ПРОГРАММИРОВАНИЕ CUDA C/C++, АНАЛИЗ ИЗОБРАЖЕНИЙ И DEEP LEARNING

Лекция №8

Спасёнов Алексей

Введение в Машинное обучение

Содержание

1. Извлечение признаков
2. Преобразование признаков

Введение в Машинное обучение

Признаки (Features)

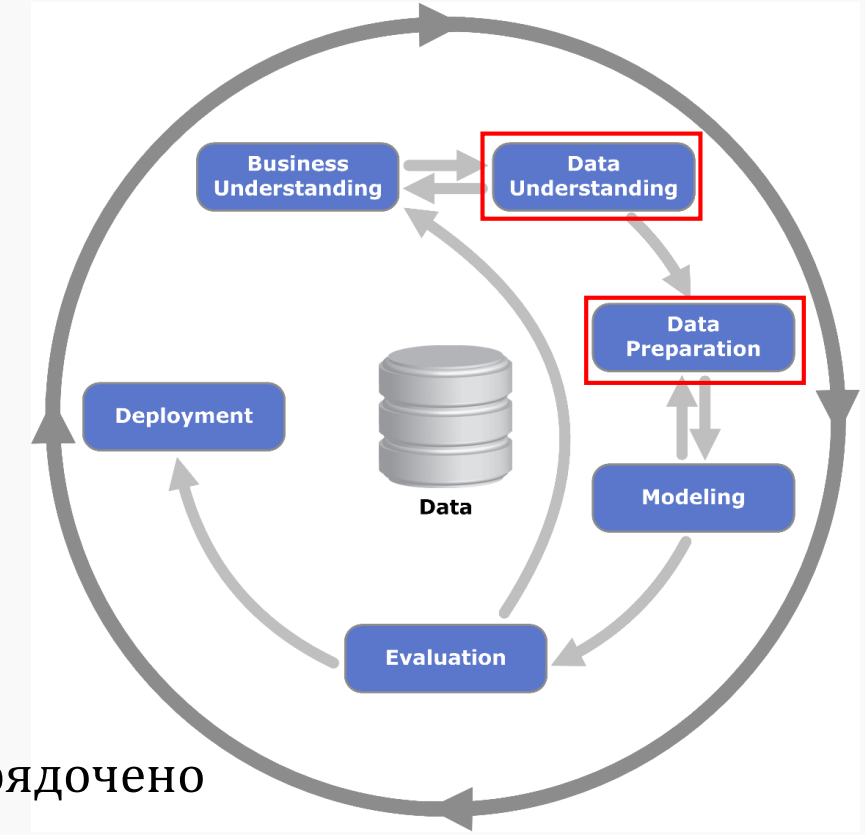
D – множество объектов (Data set)

$d \in D$ – обучающий объект

$\phi_i : D \rightarrow F_j$ - признак

Виды признаков:

1) Бинарные	Binary	$F_j = \{true, false\}$
2) Номинальные	Categorical	F_j – конечно
3) Порядковые	Ordinal	F_j – конечно упорядочено
4) Количественные	Numerical	$F_j = \mathbb{R}$



Извлечение признаков

Даты и время

1) Абсолютное время события

(2017:04:22 16:25:00)

2) Использование периодичности

(месяц, день, неделя и т.д.)

3) Временной интервал до или после особого события

(день выдачи зарплаты, праздник и т.д.)

