Домашнее задание №8 по курсу «Машинное обучение»

Колесов Алексей

10 ноября 2019 г.

1 Задания

Пожалуйста, пройдите опрос `https://forms.gle/XiHoXGmzXeEbMWU1A`. Он поможет нам понять, как можно сделать этот курс лучше. Опрос можно проходить анонимно, но нам будет приятней и полезней, если вы оставите свою фамилию (в опросе есть специальное поле). Прохождение опроса не влияет на оценку.

1. Примените k-NN алгоритм к датасету «ирисы Фишера»: `https://www.kaggle.com/vikrishnan/iris-dataset`. Какие параметры можно перебрать? Нужно сдать реализацию и отчёт о решении.

2. Примените PCA-преобразование к датасету MNIST `http://yann.lecun.com/exdb/mnist/` для разных размерностей пространства. Сравните качество k-NN алгоритма для исходного датасета, восстановленных версий изображений для разных размерностей, «сжатых» представлений для каждой размерности. Прокомментируйте результаты. Вы можете использовать любую доступную реализацию PCA, всё остальное нужно реализовать самим.

3. Реализуйте алгоритм k-means для точек из $\mathbb{R}^d$ и евклидовой метрики. Реализуйте генерацию синтетического датасета: n объектов в $\mathbb{R}^d$, $d \geqslant 3$. Каждый объект равновероятно генерируется из одного из трёх распределений: каждое распределение многомерное нормальное с единичной матрицей ковариаций и матожиданием равным se_i, где s – скаляр, e_i – i-й вектор стандартного ортонормированного базиса (вектор из всех нулей, кроме единицы на i-м месте). Для каждого $s \in (0.5, 1, 1.5, \ldots, 10)$ запустите алгоритм k-means 10 раз (каждый раз для нового датасета) при $n = 10^5$, $d = 10$, $k = 3$. Сравните кластеры и номер распределения, из которого приходили объекты. Учитывая, что кластеры эквиваленты с точностью до перестановки, предложите способ, как вычислить долю правильно "классифицированных" объектов с помощью k-means кластеризации. Как эта доля зависит от s?