Прикладная статистика и анализ данных.

Задание 4.

- Дедлайн **10 марта 23:59**. 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 4`". Квадратные скобки обязательны. Если письмо дошло, придет ответ от автоответчика.
- По задачам 3-4 прислать нужно ноутбук и его pdf-версию (без архивов). Названия файлов должны быть такими: `4.N.ipynb` и `4.N.pdf`, где `N` — ваш номер из таблицы с оценками.
- Задачи 1 и 2 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. **(2 балла)** Пусть $X \in \mathbb{R}^{n \times d}$ — матрица признаков, а $Y \in \mathbb{R}^n$ — вектор отклика. Рассмотрим $\widehat{\theta}$ — оценка коэффициентов линейной модели методом ридж-регрессии. Представив матрицу X в виде сингулярного разложения распишите $\widehat{\theta}$, а также оценку отклика на обучающей выборке $\widehat{Y}$. Что происходит при отсутствии регуляризации?

2. **(4 балла)** Пусть $X_1, ..., X_n$ — выборка в пространстве $\mathbb{R}^D$, а $Y_1, ..., Y_n$ — ее проекция на линейное подпространство размерности $d < D$. Докажите, что величина

$$\sum_{i=1}^{n} (X_i - Y_i)^2$$

   минимальна, если $Y_1, ..., Y_n$ — проекция на линейное подпространство, образованное первыми d главными компонентами. Чему она равна?

3. **(4 балла)** В этой задаче нужно пронаблюдать экспериментально проблему скученности (crowding problem) в методе SNE. Все числа в задаче носят рекомендательный характер, можно выбрать свои.

   Рассмотрите трехмерный симплекс с вершинами в точках (1,0,0), (0,1,0), (0,0,1), (1,1,1). Вокруг каждой вершины сгенерируйте 50 точек из нормального распределения с масштабом 0.1, получив тем самым выборку в пространстве большой размерности. Результаты в задаче будут наглядными если точки выборки будут отсортированы по вершинам.

Посчитайте вероятности p_{ij} и визуализируйте матрицу их логарифмов с помощью `plt.imshow`. Для избежания ошибок округления проводите все вычисления в логарифмах и используйте функцию `scipy.special.logsumexp`. Для удобства работы с вероятностями p_{ii} можно использовать функцию `numpy.nan_to_num` там, где это необходимо. Перплексию возьмите равной 10-15. Числа σ_i можно честно посчитать для каждой точки, а можно и сделать приближение, взяв их одинаковыми для всех точек в силу локальной однородности данных.

В пространстве малой размерности рассмотрите квадрат со стороной 10 и аналогичным образом сгенерируйте выборку, разброс точек вокруг вершин равен 1. Посчитайте вероятности q_{ij}, определяемые методом SNE, и визуализируйте матрицу. Какой наблюдается эффект? Посчитайте дивергенцию Кульбака-Лейблера.

Для некоторого объекта i визуализируйте зависимость p_{ij} и q_{ij} от j. При минимизации дивергенции Кульбака-Лейблера вероятности q_{ij} должны приближать p_{ij}. Что будет в таком случае?

Повторите те же операции для q_{ij}, определяемых методом t-SNE. Сторону квадрата возьмите равной 1000, разброс точек 100. Какой наблюдается эффект?

4. **(12 баллов)** Рассмотрим датасет **Leaf Classification**:

   https://www.kaggle.com/c/leaf-classification

   Данные содержат 1584 изображений образцов листьев (16 изображений для 99 видов). По ссылке доступно подробное описание данных. Для вашего удобства размер некоторых изображении был изменен, в результате чего все изображения имеют одинаковый размер 170×250. Скачайте файл с данными на сайте курса.

   - Загрузите все изображения с помощью `plt.imread` и визуализируйте некоторые из них. Каждое изображение — матрица размера 170×250.
   - В файле `train_labels.csv` указаны номера образцов листьев, которые относятся к обучающей части данных, а так же их виды. Разделите данные на обучающую и тестовую часть.
   - На обучающей части данных постройте 30 главных компонент. Какую долю дисперсии данных они объясняют? Какую долю дисперсии объясняет каждая компонента отдельно?
   - Визуализируйте главные компоненты. Можете ли вы их как-то охарактеризовать?
   - Визуализируйте обучающую часть данных в проекции на две первых главных компоненты. Цвет точки должен соответствовать виду образца. Используйте `cmap='Set1'` во избежании градации цвета по номеру вида. Наблюдаются ли какие-либо закономерности?
   - Визуализируйте данные при помощи t-SNE двумя способами — на основе исходных признаков (пиксели) и по проекциям на первые 30 главных компонент. Кластеризуются ли точки?
   - По проекциям данных на первые 30 главных компонент обучите многоклассовую классификацию. Для образцов из тестовой части данных оцените вероятности принадлежности к классам.