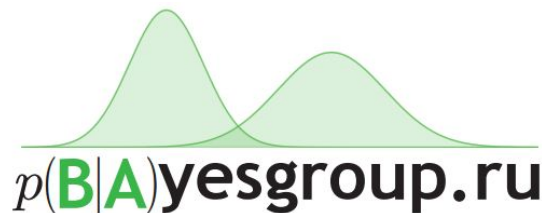


Майнор “Интеллектуальный анализ данных”

Надежда Чиркова

Международная лаборатория глубинного обучения и байесовских методов
Приглашенный преподаватель, факультет компьютерных наук



Что за майнор такой

- Первый запуск: 2015/16 учебный год
 - 2 потока, 22 группы по ~15 студентов
 - 1 группа = 1 факультет
- Второй запуск: 2016/2017 учебный год
 - 5 групп по ~35 студентов
 - факультеты кластеризованы по группам
- Третий запуск: текущий год

Курсы майнора

- Введение в программирование
- Введение в анализ данных
- Современные методы машинного обучения
- Прикладные задачи анализа данных

Программа курса Введение в анализ данных

- Вводная лекция и семинар
- Первые несколько занятий - вводные темы
 - Лекции: основы линейной алгебры, математического анализа, теории вероятностей, математической статистики, методов оптимизации
 - Семинары: знакомство с библиотеками numpy, pandas, matplotlib, sklearn
- Метрики качества
- Метрические методы
- Линейные методы
- Решающие деревья
- Композиции: бэггинг, случайный лес
- Понижение размерности и визуализация: PCA, t-SNE
- Отбор признаков
- Кластеризация
- Рекомендательные системы
- Ранжирование
- Transfer learning

Программа курса Современные методы машинного обучения

- Метод опорных векторов, ядра
 - Бустинг
 - Нейронные сети
 - Распределения и статистики
 - Проверка гипотез
 - Анализ зависимостей
 - Регрессия
 - Прогнозирование временных рядов
-
- 1 модуль
- 2 модуль

Программа курса Прикладные задачи анализа данных

- Анализ текстов
- Рекомендательные системы
- Ассоциативные правила
- Распознавание изображений

Компоновка курса по машинному обучению

Прикладная компонента:

- примеры задач
- внедрение решений

Математическая компонента:



- формальная постановка задач
- разбор алгоритмов обучения
- решение теоретических заданий

Программистская компонента:




- использование библиотек (pandas, sklearn)
- проведение экспериментальных исследований методов
- самостоятельная реализация методов (numpy, pytorch / tensorflow)

Компоненты курса Машинное обучение на ФКН




Прикладная компонента:

- примеры задач 
- внедрение решений 

Математическая компонента:



- формальная постановка задач 
- разбор алгоритмов обучения 
- решение теоретических заданий 

Программистская компонента:




- использование библиотек (pandas, sklearn) 
- проведение экспериментальных исследований методов 
- самостоятельная реализация методов (numpy, pytorch / tensorflow) 

Компоненты майнора




Прикладная компонента:

- примеры задач 
- внедрение решений 

Математическая компонента:



- формальная постановка задач 
- разбор алгоритмов обучения 
- решение теоретических заданий 

Программистская компонента:




- использование библиотек (pandas, sklearn) 
- проведение экспериментальных исследований методов 
- самостоятельная реализация методов (numpy, pytorch / tensorflow) 

Компоненты вводного курса Data Culture




Прикладная компонента:

- примеры задач 
- внедрение решений 

Математическая компонента:

- формальная постановка задач 
- разбор алгоритмов обучения 
- решение теоретических заданий 

Программистская компонента:

- использование библиотек (pandas, sklearn) 
- проведение экспериментальных исследований методов 
- самостоятельная реализация методов (numpy, pytorch / tensorflow) 

Лекции

- Со слайдами или с доской?

Градиентный спуск

- Повторять до сходимости:

$$w^t = w^{t-1} - \eta \nabla Q(w^{t-1})$$

Новая точка

Размер шага

Градиент в предыдущей точке

Проклятие размерности

- Задача: классификация пончиков на вкусные и невкусные
- 100 объектов
- Цвет: 10 вариантов
- Цвет + размер: $10 * 4 = 40$ вариантов
- Цвет + размер + форма: $10 * 4 * 4 = 160$ вариантов



Семинары

- Повторение основных понятий и формул с лекции
- Решение несложных теоретических задач на понимание метода
- Работа с методом в python
 - реализация части алгоритма или алгоритма целиком с помощью numpy
 - проведение экспериментов на конкретных данных (изучение влияния гиперпараметров на качество, визуализация результатов работы алгоритма, сравнение реализаций)

