
Введение в анализ текстов

Структура курса

1. Методология анализа данных
2. Блок данных
3. Блок обработки данных
4. Проектный блок

Информационные ресурсы

Сайт

<https://dstudent.ru>

Сегодня

1. Интеллектуальный анализ данных и методология исследований
2. Качество исходных данных и требования к данным
3. Обработка текстов
4. Пример задачи анализ текстов

Интеллектуальный анализ данных

Data Mining - собирательное название, которое используется для обозначения подходов и методов, позволяющих **извлекать ранее неизвестные знания из наборов данных**.

Первоначально **задача** ставится следующим образом:

- имеется достаточно крупная база данных;
- предполагается, что в базе данных находятся некие «скрытые знания».

Необходимо разработать методы обнаружения знаний, скрытых в больших объёмах исходных «сырых» данных.

Интеллектуальный анализ данных

Что означает «**скрытые знания**»:

- ранее неизвестные — то есть такие знания, которые должны быть новыми (не подтверждающими какие-то ранее полученные сведения);
- нетривиальные — то есть такие, которые нельзя просто так увидеть (при непосредственном визуальном анализе данных или при вычислении простых статистических характеристик);
- практически полезные — то есть такие знания, которые представляют ценность для исследователя или потребителя;
- доступные для интерпретации — то есть такие знания, которые легко представить в наглядной для пользователя форме и легко объяснить в терминах предметной области.

Методология анализа данных

CRISP-DM (Cross-Industry Standard Process for Data Mining) - наиболее распространенная методология ведения проектов по исследованию данных.

1. Понимание бизнеса

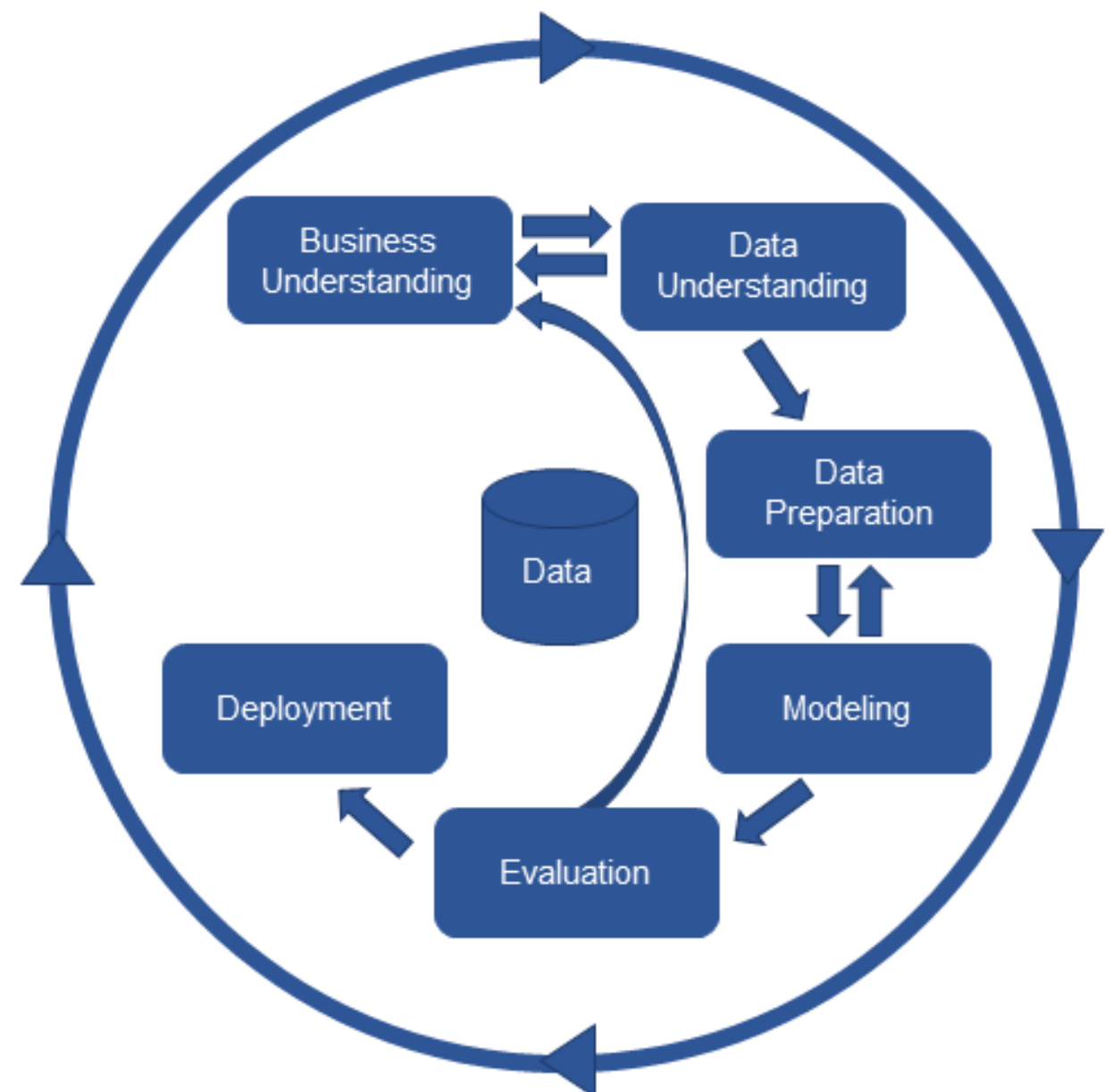
1. Определить бизнес цели
2. Оценить ситуацию
3. Определить цели анализа данных
4. Составить план проекта

