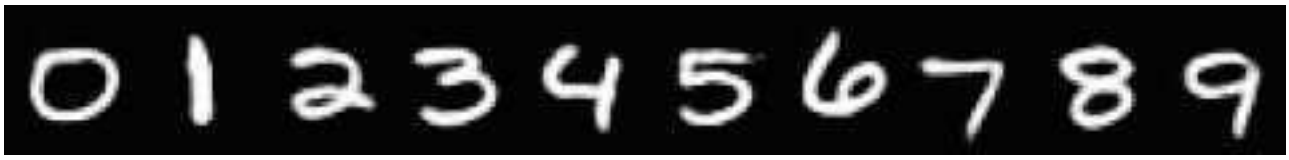
Машинное Обучение МФТИ

Практическое задание №1: Метрические Алгоритмы



В качестве первой домашней работы студентам предлагается принять участие в соревновании по распознаванию рукописных цифр на Kaggle. Для выполнения домашнего задания используйте ipython notebook с небольшими заготовками кода.
Соревнование общее – бонусы получают три первые места в каждой учебной группе.

Настройка окружения, инструкция по отправке решения

1. Вы можете загрузить `01_knn_start_code.zip`, в этом файле находится ноутбук с кодом, который необходимо доделать, следуя инструкциям =).

2. Установите https://www.continuum.io/downloads для **Python 2.7** – это важно.

3. Для настройки окружения задания выполните

```
unzip 01_knn_start_code.zip
cd 01_knn_start_code
pip install -r req.txt
ipython notebook
```

4. Сохранить ipnb с выполненным экспериментом. Нужно запаковать, переименовать `01_knn_start_coc` и отправить в приватный канал в piazza. В ноутбуке должны быть описаны эксперименты, которые показали улучшение. Приведен код который их выполняет. Можете описать интересные идеи которые не сработали.

Описание данных и метрики

Для решения задания нужно обучить модель на обучающей выборке `train.csv`, сделать предсказание на `validation.csv` и оправить свое решение на kaggle. Для ранжирования решений на `ladboard` используется метрика *accuracy* – доля правильно классифицированы объектов в валидационной выборке.

$$Accuracy = \frac{num\ right\ answers}{num\ all\ answers} = \frac{\sum_i^N [\hat{y}_i = y_i]}{N}$$

Каждый объект в обучающей выборке (`train.csv`) представляет собой черно белое квадратное изображение размера 27 пикселя и метку класса, к которому относится этот обект. Изображение построчно вытянуто в одну большую строку размера $27 \cdot 27 = 729$ в каждом элементе с индексом $row \cdot 27 + column$ находится интенсивность $I_{row,clumn}$ пикселся.

Задания которое нужно выполнить

1. Визуализируйте внутриклассовые центры

2. Реализовать метод K-ближайших соседей с метрикой L2:

   (a) Матрицей попарных расстояний используя: 2 цикла, 1 цикл, не используя циклов

   (b) KD tree (используйте класс `sklearn.neighbors.KDTree`)

В чем достоинства и недостатки предложенных методов, какие улучшения можно предложить?

3. Реализуйте алгоритм кросс валидации
4. Настройте параметры алгоритма: число ближайших соседей, метрику, ядро. Визуализируйте зависимость качества от настраиваемых параметров.
5. На основе отступа реализуйте удаление шумовых объектов, улучшилось ли-качество, почему это произошло?
6. Для того, чтобы заданнее было оценено выше 0 баллов, ваше итоговое решение должно превысить превысить KNN-бенчмарк
7. [Придумайте сами] Получите улучшение за сечет различных модификаций:
   (a) метрических алгоритмов
   (b) преобразований на объектами обучающей выборки
8. **Задание считается сданным после отправки ipython notebook, с описанием и кодом проведенных экспериментов, наглядными графиками и правильными выводами**

Методические указания

1. При подборе параметров модели рекомендуется использовать только часть обучающей выборки, для того чтобы сократить время обучения.
2. Согласно правилам соревнований нельзя делать больше 2х коммитов в систему в сутки. Из этого надо сделать следующие выводы:
   (a) Обучаться нужно локально (cross validation) и только после получения результата, который вы считаете удовлетворительным, нужно делать submit в систему.
   (b) Начать делать домашнее задание стоит заблаговременно.
3. Обратите внимание, что публичные результаты на kaggle рассчитываются только по части контрольной выборки, и будут рассчитаны по всей контрольной выборке после окончания соревнования. Будьте аккуратны с переобучением.
4. Победители получают бонусные балы – шарить решение не выгодно.

Разница между списыванием и помощью товарища иногда едва различима. Мы искренне надеемся, что при любых сложностях вы можете обратиться к семинаристам и с их подсказками самостоятельно справиться с заданием. При зафиксированных случаях списывания (одинаковый код, решение задачи), баллы за задание будут обнулены всем участникам инцидента.