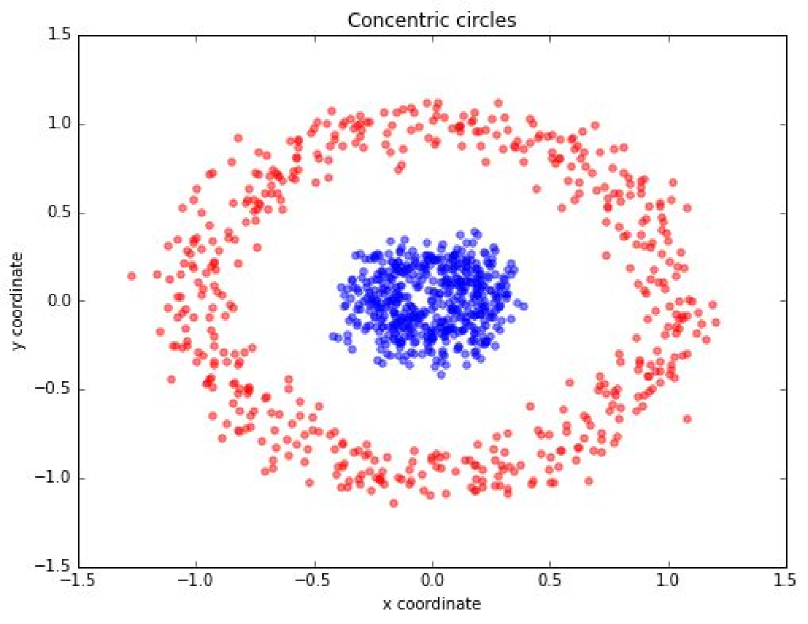
Продвинутые фичи, а также знания полезные в конкурсах на Kaggle и не только

Дмитрий Ульянов

Features based on nearest neighbors

- It is beneficial to add features, that describe geometry of the manifold
    - **Density**:



Features based on nearest neighbors

- Mean target of nearest 5, 10, 15, 500, 2000 neighbors (KNN)
    - Optionally use a weighting scheme

- Mean distance to 5, 10, ... closest neighbors

- Mean distance to 10 closest neighbors with target 1
- Mean distance to 10 closest neighbors with target 0

- How many objects are there in a ball of radius 5, 10, ...
- Mean distance to the objects in a ball of radius 5, 10, …

- How many of closest objects have the same label
- How many different labels are there among nearest neighbors

- …
- …

Different distributions in test and train

Different distributions in test and train

- We usually assume the train data is similar to test data.
- It does not always hold true.

Data:

- y -- label
- x -- a feature with 2 levels

Bayes rule:

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)} \propto p(x|y)p(y)$$

x	**y**
a	1
b	0
a	0
a	0
b	1
...	...

How predictions are made

Bayes rule: $$p(y|x) = \frac{p(x|y)p(y)}{p(x)} \propto p(x|y)p(y)$$

How predictions are made

Bayes rule: $$p(y|x) = \frac{p(x|y)p(y)}{p(x)} \propto p(x|y)p(y)$$

$p(y)$

y	**p(y)**
0	0.7
1	0.3

How predictions are made

Bayes rule: $p(y|x) = \frac{p(x|y)p(y)}{p(x)} \propto p(x|y)p(y)$

$p(x|y)$

$y \setminus x$	**a**	**b**	
0	0.8	0.2	$\leftarrow p(x\|y=0)$
1	0.4	0.6	$\leftarrow p(x\|y=1)$

$p(y)$

y	**p(y)**
0	0.7
1	0.3

How predictions are made

Bayes rule: $p(y|x) = \frac{p(x|y)p(y)}{p(x)} \propto p(x|y)p(y)$

$p(y|x)$

y \ x	a	b
0	0.83	0.43
1	0.17	0.57
	$p(y|x=a)$	$p(y|x=b)$

$p(x|y)$

y \ x	a	b	
0	0.8	0.2	$\leftarrow p(x|y=0)$
1	0.4	0.6	$\leftarrow p(x|y=1)$

$p(y)$

y	p(y)
0	0.7
1	0.3

Classifier: $p(y=1|x=a) = \frac{0.4 \cdot 0.3}{0.8 \cdot 0.7 + 0.4 \cdot 0.3}$

What if p(y) are different in test?