2. Понимание данных

1. Собрать исходные данные
2. Описать данные
3. Исследовать данные
4. Проверить качество данных

3. Подготовка данных

1. Отобрать данные
2. Очистить данные
3. Сделать производные данные
4. Объединить данные
5. Привести данные в нужный формат



Методология анализа данных

CRISP-DM (Cross-Industry Standard Process for Data Mining) - наиболее распространенная методология ведения проектов по исследованию данных.

4. Моделирование

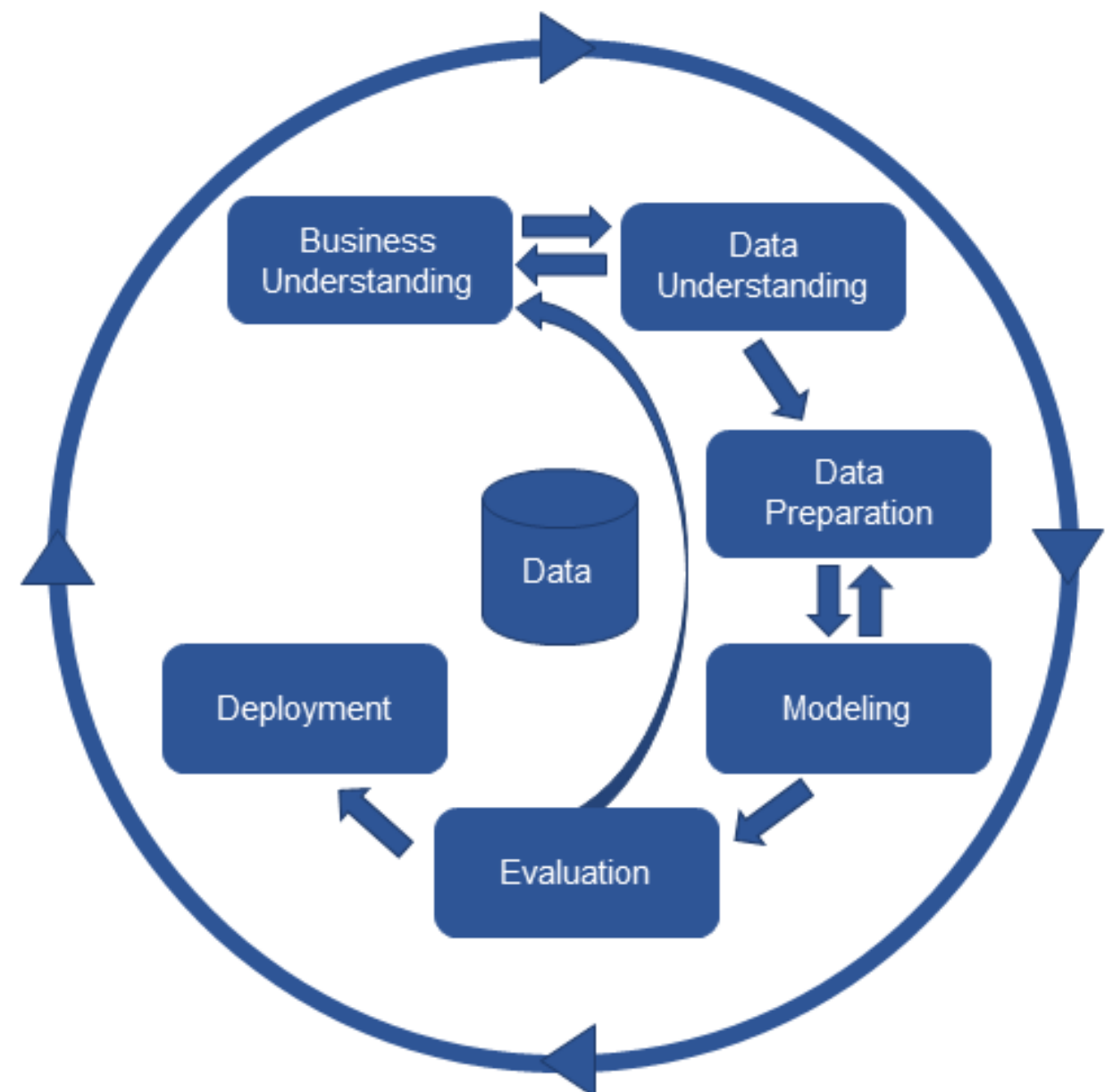
1. Выбрать методику моделирования
2. Сделать тесты для модели
3. Построить модель
4. Оценить модель