Извлечение признаков

Геоданные

Чаще всего представлены парой ширина-долгота

Задачи:

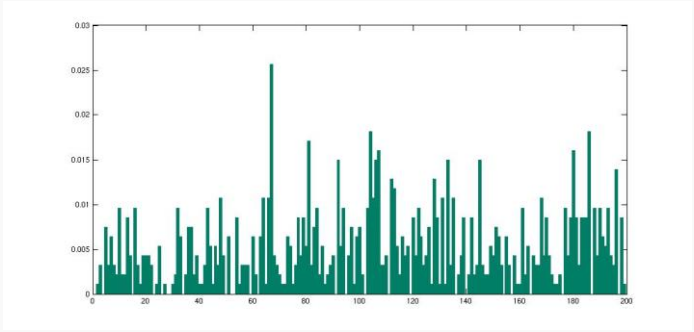
1) Восстановление точки из адреса (геокодинг)
2) Обратная задача

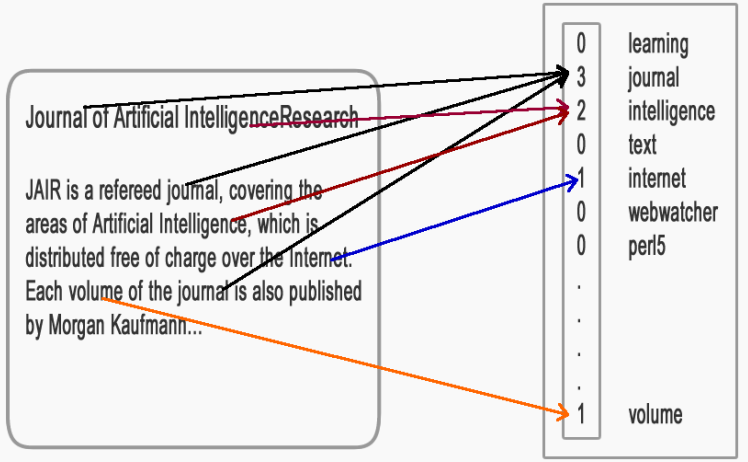
1) Расстояние до объектов внутри выборки
2) Расстояние до особых объектов

