диаграммы в интервале от 0.05-квантили до 0.95-квантили по значениям каждого признака. В качестве ширины ядра берите треть этого диапазона. Какие выводы можно сделать?

Указание. Используйте реализацию непараметрической регрессии из пакета `np`. Ширину ядра можно выставить с помощью параметра `bws`.

In [ ]:

```
1
```

6. Можно ли для каких-то из признаков, по которым не подтвердилась линейность модели, добиться линейности с помощью преобразований (в качестве преобразований можно взять логарифмирование, возведение в квадрат, взятие модуля, сдвиг, введение нескольких признаков и т.д.)?

In [ ]:

```
1
```

7. Оставьте только значимые и преобразованные признаки, обучите модель еще раз и проинтерпретируйте полученные результаты. Сильно выделяющиеся наблюдения по какому-либо признаку также имеет смысл удалить. Какой смысл имеют коэффициенты модели?

In [ ]:

```
1
```

8. Для некоторых объектов из тестовой выборки оцените вероятность наличия патологии. Постройте также доверительный интервал для этой вероятности.

Указание. См. указание в предыдущей задаче.

In [ ]:

```
1
```

9. Выберите один достаточно значимый признак. Постройте график доверительного интервала

вероятности наличия патологии при изменении этого признака. Значения других признаков зафиксируйте как средние значения по этим признакам.

In [ ]:

```
1
```

10. Подберите оптимальный порог классификации для получения максимальной точности прогноза в соответствии с некоторой метрикой качества классификации.

Задача 4

Рассмотрите пример с занятия про детектирование спама. Попытайтесь улучшить точность классификации по метрике F1 с помощью идей из бонусной части или же своих собственных идей.