5. Оценка

1. Оценить результаты
2. Сделать ревью процесса
3. Определить следующие шаги

6. Развертывание

1. Запланировать развертывание
2. Запланировать поддержку и мониторинг развернутого решения
3. Сделать финальный отчет
4. Сделать ревью проекта



Пример задачи прогнозирования

Запрос заказчика:

Определить необходимость закупки новых серверов и месяц.

Пример задачи прогнозирования

Запрос заказчика:

Определить необходимость закупки новых серверов и месяц.

Одна из задач:

По данным о посещаемости сайта спрогнозировать максимальное количество пользователей в час.

Какие данные необходимы:

Данные о посещаемости сайта за каждый час или день.

Пример задачи прогнозирования

```
numbers = pd.read_csv('../anaconda/servers.csv')  
print(numbers)
```

	UPDATE_TIME	ZONE_CODE	HOUR_ID	BANDWIDTH_TOTAL	MAX_USER
0	2017-10-01	ZONE01	0	16096.710313	212415.0
1	2017-10-01	ZONE01	1	9374.207907	166362.0
2	2017-10-01	ZONE01	2	5606.225750	146370.0
3	2017-10-01	ZONE01	3	4155.654661	141270.0
4	2017-10-01	ZONE01	4	3253.978594	139689.0
5	2017-10-01	ZONE01	5	4310.774522	165342.0
6	2017-10-01	ZONE01	4	10015.599874	
7	2017-10-01	ZONE01	7	19780.995631	302736.0
8	2017-10-01	ZONE01	8	24344.875025	336600.0
9	2017-10-01	ZONE01	9	28651.427689	356235.0
10	2017-10-01	ZONE01	10	31983.899965	361131.0
11	2017-10-01	ZONE01	11	34044.849593	360774.0
12	2017-10-01	ZONE01	12	37520.328982	355980.0
13	2017-10-01	ZONE01	13	34521.039789	344811.0
14	2017-10-01	ZONE01	14	34402.177415	343281.0
15	2017-10-01	ZONE01	15	31897.525353	348381.0
16	2017-10-01	ZONE01	16	29310.921725	355776.0
17	2017-10-01	ZONE01	17	29826.141779	
18	2017-10-01	ZONE01	18	36244.746285	407286.0
19	2017-10-01	ZONE01	19	32493.870468	370464.0
20	2017-10-01	ZONE01	20	37388.830453	375768.0
21	2017-10-01	ZONE01	21	38976.393114	378981.0
22	2017-10-01	ZONE01	22	32728.948859	348381.0
23	2017-10-01	ZONE01	23	22768.778896	254133.0
24	2017-10-02	ZONE01	0	11370.931635	182988.0
25	2017-10-02	ZONE01	1	6500.976282	147594.0

Проблемы с качеством данных

- неполнота — данные не содержат атрибутов или в них пропущены значения;
- шум — данные содержат ошибочные записи или выбросы;
- несогласованность — данные содержат конфликтующие между собой записи или расхождения.

