ML-Тренировки
Вводное занятие

Интро

Цель тренировок:
- подготовка к решению задач из разных доменов
- обмен опытом, разбор базовых подходов

Программа тренировок:
- вводное занятие
- tabular data x 2
- natural language processing x 2
- computer vision x 2
- разбор новых соревнований

Evaluation metric:
- активность
- обмен знаниями
- участие в соревнованиях
- публикация решений

Мир ML

cpmpml

CPMP

RAPIDS and deep learning at NVIDIA

France

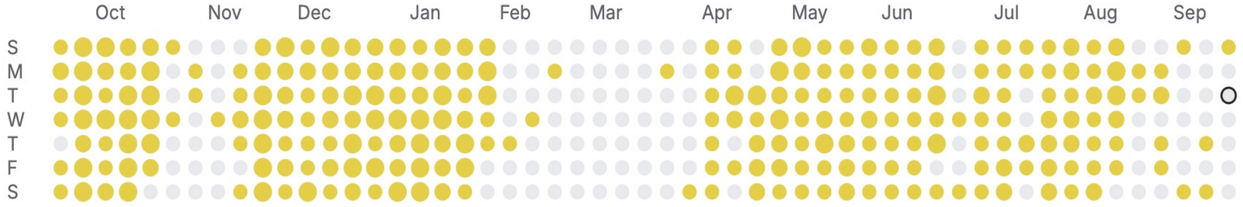
Joined 12 years ago · last seen in the past day

Competitions Grandmaster
22 of 205,792

Bio

Got a PhD in machine learning in a previous millennium, ML was very different from now, and therefore this PhD is useless... Worked on constraint programming and mathematical optimization since then until few years ago where I came back to ML. Entered my first competition on Kaggle in May 2016, and never stopped competing since then. Now at NVIDIA.

Public activity



https://www.kaggle.com/cpmpml

Мир ML

abhishek

Abhishek Thakur

AutoTrain at Hugging Face

Oslo, Oslo, Norway

Joined 14 years ago · last seen in the past day

Competitions Grandmaster
690 of 205,792

Approaching (Almost) Any NLP Problem on Kaggle

In this post I'll talk about approaching natural language processing problems on Kaggle. As an example, we will use the data from this competition. We will create a very basic first model first and then improve it using different other features. We will also see how deep neural networks can be used and end this post with some ideas about ensembling in general.

This covers:

- tfidf
- count features
- logistic regression
- naive bayes
- svm
- xgboost
- grid search
- word vectors
- LSTM
- GRU
- Ensembling

https://www.kaggle.com/code/abhishek/approaching-almost-any-nlp-problem-on-kaggle

Мир ML

abhishek

Abhishek Thakur

AutoTrain at Hugging Face

Oslo, Oslo, Norway

Joined 14 years ago · last seen in the past day

Competitions Grandmaster
690 of 205,792

Mixtral 8×7b trained on math dataset

I recently trained a mixtral8×7b on 40k math-specific dataset and it seems to perform quite well on the Open LLM LB for GSM8K benchmark.
Hopefully, someone here can use it for this competition: https://huggingface.co/abhishek/autotrain-mixtral7×8b-math

Let me know how it goes :)

Related X/Twitter post: https://twitter.com/abhi1thakur/status/1776589253847257280
Model was trained using AutoTrain: https://twitter.com/abhi1thakur/status/1776589253847257280

https://www.kaggle.com/competitions/ai-mathematical-olympiad-prize/discussion/492013

Мир ML

Выбор домена:
- tabular data
- natural language processing
- computer vision
- ..

Выбор класса задач:
- регрессия
- временные ряды
- сегментация снимков
- генерация текста
- ..

Выбор стека:
- ..

Мир ML

Соревнования (NLP):
- regression
(https://www.kaggle.com/competitions/feedback-prize-english-language-learning)
- classification
(https://www.kaggle.com/competitions/feedback-prize-effectiveness)
- ranking
(https://www.kaggle.com/competitions/jigsaw-toxic-severity-rating)
- matching
(https://www.kaggle.com/competitions/us-patent-phrase-to-phrase-matching)
- RecSys
(https://www.kaggle.com/competitions/otto-recommender-system)
- NER
(https://www.kaggle.com/competitions/coleridgeinitiative-show-us-the-data)
- prompt recovery
(https://www.kaggle.com/competitions/llm-prompt-recovery)
- question answering
(https://www.kaggle.com/competitions/ai-mathematical-olympiad-prize)
- reinforcement learning
(https://www.kaggle.com/competitions/llm-20-questions)

Мир ML

ybabakhin

Yauhen Babakhin

Principal Data Scientist at H2O.ai

Prague, Prague, Czechia

Joined 10 years ago · last seen in the past day

Competitions Grandmaster
24 of 205,792

1st Place Solution with Code

UPD V2: Paper with our solution approach is published at German Conference on Pattern Recognition (GCPR), 2019:

https://arxiv.org/abs/1904.04445

UPD: We've made our code available on github:

https://github.com/ybabakhin/kaggle_salt_bes_phalanx

First of all, I'd like to congratulate and thank my teammate phalanx for his great contribution and effort!
Also, thanks to organizers for this competition and to Heng and Peter for their insightful forum posts.