Семинары: практические аспекты

Добавление нескольких тем, связанных с обучением модели по конкретным данным:

- Правильная настройка композиций
- Методы визуализации данных
- Методы генерации признаков

Много материалов в блоге Александра Дьяконова:

<https://alexanderdyakonov.wordpress.com/>

Система оценивания

Теория:

- Проверочные работы на семинарах
- Коллоквиум (устный или письменный) в конце 3 модуля
- Экзамен

Практика:

- Практические задания
 - Соревнования по анализу данных (опционально за бонусы)
-
- 2015/16 учебный год: индивидуальные проекты

Система оценивания

Теория:

- Проверочные работы на семинарах
- Коллоквиум (устный или письменный) в конце 3 модуля
- Экзамен

Практика:

- Практические задания anytask.org
 - Соревнования по анализу данных (опционально за бонусы)
-
- 2015/16 учебный год: индивидуальные проекты

Система оценивания

Теория:

- Проверочные работы на семинарах
- Коллоквиум (устный или письменный) в конце 3 модуля
- Экзамен

Практика:

- Практические задания anytask.org жесткие дедлайны или мягкие?
- Соревнования по анализу данных (опционально за бонусы)
- 2015/16 учебный год: индивидуальные проекты

Практические задания

Заполнить jupyter notebook:

- дописать недостающий код
- сделать выводы по результатам экспериментов

Тематика заданий прошлого года:

1. Задачи на numpy и pandas (ответить на вопросы по набору данных, выполнить его преобразования - добавить признаки и т. д.)
2. Задание по линейным методам и классификации текстов (сравнить работу методов, подобрать оптимальные гиперпараметры, попробовать модификации методов)
3. Задание по визуализации, кластеризации и генерации признаков (с изображениями, чтобы проще было интерпретировать результаты).

Пример задачи для разминки

Пусть даны матрица объекты-признаки, вектор правильных ответов и вектор весов:

$$X \in \mathbb{R}^{\ell \times d}, y \in \mathbb{R}^{\ell}, w \in \mathbb{R}^d$$

Задача обучения регуляризованной линейной регрессии выглядит в векторном виде так:

$$\|Xw - y\|^2 + \lambda \|w\|_1 \rightarrow \min_w$$

Найдите ошибки в развернутой записи этого функционала:

$$\left(\sum_{i=1}^{\ell} \sum_{j=1}^d x_{ji} w_i - y_i \right)^2 + \lambda \sum_{i=1}^{\ell} w_i \rightarrow \min_w$$

Пример задачи

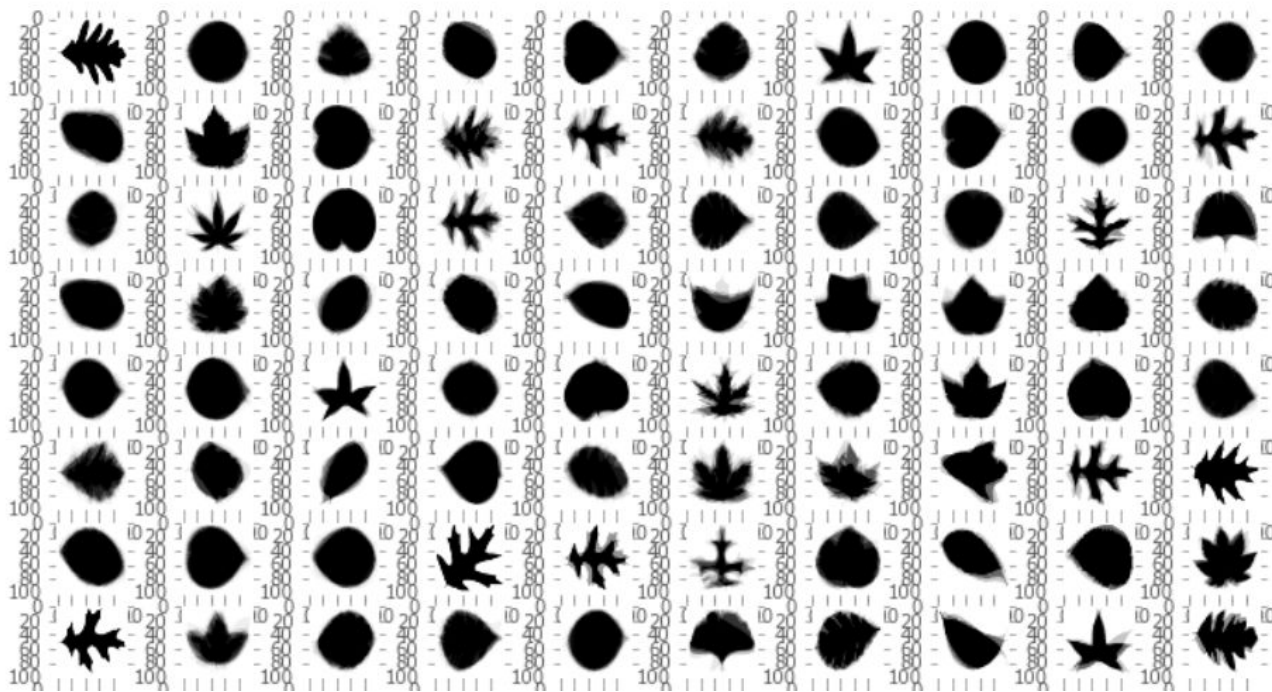
- Расстояние Жаккарда между множествами A и B :