Качественные данные - необходимое условие для создания качественных моделей прогнозирования.

Требования к данным

Четыре основных вопроса, возникающих на этапе сбора данных:

- какие входные данные следует включить?
- как получить известные значения переменной, которую необходимо предсказать?
- сколько данных требуется?
- как оценить качество данных?

Пример задачи прогнозирования

```
numbers = pd.read_csv('../anaconda/servers.csv')  
print(numbers)
```

	UPDATE_TIME	ZONE_CODE	HOUR_ID	BANDWIDTH_TOTAL	MAX_USER
0	2017-10-01	ZONE01	0	16096.710313	212415.0
1	2017-10-01	ZONE01	1	9374.207907	166362.0
2	2017-10-01	ZONE01	2	5606.225750	146370.0
3	2017-10-01	ZONE01	3	4155.654661	141270.0
4	2017-10-01	ZONE01	4	3253.978594	139689.0
5	2017-10-01	ZONE01	5	4310.774522	165342.0
6	2017-10-01	ZONE01	6	10015.599874	221901.0
7	2017-10-01	ZONE01	7	19780.995631	302736.0
8	2017-10-01	ZONE01	8	24344.875025	336600.0
9	2017-10-01	ZONE01	9	28651.427689	356235.0
10	2017-10-01	ZONE01	10	31983.899965	361131.0
11	2017-10-01	ZONE01	11	34044.849593	360774.0
12	2017-10-01	ZONE01	12	37520.328982	355980.0
13	2017-10-01	ZONE01	13	34521.039789	344811.0
14	2017-10-01	ZONE01	14	34402.177415	343281.0
15	2017-10-01	ZONE01	15	31897.525353	348381.0
16	2017-10-01	ZONE01	16	29310.921725	355776.0
17	2017-10-01	ZONE01	17	29826.141779	357918.0
18	2017-10-01	ZONE01	18	36244.746285	407286.0
19	2017-10-01	ZONE01	19	32493.870468	370464.0
20	2017-10-01	ZONE01	20	37388.830453	375768.0
21	2017-10-01	ZONE01	21	38976.393114	378981.0
22	2017-10-01	ZONE01	22	32728.948859	348381.0
23	2017-10-01	ZONE01	23	22768.778896	254133.0
24	2017-10-02	ZONE01	0	11370.931635	182988.0
25	2017-10-02	ZONE01	1	6500.976282	147594.0

Пример задачи прогнозирования

Группируем значения по дням и берем максимальное значение посещаемости в день.

```
grouped = numbers.groupby([ 'UPDATE_TIME' ]).max()  
print(grouped)
```

	ZONE_CODE	HOUR_ID	BANDWIDTH_TOTAL	MAX_USER
UPDATE_TIME				
2017-10-01	ZONE01	23	38976.393114	407286.0
2017-10-02	ZONE01	23	35790.289155	381429.0
2017-10-03	ZONE01	23	34867.323414	388518.0
2017-10-04	ZONE01	23	36784.855100	379440.0
2017-10-05	ZONE01	23	34593.017022	378318.0
2017-10-06	ZONE01	23	34110.527590	374850.0
2017-10-07	ZONE01	23	35958.015584	369240.0
2017-10-08	ZONE01	23	35530.428097	375003.0
2017-10-09	ZONE01	23	36342.441228	379440.0
2017-10-10	ZONE01	23	34945.040221	376788.0
2017-10-11	ZONE01	23	37364.970083	390558.0
2017-10-12	ZONE01	23	38821.141330	394944.0
2017-10-13	ZONE01	23	38987.724287	391272.0
2017-10-14	ZONE01	23	37655.994665	396984.0
2017-10-15	ZONE01	23	39298.685510	400860.0
2017-10-16	ZONE01	23	38234.038383	395046.0
2017-10-17	ZONE01	23	38730.646401	392853.0
2017-10-18	ZONE01	23	37486.729469	403461.0
2017-10-19	ZONE01	23	35429.819635	401982.0
2017-10-20	ZONE01	23	39522.477218	417282.0
2017-10-21	ZONE01	23	42692.525140	400503.0

