

Машинное обучение и интеллектуальные системы.

Практикум.

Задания.

1. (*) Набор данных [Davis](#) содержит рост, вес и пол 200 человек. Построить модель, показывающую приблизительную зависимость веса от роста и пола. Оценить вклад роста и пола в вес.
Построить гистограмму величины (repwt - weight), т.е. разницы между заявленным человеком весом и его реальным весом. Аналогично для двух разных полов.
2. (*) Набор данных [Wifi localization](#) содержит в себе качество сигнала для каждой из семи сетей wi-fi и комнату, в которой было снято наблюдение. Построить модель, предсказывающую комнату по семи сигналам.
3. (**) Набор данных [mushrooms](#) содержит в себе описание множества грибов по 22 атрибутам (особенности шляпки, запах, особенности пластинок, ножки, покрывала, среда обитания и др.). Построить модель, предсказывающую съедобность и ядовитость гриба по этим признакам. Отобрать наиболее важные признаки. Попытаться построить модель без самого доминирующего признака.
4. (**) Набор данных [California housing](#) описывает множество районов Калифорнии через такие признаки, как географическое положение (lon/lat), медианный возраст домов, население, общее количество комнат и спален, близость к океану (категориальная), количество домов, медианный доход и медианная стоимость дома. Построить модель, предсказывающую медианную стоимость по остальным признакам. Определить наиболее информативные признаки. Построить модель исключив географическое положение.
5. (**) Используя набор данных [Rosetta](#), построить классификатор, определяющий на каком из 30-ти языков программирования написан код. Определить, какие языки чаще путаются между собой и какие признаки доминируют в принятии классификатором решений.
6. (***) Набор данных [20 newsgroups](#) содержит в себе посты из 20 тем форума. Назначить 50% постов из тестовой выборки неправильную тему и построить модель, которая определяет, является ли заданный пост офф-топиком для заданной темы. Найти баланс между количеством пропущенного офф-топика и количеством “ложных тревог”.

Примечания к заданиям.

5. Код взят с сайта rosetta-code, и представлен в двух версиях - сырой и токенизированной простейшим образом (выделение чисел и си-подобных идентификаторов, удаление пробелов). Не факт, что эта токенизация нужна. Файл rosetta_tokenize.py описывает токенизацию, если нужно её использовать на новых данных.

Общая ссылка на данные

https://drive.google.com/open?id=1YfzufFH0of48TKgZFzjO7cBtr_RLiB8X