Извлечение признаков

Тексты

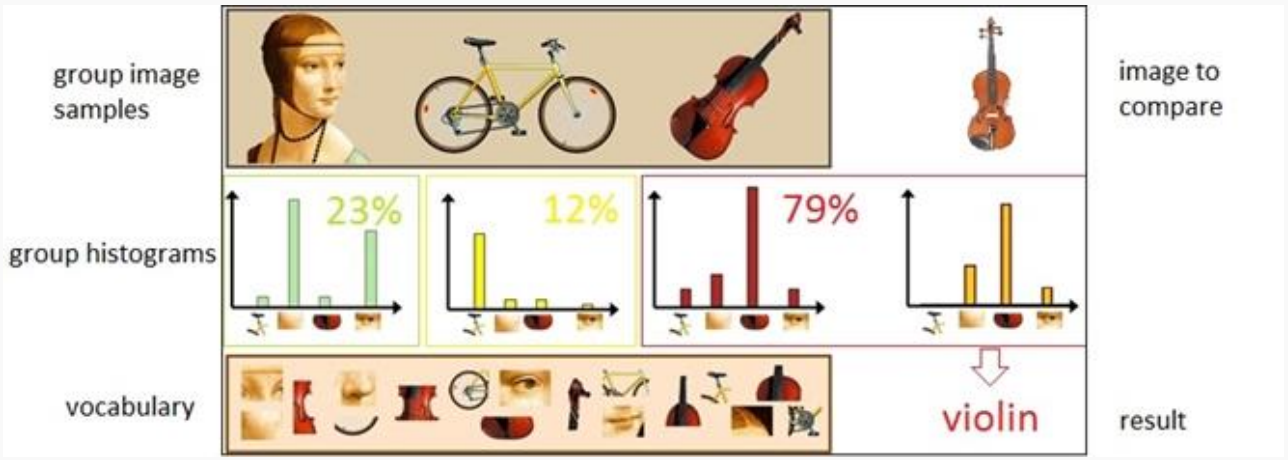
Bag-of-words





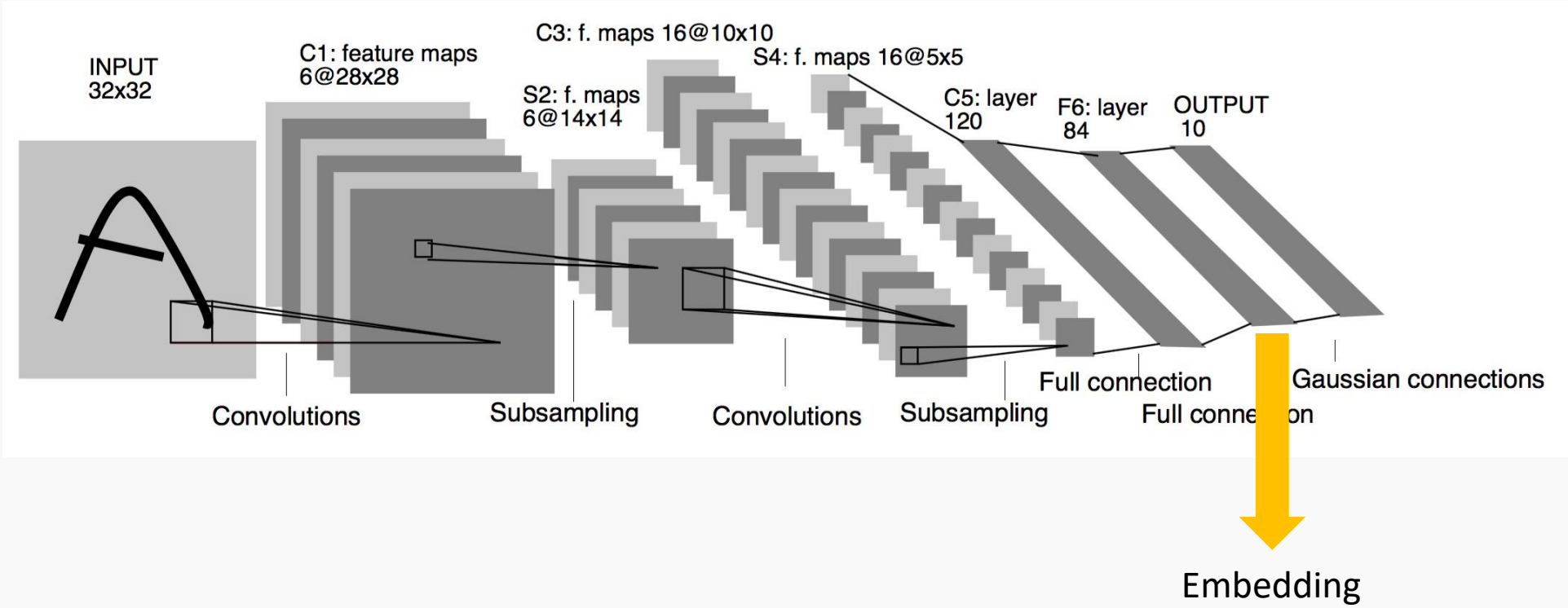
Извлечение признаков

Изображения



Извлечение признаков

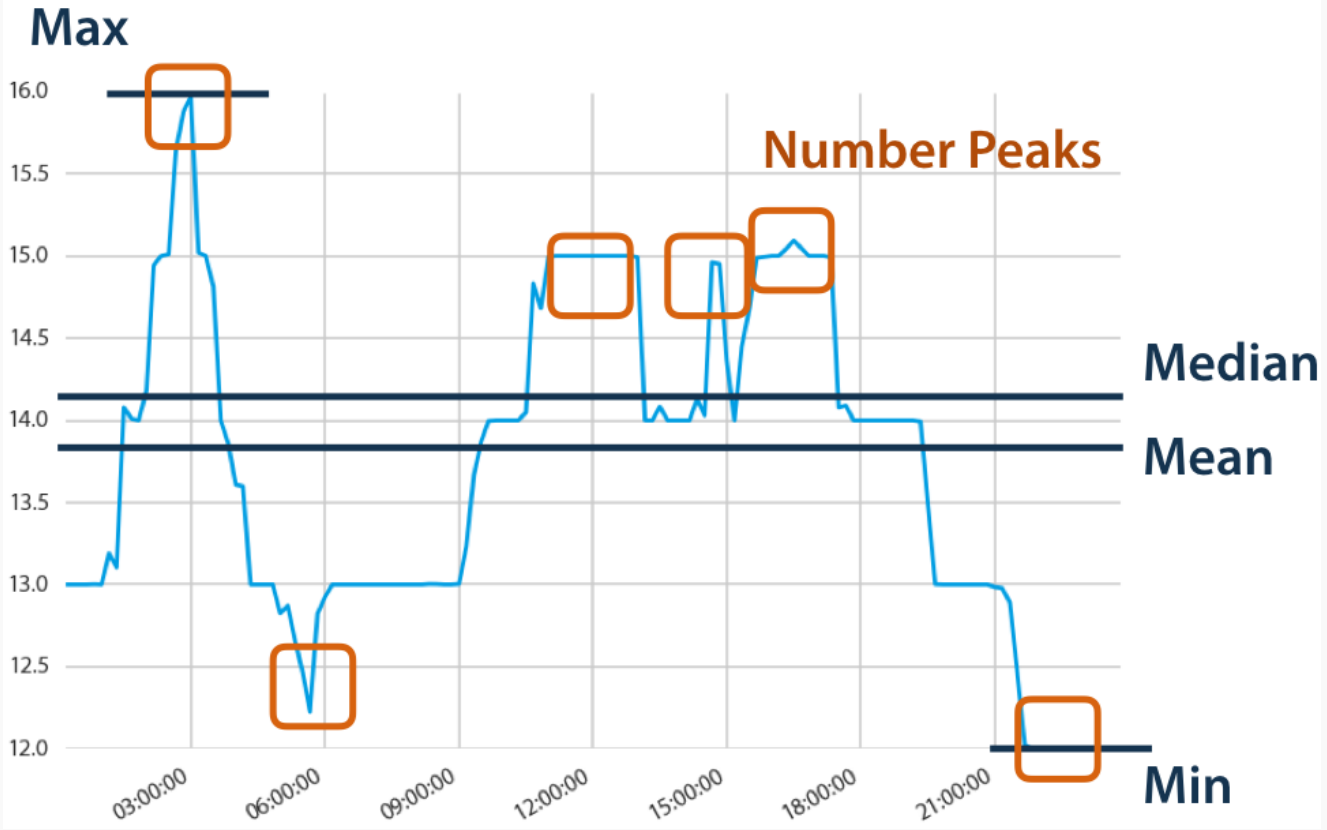
Изображения



Извлечение признаков

Временные ряды

https://github.com/blue-yonder/tsfresh



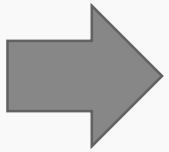
Преобразование признаков

Категориальные признаки

Label Encoding

Пример: имеется текстовое описание признаков

	Feature
1	School
2	Basic
3	University
4	School



	Feature
1	1
2	0
3	2
4	1

Не подходит для линейных моделей

http://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.LabelEncoder.html

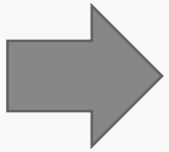
Преобразование признаков

Категориальные признаки

One-Hot Encoding

Пример: имеется текстовое описание признаков

	Feature
1	School
2	Basic
3	University
4	School



	F=School	F=Basic	F=University
1	1	0	0
2	0	1	0
3	0	0	1
4	1	0	0

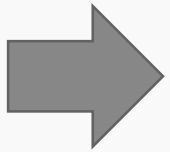
http://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html