Пример задачи прогнозирования

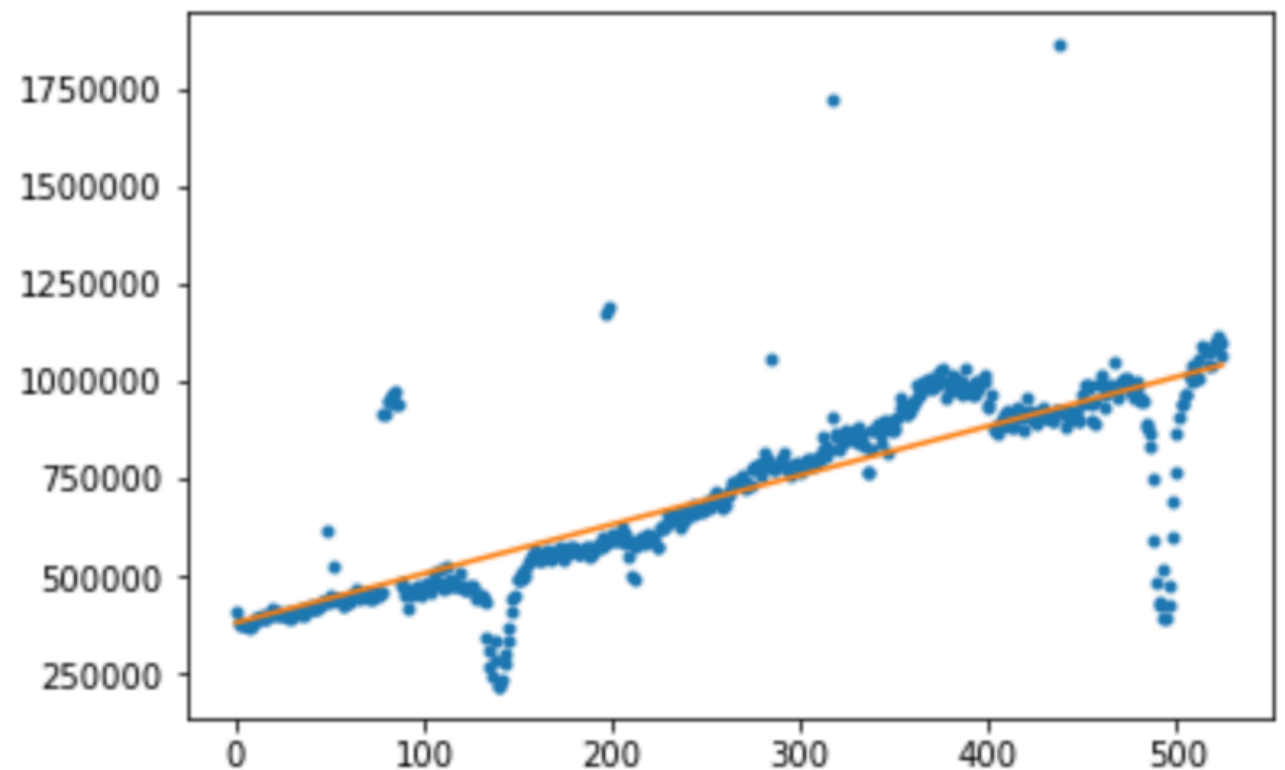
Найдем тренд по значениям посещаемости сайта в день

```
x = np.arange(0, len(grouped.MAX_USER))
y = np.array(grouped.MAX_USER)
z = np.polyfit(x, y, 1)
print("{0}x + {1}".format(*z))

from matplotlib import pyplot as plt

plt.plot(x, y, '. ')
trendpoly = np.poly1d(z)
plt.plot(x, trendpoly(x))
plt.show()
```

1258.7273731087068x + 381743.1825312329



Типы исходных данных

Структурированные данные

	UPDATE_TIME	ZONE_CODE	HOUR_ID	BANDWIDTH_TOTAL	MAX_USER
0	2017-10-01	ZONE01	0	16096.710313	212415.0
1	2017-10-01	ZONE01	1	9374.207907	166362.0
2	2017-10-01	ZONE01	2	5606.225750	146370.0
3	2017-10-01	ZONE01	3	4155.654661	141270.0
4	2017-10-01	ZONE01	4	3253.978594	139689.0
5	2017-10-01	ZONE01	5	4310.774522	165342.0
6	2017-10-01	ZONE01	6	10015.599874	221901.0
7	2017-10-01	ZONE01	7	19780.995631	302736.0
8	2017-10-01	ZONE01	8	24344.875025	336600.0
9	2017-10-01	ZONE01	9	28651.427689	356235.0
10	2017-10-01	ZONE01	10	31983.899965	361131.0
11	2017-10-01	ZONE01	11	34044.849593	360774.0
12	2017-10-01	ZONE01	12	37520.328982	355980.0
13	2017-10-01	ZONE01	13	34521.039789	344811.0
14	2017-10-01	ZONE01	14	34402.177415	343281.0
15	2017-10-01	ZONE01	15	31897.525353	348381.0
16	2017-10-01	ZONE01	16	29310.921725	355776.0
17	2017-10-01	ZONE01	17	29826.141779	357918.0
18	2017-10-01	ZONE01	18	36244.746285	407286.0

