

Проект по дисциплине "Методы искусственного интеллекта в анализе данных"

Этап 1

Бобровских Глеб, Иванов Дмитрий, Угадяров Леонид
<https://github.com/ugadiarov-la-phystech-edu/aimda-project>

Группа 4

4 ноября 2020 г.

Набор данных и применённые модели

Рассматривается подвыборка за 2014 год из набора данных об убийствах в США [Homicide Reports, 1980-2014](#).

Задача классификации — предсказание значения бинарного признака Crime Solved.

Метрика качества — F1 мера для класса нераскрытых преступлений.

Метод опорных векторов

В качестве реализации алгоритма использовался класс `sklearn.svm.SVC` библиотеки `Scikit-learn`. Наилучшие гиперпараметры¹:

```
C = 0.1, class_weight = None, coef0 = 0, gamma = 'scale', kernel = 'linear'
```

Случайный лес

В качестве реализации алгоритма использовался класс `sklearn.ensemble.RandomForestClassifier` библиотеки `Scikit-learn`. Наилучшие гиперпараметры¹:

```
criterion = 'gini', max_depth = 10, max_features = 'auto', n_estimators = 500
```

Бустинг

В качестве реализации алгоритма использовалась класс `catboost.CatBoostClassifier` библиотеки `CatBoost`. Наилучшие гиперпараметры¹:

```
iterations = 300, depth = 6, loss_function = 'Logloss', learning_rate = 0.1, l2_leaf_reg = 4.5
```

¹Значения остальных гиперпараметров оставлены по умолчанию

Результаты экспериментов

Эксперименты проводились на платформе Google Colaboratory. Характеристики предоставляемого оборудования:

- 2 ядра процессора Intel Xeon E5-2699 v4 2.20 ГГц
- 12ГБ оперативной памяти

Метрики качества классификации обученных моделей

	F1	Precision	Recall
SVC	0.735 ± 0.007	0.703 ± 0.005	0.770 ± 0.017
RandomForest	0.737 ± 0.009	0.725 ± 0.005	0.75 ± 0.02
CatBoost	0.748 ± 0.010	0.730 ± 0.008	0.767 ± 0.016

Быстродействие обученных моделей

	Время обучения, с	$\frac{\text{Время обучения}}{\text{Количество объектов}}, \text{мс}$	Время предсказания на одном объекте, мс
SVC	16.4	1.89	1.3
RandomForest	3.4	0.39	37.2
CatBoost	4.7	0.54	1.5