Преобразование признаков

Категориальные признаки

Hashing trick

	Feature
1	School
2	Basic
3	University
4	School



	F=S	F=B,U
1	1	0
2	0	1
3	0	1
4	1	0

http://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.FeatureHasher.html

Метапризнаки

В качестве признаков можно использовать предсказания обученных моделей

	Random forest	Logistic regression	xgboost
train	0.634	0.726	0.801
train	0.461	0.294	0.310
test	0.717	0.582	0.847

Преобразование признаков

Нормализация

Различные модели по-разному реагируют на возможные значения входных признаков

1) Standart Scaling $z = \frac{x-\mu}{\sigma}$

http://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html

2) MinMax Scaling $x_N = \frac{x-x_{min}}{x_{max}-x_{min}}$

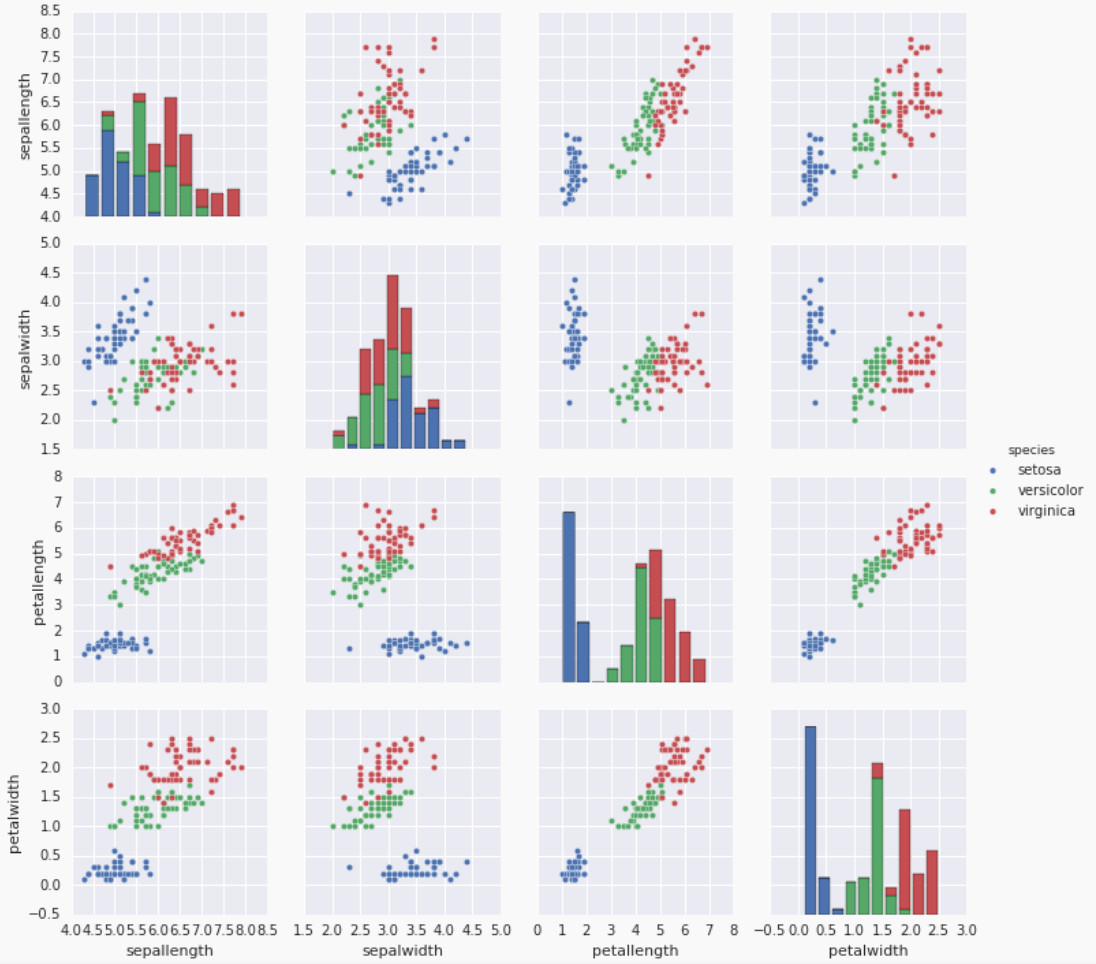
http://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html