Неструктурированные данные - информация, не имеющая определенную структуру, или не организована в формате, который требуется для решения задачи.

Типы исходных данных



Неструктурированные



Изображения

Видео



Типы исходных данных

Efficient and Robust Automated Machine Learning

Matthias Feurer
Jost Tobias Springenberg

Aaron Klein
Manuel Blum
Department of Computer Science
University of Freiburg, Germany

Katharina Eggensperger
Frank Hutter

{feurerm, kleinaa, eggensp, springj, mblum, fh}@cs.uni-freiburg.de

Abstract

The success of machine learning in a broad range of applications has led to an ever-growing demand for machine learning systems that can be used off the shelf by non-experts. To be effective in practice, such systems need to automatically choose a good algorithm and feature preprocessing steps for a new dataset at hand, and also set their respective hyperparameters. Recent work has started to tackle this *automated machine learning (AutoML)* problem with the help of efficient Bayesian optimization methods. Building on this, we introduce a robust new AutoML system based on scikit-learn (using 15 classifiers, 14 feature preprocessing methods, and 4 data preprocessing methods, giving rise to a structured hypothesis space with 110 hyperparameters). This system, which we dub AUTO-SKLEARN, improves on existing AutoML methods by automatically taking into account past performance on similar datasets, and by constructing ensembles from the models evaluated during the optimization. Our system won the first phase of the ongoing ChaLearn AutoML challenge, and our comprehensive analysis on over 100 diverse datasets shows that it substantially outperforms the previous state of the art in AutoML. We also demonstrate the performance gains due to each of our contributions and derive insights into the effectiveness of the individual components of AUTO-SKLEARN.

1 Introduction

Machine learning has recently made great strides in many application areas, fueling a growing demand for machine learning systems that can be used effectively by novices in machine learning. Correspondingly, a growing number of commercial enterprises aim to satisfy this demand (e.g., BigML.com, Wise.io, SkyTree.com, RapidMiner.com, Dato.com, Prediction.io, DataRobot.com, Microsoft's Azure Machine Learning, Google's Prediction API, and Amazon Machine Learning). At its core, every effective machine learning service needs to solve the fundamental problems of deciding which machine learning algorithm to use on a given dataset, whether and how to preprocess its features, and how to set all hyperparameters. This is the problem we address in this work.

More specifically, we investigate automated machine learning (AutoML), the problem of automatically (without human input) producing test set predictions for a new dataset within a fixed computational budget. Formally, this AutoML problem can be stated as follows:

Тексты на естественном языке



ИМиКН. ML&BigData запись закреплена

15 апр 2018

Клуб занимается развитием направлений науки Машинное обучение и Анализ больших данных среди студентов ИМиКН ТюмГУ. Участники Клуба при помощи организаторов осваивают компетенции профессии Data Scientist.

Фактически технология машинного обучения стала одним из самых востребованных инструментов бизнеса. Сейчас уже трудно найти крупную компанию-лидера в своем направлении, которая не заявила бы об адаптации методов машинного обучения к своим бизнес-процессам

Занятия проводятся дополнительно вне учебных занятий.

Каждый желающий может приходить на занятия Клуба, получать новые знания и решать вместе с нами задачи.

Ответы на организационные вопросы есть в Информации группы.

О том, что такое Машинное обучение, Big Data, кто такой Data Scientist, интересные факты из мира Data Science регулярно появляется информация в группе.



Обработка текстов

Обработка текстов на естественном языке (Natural Language Processing) - подходы и методы для обработки текстов на естественном языке.

Примеры задач, которые решаются с помощью NLP:

- Синтаксический разбор, токенизация предложений (предложение);
- Извлечение отношений, определение языка, анализ эмоциональной окраски (абзац);
- Аннотация документа, анализ тематики (документ);
- Информационный поиск (корпус).