$$J(A, B) = 1 - \frac{|A \cap B|}{|A \cup B|}$$

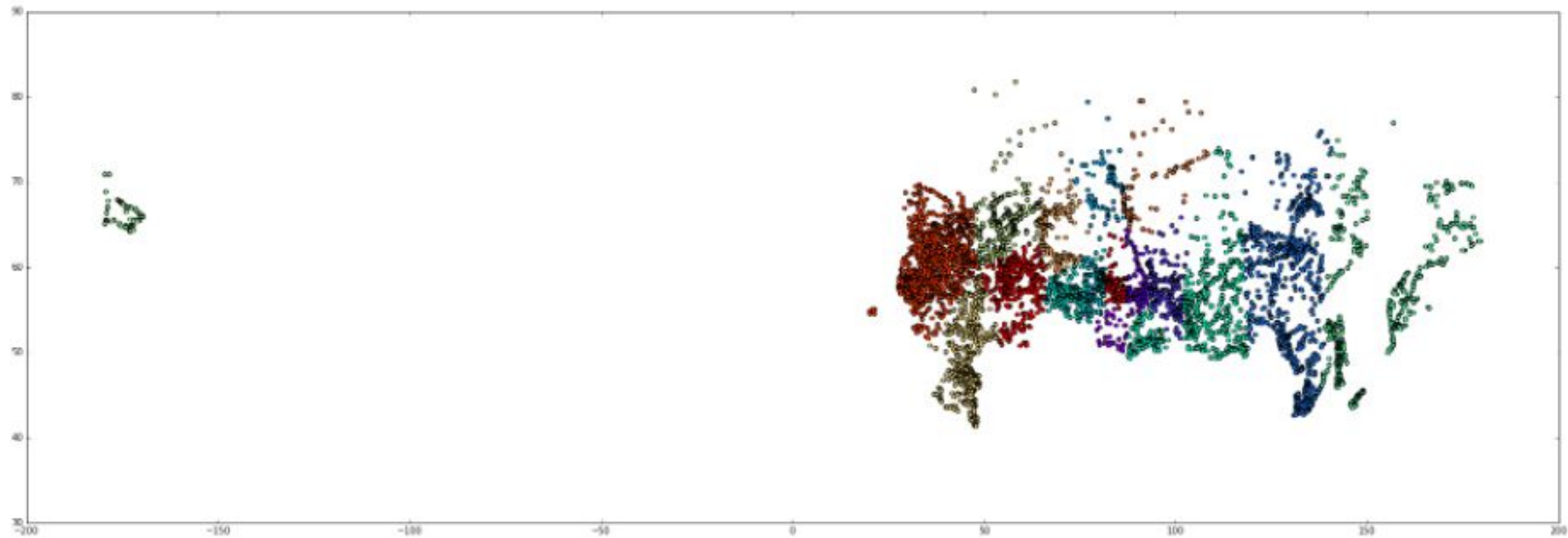
- Выберем супермножество, содержащие все элементы A и B .
- Закодируем каждое множество бинарным вектором: 1 ставится в том случае, если соответствующий элемент принадлежит множеству. Получим вектора a и b .
- Запишите расстояние Жаккарда через скалярные произведения и нормы векторов a и b .

Задание с изображениями листьев

```
1 km = KMeans(n_clusters = 100, random_state=random_seed)
2 km.fit(imgs_train)
3 plot_centers(km.cluster_centers_, np.sqrt(km.cluster_centers_.shape[0]))
```



Семинар по k-Means



Заключение

- Майнор дает достаточно широкие познания в области анализа данных, и при желании по его окончании можно пойти углублять знания в ШАД или с помощью онлайн-курсов
- На майноре много практики не только в домашней работе, но и на семинарах
- На майноре студентам приходится тратить много времени на самостоятельную работу, но по их отзывам оно того стоит