http://scikit-learn.org/stable/modules/preprocessing.html

Преобразование признаков

Понижение размерности

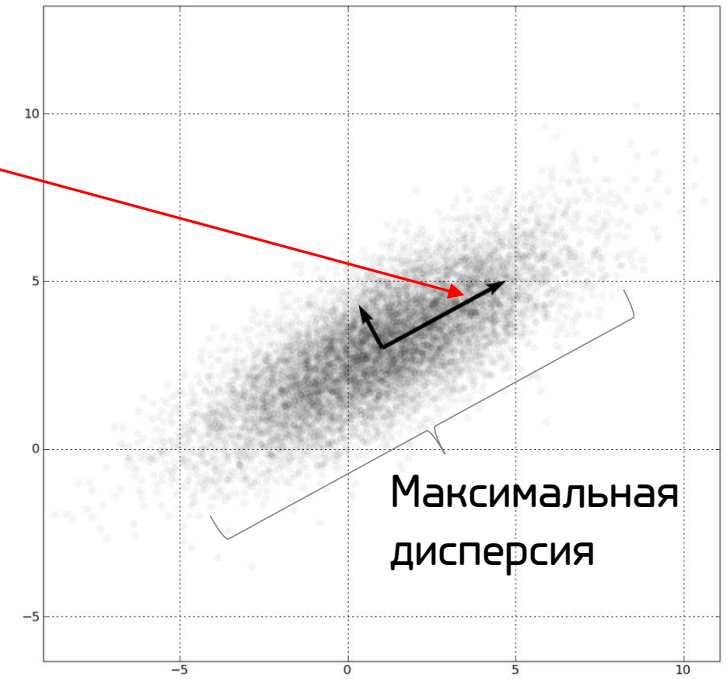
Признаков может быть намного больше!



Преобразование признаков

Понижение размерности – PCA (Principal component analysis)

Первая главная компонента

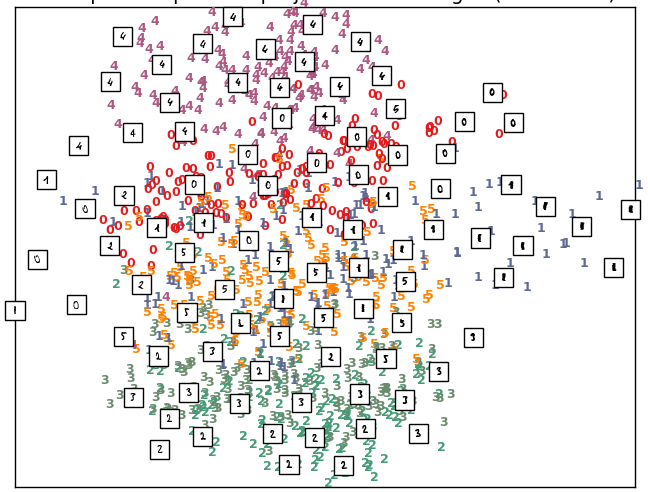


Преобразование признаков

Понижение размерности – PCA (Principal component analysis)

A selection from the 64-dimensional digits dataset

Principal Components projection of the digits (time 0.01s)



http://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html

Заполнение пропусков

Возможные операции:

1) Замена пропуска значением вида n/a

2) Выбор наиболее вероятного значения (выбор среднего или медианы)

3) Выбор наиболее невероятного значения

4) Выбор ближайшего значения (следующего или предыдущего, например для временных рядов)

Контакты:
a.spasenov@corp.mail.ru
alex_spasenov (Skype)

Спасибо за внимание!