Разделение на предложения

NLTK (Natural Language ToolKit) - платформа для разработки программ на языке Python для обработки текстов на естественном языке.

```
#Разделение текста на предложения  
from nltk.tokenize import sent_tokenize
```

```
text = 'Первое предложение. Второе предложение, уже с запятой. Восклицательное предложение! Вопросительное предложение'  
sentences = sent_tokenize(text)  
print(sentences)
```

```
['Первое предложение.', 'Второе предложение, уже с запятой.', 'Восклицательное предложение!', 'Вопросительное предложение']
```

Разделение на слова

Под словом понимается в том числе и знак препинания

```
from nltk import word_tokenize
```

```
text = "В этом предложении есть много слов, мы их разделим."  
print(word_tokenize(text))
```

```
['В', 'этом', 'предложении', 'есть', 'много', 'слов', ',', 'мы', 'их', 'разделим', '.']
```

Знаки препинания исключаются

```
from nltk.tokenize import RegexpTokenizer
```

```
text = "В этом предложении есть много слов, мы их разделим."  
tokenizer = RegexpTokenizer(r'\w+')  
print(tokenizer.tokenize(text))
```

```
['В', 'этом', 'предложении', 'есть', 'много', 'слов', 'мы', 'их', 'разделим']
```


Удаление стоп-слов

```
from nltk.corpus import stopwords
```

```
stop_words=set(stopwords.words('english'))  
print(stop_words)
```

```
{'theirs', 'hadn', 'when', 'there', 'mightn', 'isn', 'further', 'shan', 'hasn't', 'mightn't', 'out', 'few', 's', 'her  
e', 'off', 'does', 'you've', 'where', 'won', 'him', 'doesn't', 'can', 'she's', 'herself', 'yours', 'in', 'isn't', 'th  
en', 'don't', 'that', 'did', 'down', 'wouldn', 'don', 'before', 'wouldn't', 'didn', 'over', 'your', 'it's', 'against'  
, 'haven', 'being', 'again', 'been', 'needn', 'my', 'me', 'ma', 'if', 'than', 'this', 'ours', 'only', 'shouldn't', 'a  
in', 'it', 'you', 'such', 'how', 'should've', 'didn't', 'had', 'any', 'between', 'once', 'weren't', 'hasn', 'having',  
'own', 'no', 'whom', 'about', 'couldn', 'am', 'who', 'that'll', 'its', 'most', 'our', 'yourself', 'of', 'do', 'becaus  
e', 'hadn't', 'which', 'above', 'you're', 'them', 'as', 'by', 'until', 'too', 'mustn't', 'their', 'we', 'at', 'was',  
'should', 'o', 'what', 'her', 'the', 'wasn', 'yourselves', 'now', 'doing', 'is', 'he', 'below', 'you'd', 'be', 'all',  
'and', 'm', 'just', 'are', 'some', 'himself', 'on', 'shouldn', 'aren', 'hers', 're', 'aren't', 'those', 'weren', 'a',  
'while', 'after', 'up', 'will', 'for', 'won't', 'an', 't', 'or', 'haven't', 'from', 'have', 've', 'each', 'other', 'm  
ustn', 'll', 'his', 'into', 'why', 'wasn't', 'shan't', 'themselves', 'very', 'needn't', 'has', 'to', 'ourselves', 'mo  
re', 'i', 'same', 'couldn't', 'nor', 'with', 'not', 'but', 'they', 'you'll', 'doesn', 'were', 'during', 'so', 'itself'  
, 'd', 'both', 'y', 'myself', 'these', 'under', 'through', 'she'}
```

```
stop_words=set(stopwords.words('russian'))  
print(stop_words)
```

```
{'чтоб', 'до', 'конечно', 'бы', 'но', 'через', 'совсем', 'от', 'если', 'теперь', 'вот', 'еще', 'всего', 'нибудь', 'на  
конец', 'потому', 'вдруг', 'тот', 'во', 'для', 'здесь', 'потом', 'чтобы', 'только', 'под', 'того', 'этого', 'какая',  
'за', 'ни', 'со', 'она', 'этот', 'у', 'ли', 'в', 'кто', 'то', 'ей', 'к', 'том', 'раз', 'сейчас', 'такой', 'они', 'не'  
, 'да', 'хорошо', 'было', 'все', 'я', 'что', 'эту', 'более', 'тебя', 'даже', 'вы', 'их', 'моя', 'есть', 'были', 'или'  
, 'перед', 'вам', 'чего', 'тем', 'же', 'им', 'как', 'нас', 'так', 'про', 'уже', 'всех', 'всегда', 'уж', 'нельзя', 'ту  
т', 'этом', 'мне', 'ему', 'ничего', 'хоть', 'себя', 'при', 'иногда', 'ведь', 'будто', 'после', 'мой', 'будет', 'когда'  
, 'этой', 'свою', 'была', 'по', 'них', 'без', 'ним', 'ней', 'мы', 'нет', 'можно', 'никогда', 'меня', 'может', 'его',  
'себе', 'больше', 'опять', 'надо', 'сам', 'разве', 'между', 'быть', 'ж', 'а', 'с', 'один', 'зачем', 'тогда', 'впрочем'  
, 'там', 'два', 'ты', 'него', 'почти', 'лучше', 'из', 'всю', 'эти', 'ее', 'и', 'об', 'был', 'ну', 'куда', 'какой', '  
много', 'над', 'нее', 'на', 'чуть', 'о', 'три', 'где', 'вас', 'он', 'чем', 'другой', 'тоже'}
```