Bayes rule: $p(y|x) = \frac{p(x|y)p(y)}{p(x)} \propto p(x|y)p(y)$

$p(y|x)$

$y \setminus x$	a	b
0	0.83	0.43
1	0.17	0.57
	$p(y\|x=a)$	$p(y\|x=b)$

$p(x|y)$

$y \setminus x$	a	b	
0	0.8	0.2	$\leftarrow p(x\|y=0)$
1	0.4	0.6	$\leftarrow p(x\|y=1)$

$p(y)$

y	p(y)
0	1
1	0

Classifier: $p(y=1|x=a) = 0.17 \neq \frac{0.4 \cdot 0}{0.8 \cdot 1 + 0.4 \cdot 0} = 0$

Efficiency

- Learn to implement everything efficiently.

Joblib

- Learn to implement everything efficiently.
  - **Joblib**

```python
# Simple loop
for i in range(1000):
    b[i] = a[i] ** 2


# The same, but using a closure
def f(x):
    return x ** 2

for i in range(1000):
    b[i] = f(a[i])


# Parallel version
Parallel(n_jobs=32)(delayed(f)(a[i]) for i in range(1000))
```

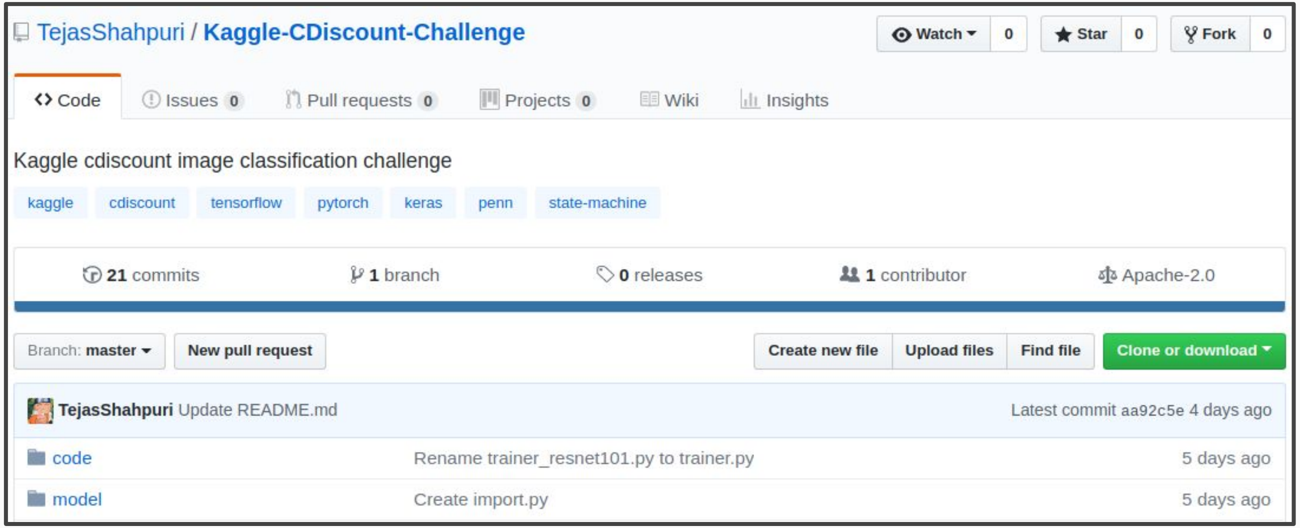
Numba

- Learn to implement everything efficiently.
  - **Numba**

```python
from numba import jit

# Pure python
def sum_python(arr):
    s = 0.0
    for i in xrange(arr.shape[0]):
        s += arr[i]
    return s

%timeit sum_python(a) # 138 ms

# Numba
sum_numba = jit(sum_python)

%timeit sum_numba(a) # 1 ms

# >100x boost
```

