

Redes neurais sem peso

abril/2018

Grupo plain-text

- Rafael Nunes (rnunes@cos.ufrj.br)
- Marcelle Ramos (panzariello@cos.ufrj.br)
- Rodrigo Azevedo (rodrigoasantos@cos.ufrj.br)
- Vinicius Deodoro (vinicius.deodoro@ufrj.br)
- Débora Lima (teredeby@gmail.com)

Texto

Classificação

Múltiplas classes

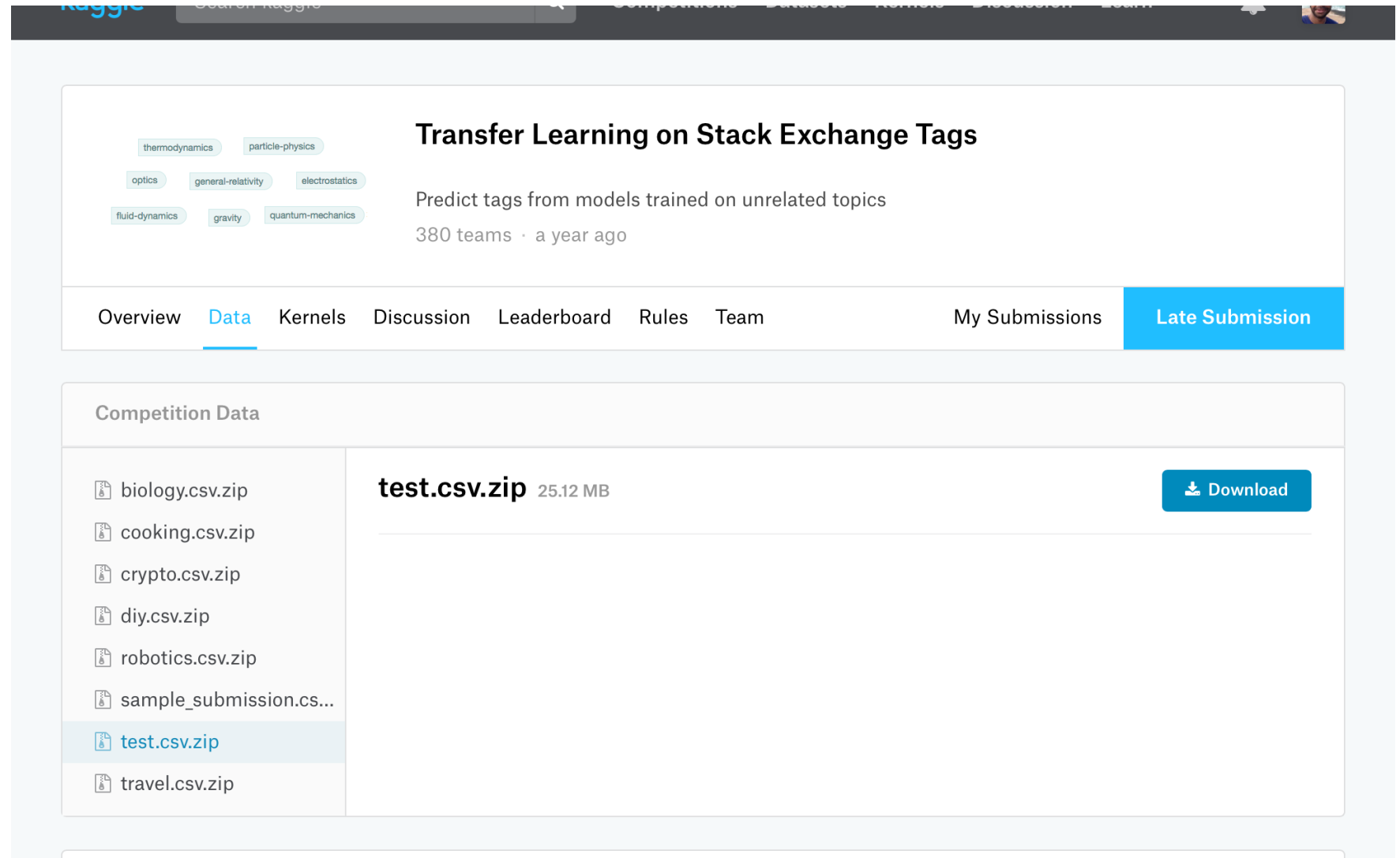
The screenshot shows a web browser on an iPad displaying the Kaggle website. The address bar shows 'kaggle.com'. The page title is 'Transfer Learning on Stack Exchange Tags'. Below the title, there are tags: 'thermodynamics', 'particle-physics', 'optics', 'general-relativity', 'electrostatics', 'fluid-dynamics', 'gravity', and 'quantum-mechanics'. The description states: 'Predict tags from models trained on unrelated topics' and '380 teams · a year ago'. The navigation bar includes 'Overview', 'Data', 'Kernels', 'Discussion', 'Leaderboard', 'Rules', 'Team', 'My Submissions', and 'Late Submission'. The 'Overview' section is active, showing a table with 'Description' and 'Evaluation' columns. The 'Description' column contains the text: 'What does physics have in common with biology, cooking, cryptography, diy, robotics, and travel? If you answered "all pursuits are governed by the immutable laws of physics" we'll begrudgingly give you partial credit. If you answered "all were chosen randomly by a scheming Kaggle employee for a twisted transfer learning competition", congratulations, we accept your answer and mark the question as solved.' The 'Evaluation' column contains the text: 'In this competition, we provide the titles, text, and tags of Stack Exchange questions from six different sites. We then ask for tag predictions on unseen physics questions. Solving this problem via a standard machine approach might involve training an algorithm on a corpus of related text. Here, you are challenged to train on material from outside the field. Can an algorithm learn appropriate physics tags from "extreme-tourism Antarctica"? Let's find out.' The 'Late Submission' button is highlighted in blue.

