

Machine Learning Tutorial - Kaggle

Seil Na, Youngjae Yu

Sep 6, 2017

<https://seilna.github.io/ml-tutorial>



SEOUL NATIONAL UNIV.
VISION & LEARNING




About

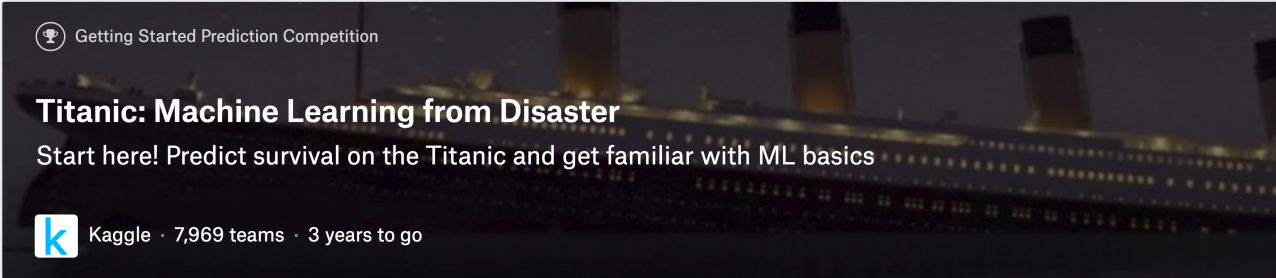
- Kaggle: Titanic
- Overview
- Data I/O (Today!)
- Model
- Loss & Optimization
- Visualize training stats
- Inference: Saving and Loading Variables


Clone

```
git clone https://github.com/seilna/ml-practice.git
```

Kaggle: Titanic


 Search kaggle  Competitions Datasets Kernels Discussion Jobs ... 



 Getting Started Prediction Competition

Titanic: Machine Learning from Disaster

Start here! Predict survival on the Titanic and get familiar with ML basics

 Kaggle · 7,969 teams · 3 years to go

[Overview](#) [Data](#) [Kernels](#) [Discussion](#) [Leaderboard](#) [Rules](#) [Team](#) [My Submissions](#) [Submit Predictions](#)

Overview

[Description](#)
[Evaluation](#)
[Frequently Asked Questions](#)
[Tutorials](#)

Start here if...

You're new to data science and machine learning, or looking for a simple intro to the Kaggle prediction competitions.

Competition Description

The sinking of the RMS Titanic is one of the most infamous shipwrecks in history. On April 15, 1912, during her maiden voyage, the Titanic sank after colliding with an iceberg, killing 1502 out of 2224 passengers and crew. This sensational tragedy shocked the international community and led to better safety regulations for ships.

One of the reasons that the shipwreck led to such loss of life was that there were not enough lifeboats for the passengers and crew. Although there was some element of luck involved in surviving the sinking, some groups of people were more likely to survive than others, such as women, children, and the upper-class.

In this challenge, we ask you to complete the analysis of what sorts of people were likely to survive. In particular, we ask you to apply the tools of machine learning to predict which passengers survived the tragedy.

Kaggle: Titanic

Data

Overview

Data

Kernels

Discussion

Leaderboard

Rules

Team

My Submissions

Submit Predictions

Additional Files

gender_submission.cs...

test.csv

train.csv

train.csv

59.76 KB

Download

Data Introduction

Overview

The data has been split into two groups:

- training set (train.csv)
- test set (test.csv)

The **training set** should be used to build your machine learning models. For the training set, we provide the outcome (also known as the “ground truth”) for each passenger. Your model will be based on “features” like passengers’ gender and class. You can also use **feature engineering** to create new features.

The **test set** should be used to see how well your model performs on unseen data. For the test set, we do not provide the ground truth for each passenger. It is your job to predict these outcomes. For each passenger in the test set, use the model you trained to predict whether or not they survived the sinking of the Titanic.