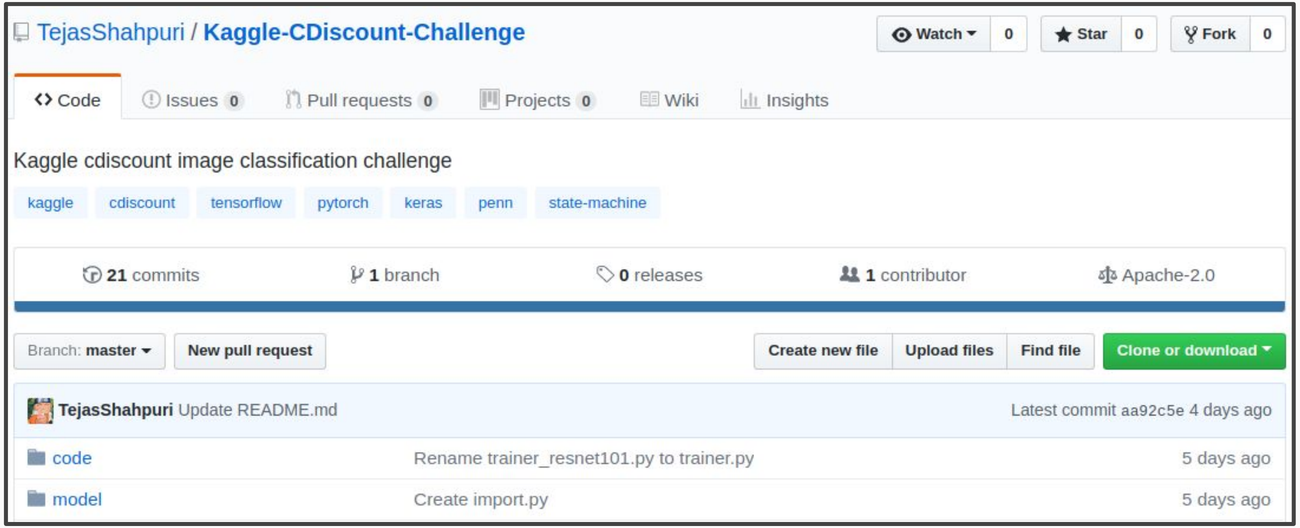
Github

- Learn to implement everything efficiently.
- Search github for solutions and inspiration



Github

- Learn to implement everything efficiently.
- Search github for solutions and inspiration



Вопрос про ГБМ

Вопрос про ГБМ

- Что будет, если из обученной GBDT модели (например XGboost) выкинуть первое дерево?

    a. Все сломается к чертям (почти рандом)
    b. Качество упадет, но не сильно
    c. Качество не изменится
    d. Качество улучшится, но не сильно
    e. Качество станет 146

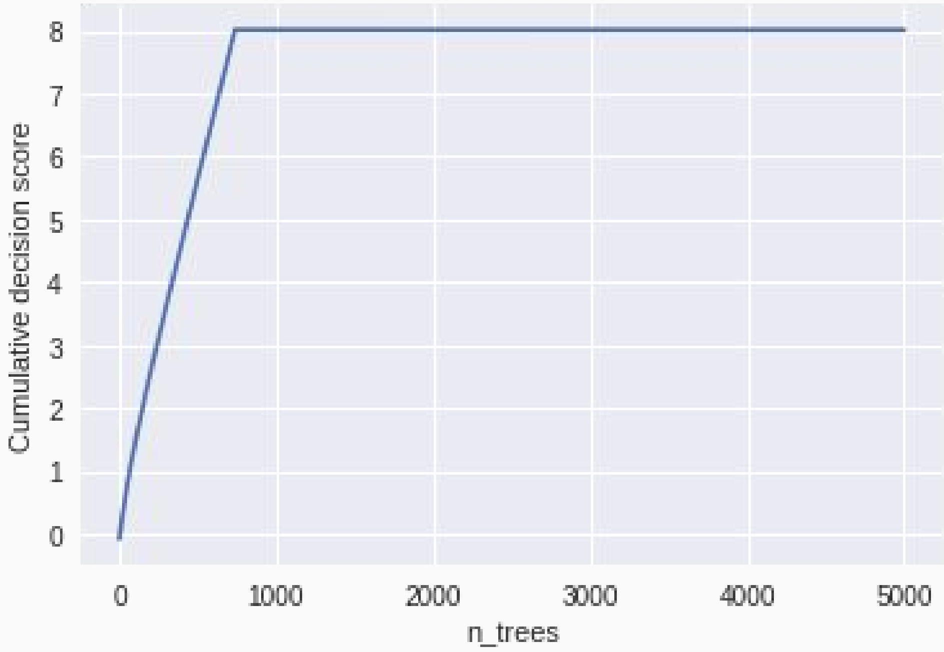
Вопрос про ГБМ

```
X_all = np.random.randn(5000, 1)
y_all = (X_all[:, 0] > 0)*2 - 1
```

```
clf = GradientBoostingClassifier(n_estimators=5000, learning_rate=0.01, max_depth=3,
clf.fit(X_train, y_train)
```

```
Logloss using all trees:           0.00031358024842548 6
Logloss using all trees but last:  0.000313580248426575 5
Logloss using all trees but first: 0.000320536825223975 3
```