| Overview | |
|-------------|--|
| Description | What does physics have in common with biology, cooking, cryptography, diy, robotics, and travel? If you answered "all pursuits are governed by the immutable laws of physics" we'll begrudgingly give you partial credit. If you answered "all were chosen randomly by a scheming Kaggle employee for a twisted transfer learning competition", congratulations, we accept your answer and mark the question as solved. |
| Evaluation | In this competition, we provide the titles, text, and tags of Stack Exchange questions from six different sites. We then ask for tag predictions on unseen physics questions. Solving this problem via a standard machine approach might involve training an algorithm on a corpus of related text. Here, you are challenged to train on material from outside the field. Can an algorithm learn appropriate physics tags from "extreme-tourism Antarctica"? Let's find out. |

dataset

6 classes:

- biology
- cooking
- travel
- diy
- robotics
- crypto



The screenshot shows the Kaggle interface for a competition titled "Transfer Learning on Stack Exchange Tags". The competition description is "Predict tags from models trained on unrelated topics" and it was created "380 teams · a year ago". The navigation bar includes "Overview", "Data", "Kernels", "Discussion", "Leaderboard", "Rules", "Team", "My Submissions", and "Late Submission". The "Data" tab is selected, showing a list of data files under "Competition Data". The files are: biology.csv.zip, cooking.csv.zip, crypto.csv.zip, diy.csv.zip, robotics.csv.zip, sample_submission.cs..., test.csv.zip (highlighted), and travel.csv.zip. The "test.csv.zip" file is 25.12 MB and has a "Download" button next to it.

| File Name | Size | Action |
|-------------------------|-----------------|-----------------|
| biology.csv.zip | | |
| cooking.csv.zip | | |
| crypto.csv.zip | | |
| diy.csv.zip | | |
| robotics.csv.zip | | |
| sample_submission.cs... | | |
| test.csv.zip | 25.12 MB | Download |
| travel.csv.zip | | |

datasets available on [kaggle.com](https://www.kaggle.com)

dataset – características

```
In [4]: for _class in df:
        print(_class, df[_class].shape)
```

```
cooking (15404, 4)
crypto (10432, 4)
robotics (2771, 4)
biology (13196, 4)
travel (19279, 4)
diy (25918, 4)
```

```
In [5]: df['physics'].iloc[1]
```

```
Out[5]: id                                2
        title      What is your simplest explanation of the strin...
        content    <p>How would you explain string theory to non ...
        Name: 1, dtype: object
```

```
In [6]: for file in df:
        display(df[file].head())
        print('{0}: {1} questions'.format(file, df[file].shape[0]))
```

| | id | title | content | tags |
|---|----|---|---|---|
| 0 | 1 | How can I get chewy chocolate chip cookies? | <p>My chocolate chips cookies are always too c... | baking cookies texture |
| 1 | 2 | How should I cook bacon in an oven? | <p>I've heard of people cooking bacon in an ov... | oven cooking-time bacon |
| 2 | 3 | What is the difference between white and brown... | <p>I always use brown extra large eggs, but I ... | eggs |
| 3 | 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... |
| 4 | 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 |

objetivo

- **Validação cruzada com 10 folds:**
 - **Classificador Naïve-Bayes (baseline);**
 - **Classificador WiSARD;**
 - **Classificador SVM;**
 - **Classificador XGBoost;**
- **Comparar a performance dos vários classificadores;**

github

- <https://github.com/rafaelscnunes/mab786>
- Repositório privado para colaboração dos participantes do grupo