Data Dictionary

Variable	Definition	Key
survival	Survival	0 = No, 1 = Yes
pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
sex	Sex	
Age	Age in years	
sibsp	# of siblings / spouses aboard the Titanic	

Kaggle: Titanic

train.csv : 891 명에 대한 정보와 생존 여부(0, 1)

	A	B	C	D	E	F	G	H	I	J	K	L
1	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
2	1	0	3	Braund, Mr. C	male	22	1	0	A/5 21171	7.25		S
3	2	1	1	Cumings, Mrs	female	38	1	0	PC 17599	71.2833	C85	C
4	3	1	3	Heikkinen, M	female	26	0	0	STON/O2. 31	7.925		S
5	4	1	1	Futrelle, Mrs.	female	35	1	0	113803	53.1	C123	S
6	5	0	3	Allen, Mr. Wil	male	35	0	0	373450	8.05		S
7	6	0	3	Moran, Mr. J	male		0	0	330877	8.4583		Q
8	7	0	1	McCarthy, Mr	male	54	0	0	17463	51.8625	E46	S
9	8	0	3	Palsson, Mast	male	2	3	1	349909	21.075		S
10	9	1	3	Johnson, Mrs	female	27	0	2	347742	11.1333		S
11	10	1	2	Nasser, Mrs.	female	14	1	0	237736	30.0708		C
12	11	1	3	Sandstrom, M	female	4	1	1	PP 9549	16.7	G6	S
13	12	1	1	Bonnell, Miss	female	58	0	0	113783	26.55	C103	S
14	13	0	3	Saunderscock,	male	20	0	0	A/5. 2151	8.05		S
15	14	0	3	Andersson, M	male	39	1	5	347082	31.275		S
16	15	0	3	Vestrom, Mis	female	14	0	0	350406	7.8542		S
17	16	1	2	Hewlett, Mrs.	female	55	0	0	248706	16		S
18	17	0	3	Rice, Master.	male	2	4	1	382652	29.125		Q
19	18	1	2	Williams, Mr.	male		0	0	244373	13		S
20	19	0	3	Vander Plank	female	31	1	0	345763	18		S
21	20	1	3	Masselmani,	female		0	0	2649	7.225		C
22	21	0	2	Fynney, Mr. J	male	35	0	0	239865	26		S
23	22	1	2	Beesley, Mr. L	male	34	0	0	248698	13	D56	S
24	23	1	3	McGowan, M	female	15	0	0	330923	8.0292		Q

Kaggle: Titanic

test.csv : 약 400 명에 대한 정보가 주어졌을 때, 생존 여부를 추측

	A	B	C	D	E	F	G	H	I	J	K
1	PassengerId	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
2	892	3	Kelly, Mr. Jan	male	34.5	0	0	330911	7.8292		Q
3	893	3	Wilkes, Mrs. J	female	47	1	0	363272	7		S
4	894	2	Myles, Mr. Th	male	62	0	0	240276	9.6875		Q
5	895	3	Wirz, Mr. Alb	male	27	0	0	315154	8.6625		S
6	896	3	Hirvonen, Mr.	female	22	1	1	3101298	12.2875		S
7	897	3	Svensson, Mr	male	14	0	0	7538	9.225		S
8	898	3	Connolly, Mis	female	30	0	0	330972	7.6292		Q
9	899	2	Caldwell, Mr.	male	26	1	1	248738	29		S
10	900	3	Abraham, Mrs	female	18	0	0	2657	7.2292		C
11	901	3	Davies, Mr. J	male	21	2	0	A/4 48871	24.15		S
12	902	3	Ilieff, Mr. Yli	male		0	0	349220	7.8958		S
13	903	1	Jones, Mr. Ch	male	46	0	0	694	26		S
14	904	1	Snyder, Mrs. .	female	23	1	0	21228	82.2667	B45	S
15	905	2	Howard, Mr. I	male	63	1	0	24065	26		S
16	906	1	Chaffee, Mrs.	female	47	1	0	W.E.P. 5734	61.175	E31	S
17	907	2	del Carlo, Mr.	female	24	1	0	SC/PARIS 216	27.7208		C
18	908	2	Keane, Mr. D	male	35	0	0	233734	12.35		Q
19	909	3	Assaf, Mr. Ge	male	21	0	0	2692	7.225		C
20	910	3	Ilmakangas, M	female	27	1	0	STON/O2. 31	7.925		S
21	911	3	Assaf Khalil,	female	45	0	0	2696	7.225		C