Вопрос про ГБМ



Вопрос про ГБМ

```
clf = GradientBoostingClassifier(n_estimators=5000, learning_rate=8, max_depth=3,
clf.fit(X_train, y_train)
```

```
Logloss using all trees:              3.03310165292726e-06
Logloss using all trees but last:     2.846209929270204e-06
Logloss using all trees but first:    2.3463091271266125
```

Вопрос про ГБМ

$$F(x) = const + \sum_{i=1}^{n} \gamma_i h_i(x)$$

Ad time

Pavel Serdyukov

December 5 at 4:35pm

Yandex.Research (https://research.yandex.com/) ищет research interns. Это уникальная возможность присоединиться к нашей команде, заняться исследованиями мирового уровня, опубликовать результаты на ведущих конференциях (https://research.yandex.com/lib/publications) и внести вклад в наукоемкие сервисы Яндекса. Например, в технологии компьютерного зрения, диалоговые системы (Алиса), нейросетевой Перевод, алгоритмы Поиска на нейронных сетях (Королев), алгоритмы обучения на ансамблях деревьев (CatBoost), и не только.

Если вы интересуетесь глубинным обучением, алгоритмами, следите за последними статьями, хорошо знаете математику и питон, а также верите в себя как исследователя, мы будем рады пообщаться. С нашей стороны — внимание к вашим идеям, комфортный офис на Парке Культуры, продуктивная среда, дружелюбная и демократичная рабочая атмосфера. Присылайте ваши резюме и короткий рассказ о научных интересах на pavser@yandex-team.ru.

Ad time II

Home > Data Science > Data Analysis

How to Win a Data Science Competition: Learn from Top Kagglers

- Overview
- Syllabus
- FAQs
- Creators
- Ratings and Reviews

How to Win a Data Science Competition: Learn from Top Kagglers

Go to Course

Already enrolled

Financial Aid is available for learners

About this course: If you want to break into competitive data science, then this course is for you! Participating in predictive modelling competitions can help you gain practical experience, improve and harness your data modelling skills in various domains such as credit, insurance, marketing, natural language processing, sales' forecasting and computer vision to name a few. At the same time you get to do

More

Created by: National Research University Higher School of Economics

What the course is about?

- **Week1**
  - Intro to competitions & Recap
  - Feature preprocessing & extraction
- **Week2**
  - EDA
  - Validation
  - Data leaks
- **Week3**
  - Metrics
  - Mean-encodings
- **Week4**
  - Advanced features
  - Hyperparameter optimization
  - Ensembles
- **Week5**
  - Final project
  - Winning solutions

How it goes?

★★★★★ 21 Ratings

5 out of 5 stars

5 star	95 %
4 star	5 %
3 star	0 %
2 star	0 %
1 star	0 %

All reviews

3 Reviews

To flag an abusive review for removal, please contact partner-support@coursera.org

★★★★★ 10 Nov 2017

This course is fantastic. It's chock full of practical information that is presented clearly and concisely. I would like to thank the team for sharing their knowledge so generously.

Reply

The last slide

Thank you!