Поиск похожих текстов

Дано: набор (корпус) текстов.

Найти: для заданного текстового сообщения (текста) найти ближайший текст из исходного набора.

	title	content	tags
id			
1	What are some Caribbean cruises for October?	<p>My fiancée and I are looking for a good Car...	caribbean cruising vacations
2	How can I find a guide that will take me safel...	<p>This was one of our definition questions, b...	guides extreme-tourism amazon-river amazon-jungle
4	Does Singapore Airlines offer any reward seats...	<p>Singapore Airlines has an all-business clas...	loyalty-programs routes ewr singapore-airlines...
5	What is the easiest transportation to use thro...	<p>Another definition question that interested...	romania transportation
6	How can I visit Antarctica?	<p>A year ago I was reading some magazine, and...	extreme-tourism antarctica
8	Best way to get from SeaTac airport to Redmond?	<p>Can anyone suggest the best way to get from...	usa airport-transfer taxis seattle

Сообщение: «safely travel tour Amazon jungle»

Поиск похожих текстов

	title	content	tags
id			
1	What are some Caribbean cruises for October?	<p>My fiancée and I are looking for a good Car...	caribbean cruising vacations
2	How can I find a guide that will take me safel...	<p>This was one of our definition questions, b...	guides extreme-tourism amazon-river amazon-jungle
4	Does Singapore Airlines offer any reward seats...	<p>Singapore Airlines has an all-business clas...	loyalty-programs routes ewr singapore-airlines...
5	What is the easiest transportation to use thro...	<p>Another definition question that interested...	romania transportation
6	How can I visit Antarctica?	<p>A year ago I was reading some magazine, and...	extreme-tourism antarctica
8	Best way to get from SeaTac airport to Redmond?	<p>Can anyone suggest the best way to get from...	usa airport-transfer taxis seattle

	title	content	tags
id			
1	How can I get chewy chocolate chip cookies?	<p>My chocolate chips cookies are always too c...	baking cookies texture
2	How should I cook bacon in an oven?	<p>I've heard of people cooking bacon in an ov...	oven cooking-time bacon
3	What is the difference between white and brown...	<p>I always use brown extra large eggs, but I ...	eggs
4	What is the difference between baking soda and...	<p>And can I use one in place of the other in ...	substitutions please-remove-this-tag baking-so...
5	In a tomato sauce recipe, how can I cut the ac...	<p>It seems that every time I make a tomato sa...	sauce pasta tomatoes italian-cuisine
6	What ingredients (available in specific region...	<p>I have a recipe that calls for fresh parsle...	substitutions herbs parsley

Поиск похожих текстов

Примеры задач поиска похожих текстов:

- ПОИСКОВЫЕ СИСТЕМЫ;
- анализ постов в социальных сетях;
- поиск похожих статей.

Задание 1

Совместно в группах выбрать и описать 3-5 гипотез и сформулировать, где еще могут использоваться задачи поиска похожих текстов. Для каждой задачи:

- описать, кто и в каких ситуациях сталкивается с данной задачей (это может быть как одна из подзадач целого комплекса);
- описать данные (тексты), которые необходимы для ее решения;
- описать, какие подзадачи предположительно будут включены в обработку текстов для решения задачи;
- найти необходимые данные.