Kaggle: Titanic

gender_submission.csv

Download

```
train.csv: wget https://www.kaggle.com/c/titanic/download/train.csv
```

```
test.csv: wget https://www.kaggle.com/c/titanic/download/test.csv
```

```
git clone https://github.com/seilna/ml-practice.git
```

Code

`reader.py` : raw_data `train.csv` / `test.csv` 를 읽어 numpy array로 변환하고,
batching하는 코드 포함

`reader_answer.py` : Quiz에 대한 정답(나중에 보시는 것을 추천)

reader.py

- 보통 python class style 로 구성한다
- *.csv 파일을 읽고, 이를 numpy array 형태로 변환한다
- 구성한 numpy array로부터 batch_size 개 만큼의 example로 이루어진 batch 를 가져오는 next_batch 함수를 method로 갖는다

.csv → **np.array (raw_to_vector())**

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. C	male	22	1	0	A/5 21171	7.25		S

Python csv reader

```
['PassengerId', 'Survived', 'Pclass', 'Name', 'Sex', 'Age', 'SibSp', 'Parch', 'Ticket', 'Fare', 'Cabin', 'Embarked']  
['1', '0', '3', 'Braund, Mr. Owen Harris', 'male', '22', '1', '0', 'A/5 21171', '7.25', '', 'S']
```

x: Pclass, Sex, Age, SibSp, Parch, Fare
y: Survived
Id: PassengerId

```
x: [ 3.  0. 22.  1.  0.  7.25]  
y: 0  
id: 1
```

.csv → **np.array (raw_to_vector())**

전체 training examples 개수가 891개이고,

각 example당 **x.shape=(6), y.shape=(1), id.shape=(1)**

이므로, 모든 example에 대한 shape은 **x.shape=(891, 6), y.shape=(891), id.shape=(891)**

Batching

다음과 같은 기능을 하는 함수 `next_batch()`를 `reader` class의 method로 만들자

- 모든 training example `x = (891, 6)`, `y = (891)`, `id = (891)` 중에서
- batch를 구성하기 위해 `batch_size`개 만큼의 example을 **랜덤으로** 가져온다
- 따라서, `x_batch = (batch_size, 6)`, `y_batch = (batch_size)`,
`id_batch = (batch_size)`

Batching

다음과 같은 기능을 하는 함수 `next_batch()` 를 `reader` class의 method로 만들자

- 단, 같은 epoch 내에서는 random으로 가져오는 example들이 겹쳐서는 안된다
- 또한, 1 epoch를 돌 때 마다, 모든 example들에 대한 순서를 섞어주어야 (shuffle) 한다

Practice (1)

여러분만의 `reader.py` 를 구현해보세요!

- sample code의 `reader.py`는 참고만하여 직접 새로 구현하시는 것을 추천드립니다

Practice (2)

`reader.py` 이 다음과 같은 기능을 갖도록 수정해보세요!

- 891 개 training example을 800개 train/ 100개 validation example로 split하여 따로 관리(보관)
- 따라서, `next_batch()` 함수도 train batch를 불러올 때/ validation batch를 불러올 때 각각 다르게 동작해야 함
- 충분히 고민해보고, 잘 감이 안잡힌다면 `reader_answer.py` 를 [참고](#)해보세요

Thank You!

@Seil Na, @Youngjae Yu

Special Thanks to: Byeongchang Kim, Jongwook Choi, Cesc Park, Sungjoon Choi

Slideshow created using [remark](#).