It is my first problem in image segmentation, just 3 months ago I knew nothing about segmentation. So, this 1st place is a tremendous bonus for all the knowledge and experience we've gained. I guess, it's a good example for novices: if you work hard, you could achieve high results even with little background.

https://www.kaggle.com/competitions/tgs-salt-identification-challenge/discussion/69291

Мир ML


```
test["pred_answer"] = "Rephrase paragraph " +
test["answers_fs"] + " " + test["pred_answer_llm4"] + " " + test["pred_answer_llm1"] +
'.lucrarealucrarealucrarea sentence appealinglucrarea
Improve respond storytelling tonelucrareaimplication. write someoneran. lucrarea].
Consider clarify paragraphlucrarea similarly serious themed way temporarily.! ElePT'
```

https://www.kaggle.com/competitions/llm-prompt-recovery/discussion/494343
https://www.kaggle.com/code/ilu000/2nd-place-team-danube-llm-prompt-recovery

Знакомство с Kaggle

Важные шаги:
- прочитать описание проблемы
- разобраться, как сделать submission
- привыкнуть к OutOfMemory
- ..

- месяц бороться за попадание в топ50%
- открыть секцию с high-scoring kernels
- посмотреть топ1 решение
- посмотреть решения
- кросс-валидация
- ..

- принять участие в code competition
- привыкнуть к OutOfMemory
- ..

- начать считаться в облаке
- привыкнуть к OutOfMemory
- ..

Знакомство с Kaggle

tvdwiele

Tom Van de Wiele

Senior Quantitative Researcher at G-Research

Gran, Innlandet, Norway

Joined 8 years ago · last seen 4 months ago

Competitions Grandmaster
903 of 205,792

ttvand Undo new tab hyperlink 0bb7165 · 5 years ago 20 Commits

Candidate selection	Update getTopKNNDT.R	6 years ago
Common	Cleaned Facebook V code	8 years ago
Data	Cleaned Facebook V code	8 years ago
Downsampling	Cleaned Facebook V code	8 years ago
Evaluate predictions	Cleaned Facebook V code	8 years ago
Exploratory analysis	Removed irrelevant exploratory plots	8 years ago
Feature engineering	Cleaned Facebook V code	8 years ago
First level learners	Cleaned Facebook V code	8 years ago
References	Cleaned Facebook V code	8 years ago
Second level learners	Cleaned Facebook V code	8 years ago
Strategy	Cleaned Facebook V code	8 years ago
Submission	Removed outdated function from getTopKNNDT	8 years ago

https://github.com/ttvand/Facebook-V

Знакомство с Kaggle

lopuhin

Konstantin Lopuhin

ML Engineer at Zyte

Tbilisi, Tbilisi, Georgia

Joined 13 years ago · last seen 2 months ago

Competitions Grandmaster
4,585 of 205,792

```python
def main():
    vectorizer = make_union(
        on_field('name', Tfidf(max_features=100000, token_pattern='\w+')),
        on_field('text', Tfidf(max_features=100000, token_pattern='\w+', ngram_range=(1, 2))),
        on_field(['shipping', 'item_condition_id'],
                 FunctionTransformer(to_records, validate=False), DictVectorizer()),
        n_jobs=4)
    y_scaler = StandardScaler()
    with timer('process train'):
        train = pd.read_table('../input/train.tsv')
        train = train[train['price'] > 0].reset_index(drop=True)
        cv = KFold(n_splits=20, shuffle=True, random_state=42)
        train_ids, valid_ids = next(cv.split(train))
        train, valid = train.iloc[train_ids], train.iloc[valid_ids]
        y_train = y_scaler.fit_transform(np.log1p(train['price'].values.reshape(-1, 1)))
        X_train = vectorizer.fit_transform(preprocess(train)).astype(np.float32)
        print(f'X_train: {X_train.shape} of {X_train.dtype}')
        del train
    with timer('process valid'):
        X_valid = vectorizer.transform(preprocess(valid)).astype(np.float32)
    with ThreadPool(processes=4) as pool:
        Xb_train, Xb_valid = [x.astype(np.bool).astype(np.float32) for x in [X_train, X_valid]]
        xs = [[Xb_train, Xb_valid], [X_train, X_valid]] * 2
        y_pred = np.mean(pool.map(partial(fit_predict, y_train=y_train), xs), axis=0)
    y_pred = np.expm1(y_scaler.inverse_transform(y_pred.reshape(-1, 1))[:, 0])
    print('Valid RMSLE: {:.4f}'.format(np.sqrt(mean_squared_log_error(valid['price'], y_pred))))

if __name__ == '__main__':
    main()
```

https://www.kaggle.com/code/lopuhin/mercari-golf-0-3875-cv-in-75-loc-1900-s

Знакомство с Kaggle

cdeotte

Chris Deotte

Data Scientist & Researcher at NVIDIA

San Diego, California, United States

Joined 7 years ago · last seen in the past day

Competitions Grandmaster
15 of 205,780

Winning Solution in 7 lines of code!

