Прикладная статистика и анализ данных.

Задание 5.

- Дедлайн **16 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 5`". Квадратные скобки обязательны. Если письмо дошло, придет ответ от автоответчика.
- По задачам 1-4, 6 прислать нужно ноутбук и его pdf-версию (без архивов). Названия файлов должны быть такими: `5.N.ipynb` и `5.N.pdf`, где `N` — ваш номер из таблицы с оценками.
- Задачу 5 необходимо оформить в tex'е и прислать `pdf` или же прислать фотку в правильной ориентации рукописного решения, где все четко видно.
- Решения, размещенные на каких-либо интернет-ресурсах не принимаются. Кроме того, публикация решения в открытом доступе может быть приравнена к предоставлении возможности списать.
- Не забывайте делать пояснения и выводы.

1. **(1 балл)** Можно ли на уровне значимости 0.05 считать, что последовательность чисел 1.05, 1.12, 1.37, 1.50, 1.51, 1.73, 1.85, 1.98, 2.03, 2.17 является реализацией случайного вектора, все 10 компонент которого независимые одинаково распределенные случайные величины? Вывод сделайте на основе коэффициентов корреляции.
2. **(1 балл)** Медицинская лаборатория проводит испытания нового препарата для лечения некоторого заболевания. Для исследований были отобраны 2500 больных. Некоторые из них принимали новый препарат, а другие — плацебо. В первой группе значимое улучшение состояния наблюдается среди 853 пациентов из 1719 пациентов, принимавших новый препарат. Во второй группе значимое улучшение наблюдается среди 369 пациентов из 781 пациентов, принимавших плацебо. Влияет ли новый препарат на улучшение состояния у пациентов?
3. **(1 балл)** Рассмотрим изображения образцов листьев из предыдущего домашнего задания, в котором вы построили проекцию данных на несколько первых главных компонент. Посчитайте и визуализируйте матрицу коэффициентов корреляции Пирсона для полученных проекций. То есть в клетке с индексами (i, j) должен быть коэффициент корреляции Пирсона между проекциями на i-ую и j-ую главные компоненты. Можете ли вы объяснить полученный эффект теоретически?
4. **(4 балла)** Примените UMAP к датасету из образцов листьев. Подберите оптимальные параметры исходя из четкости визуализации. Сравните результат с t-SNE.

   При обучении метод UMAP строит неориентированный случайный граф в исходном пространстве объектов. На датасете по образцам листьев для метода UMAP на 30 соседях:

(a) визуализируйте матрицу вероятностей ребер с помощью `plt.imshow`.

(b) посчитайте математическое ожидание количества ребер в графе.

(c) сгенерируйте реализацию случайного графа и посчитайте количество ребер в нем.

Вычисления проводите в логарифмах настолько, насколько это возможно. При реализации обратите внимание на функции `scipy.special.logsumexp`, `numpy.searchsorted`, `numpy.tril_indices` и класс `sklearn.neighbors.NearestNeighbors`.

5. **(7 баллов)** Даны выборки $X = (X_1, ..., X_n), Y = (Y_1, ..., Y_n)$. Рассмотрим произвольную функцию $f : \mathbb{R}^2 \to \mathbb{R}$, для которой выполнено $f(x, y) = -f(y, x)$. Для любых $1 \leqslant i < j \leqslant n$ положим $c_{ij}(X) = f(X_i, X_j)$. Обобщенным коэффициентом корреляции называется следующая статистика:

$$\widehat{r} = \frac{\sum\limits_{i<j} c_{ij}(X)c_{ij}(Y)}{\sqrt{\sum\limits_{i<j} c_{ij}(X)^2 \sum\limits_{i<j} c_{ij}(Y)^2}}$$

Докажите, что это действительно выборочный коэффициент корреляции, т.е. что $|\widehat{r}| \leqslant 1$ и значения ± 1 достигаются, а также что при верной гипотезе о независимости выборок $\mathsf{E}\widehat{r} = 0$.

Найдите коэффициенты корреляции Пирсона, Спирмэна и Кэндалла с помощью обобщенного коэффициента корреляции.

6. **(10 баллов)** Скачайте данные маркетинговых кампаний португальского банковского учреждения:

http://archive.ics.uci.edu/ml/datasets/Bank+Marketing

Цель задачи – с целью дальнейшего исследования понять, какие условия (среди 20 признаков) влияют на подписание клиентом срочного депозита (величина y). Для решения задачи воспользуйтесь методами анализа зависимостей, выяснив, какие характеристики клиента оказывают влияние на целевую переменную, и указав степень влияния.

Полученные результаты сравните с важностью признаков, полученной с помощью Random Forest, разбив предварительно данные на обучающую и тестовую части. Постройте также график точности классификации методом Random Forest в зависимости от количества деревьев в композиции для обучающей и тестовой частей данных.

Как меняется точность классификации, если для обучения брать только значимые признаки, полученные разными методами?

Напоминание: в случае отбора признаков с целью дальнейшего исследования можно применять методы, контролирующие FDR на уровне не более 0.1.