Hindsight is 20/20. Here's 7 lines of codes that scores 0.681 Private LB and wins first place!! It doesn't even use the training data.

```
import pandas as pd, numpy as np
submit = pd.read_csv('test.csv', usecols=['MachineIdentifier','AvSigVersion'])
submit['HasDetections'] = 0.5
submit['ASV2'] = submit['AvSigVersion'].map(lambda x: np.int(x.split('.')[1]) )
submit['ASV3'] = submit['AvSigVersion'].map(lambda x: np.int(x.split('.')[2]) )
submit.loc[ (submit['ASV2']==281) &amp; (submit['ASV3'] >= 451),'HasDetections'] = 0.0
submit[['MachineIdentifier','HasDetections']].to_csv('WinningSolution.csv', index=False)
```

https://www.kaggle.com/competitions/microsoft-malware-prediction/discussion/84096

Развитие решения

Этапы соревнования:
- период до публикации решения из зоны медалей (.1 -> .25 -> ..)
- стадия тюнинга гиперпараметров (.6789 -> .679 -> .679 -> ..)
- неделя ансамблей (штурм сектора медалей)

Базовые советы:
- исследование метрики (https://www.kaggle.com/competitions/statoil-iceberg-classifier-challenge)
- проверка корреляции CV и LB (https://www.kaggle.com/competitions/porto-seguro-safe-driver-prediction)
- дизайн экспериментов (https://www.kaggle.com/johnpateha)
- формулирование гипотез (https://www.kaggle.com/sggpls)
- переиспользование наработок (новые вводные)
- диверсификация решений (новое направление)
- работа в команде

Развитие решения

Pipeline stages:
- ..

- data collection
- data split
- features engineering
- features transformation
- features treatment
- features selection
- model trainer
- ..

Развитие решения

Microsoft Malware Prediction

Overview Data Code Models Discussion Leaderboard Rules Team Submissions

Prize Winners

#	△	Team	Members	Score
1	▲ 1209	abuurista		0.67585
2	▲ 1064	Confiniti		0.66535
3	▲ 1082	ken10ML		0.66523
4	▲ 1353	John DiMarco		0.66474
5	▲ 1524	khas_ccip		0.66403
6	▲ 11	ThunderBYTE		0.66393
7	▲ 1268	SanderF		0.66370
8	▲ 1684	Jose		0.66340

https://www.kaggle.com/competitions/microsoft-malware-prediction/leaderboard

Развитие решения

PAVEL PLESKOV · 13TH IN THIS COMPETITION · POSTED 7 YEARS AGO

Try to improve your solution with the leak

As I mentioned earlier the was a leak in the data: points with the same angle almost surely had the same label for both the test and the train sets.

There are several ways how to exploit it:

1. When angle is in the train set and train label does not contradict your model (rounded probability is the same as label) then assign maximum probability to this point. This probability should be equal to your clip value, of course. Never set 0 or 1 when using logloss metric since every model can be wrong and punishment for the wrong label is huge (as many participants have noticed already).
2. When the angle is not in the train set but all the data points with the same angle do not contradict to each other (have the same label) then again assign the maximum probability to all of them. This step improved public score but not private score so can be omitted.
3. When the angle is in the train set but your model is contradicting to the train label then use majority vote for all point with the same angle.

So I did not train any models, just took the best public kernel from @golubev and improved it.

Try to improve your own solution as well! Change df['is_iceberg'] in the beginning of this kernel to your probabilities and post the results: https://www.kaggle.com/ppleskov/leaky-solution-14th-place-lb-0-1038-0-0960

PS: by the way, there is a great online course about data leaks where the same technique was described - How to Win a Data Science Competition https://www.coursera.org/learn/competitive-data-science

https://www.kaggle.com/competitions/statoil-iceberg-classifier-challenge/discussion/48224

Развитие решения

TGS Salt Identification Challenge

Late Submission

Overview | Data | Code | Models | Discussion | Leaderboard | Rules | Team | Submissions

34	▼ 18	Dmitriy, Terence & Takato	0.88732	363	6y
35	▲ 14	soywu	0.88710	135	6y
36	▼ 8	DavidGbodiOdaibo	0.88707	120	6y
37	—	⌘	0.88702	14	6y
38	▲ 6	Igor Iwaskiv	0.88667	69	6y
39	▲ 9	**Oneiros**	0.88656	402	6y

You won a silver medal!
Your team placed 39th out of 3219 teams.

View Certificate

https://www.kaggle.com/competitions/tgs-salt-identification-challenge/leaderboard

Развитие решения

В результате 98 / 100 гипотез не принесут результата.
Важно продолжать, корректироваться.
Сохранять положительный настрой.

Открытые вопросы

Почему GMs всегда догоняют лидеров таблицы?
Скорость сборки стартера?