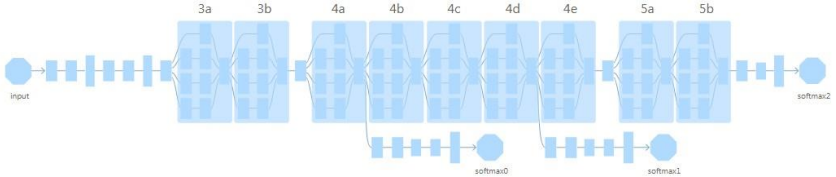
Google TensorFlow™

浙江大学城市学院 ZHEJIANG UNIVERSITY CITY COLLEGE



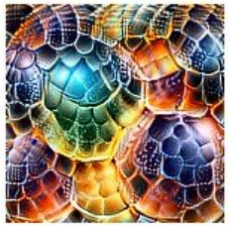

# 深度学习应用开发
## 基于TensorFlow的实践

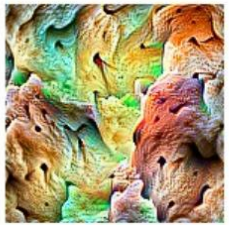
吴明晖 李卓蓉 金苍宏

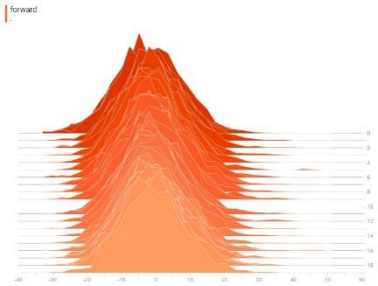
浙江大学城市学院

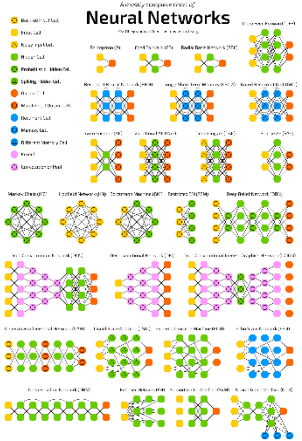
计算机与计算科学学院

Dept. of Computer Science
Zhejiang University City College







# 泰坦尼克号旅客生存预测

## TensorFlow高级API：Keras应用实践

浙江大学城市学院
ZHEJIANG UNIVERSITY CITY COLLEGE


A machine learning platform

**for everyone**

to solve real problems

# 低阶灵活、高阶易用

开箱即用模型 → Premade Estimators

分布式执行，
集成了TensorBoard，
tensorflow serving，...
→ Estimator

建模的高阶接口 → tf.keras

建模的低阶接口 → tf.*

# 泰坦尼克号上的旅客生存概率预测：TensorFlow的高级框架Keras

```
In [33]: pd[-2:]
```

Out[33]:

| | survived | name | pclass | sex | age | sibsp | parch | fare | embarked | probability |
|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 0 | Jack | 3 | male | 23.0 | 1 | 0 | 5.0 | S | 0.150541 |
| **1** | 1 | Rose | 1 | female | 20.0 | 1 | 0 | 100.0 | S | 0.969736 |


| 字段 | 字段说明 | 数据说明 |
|---|---|---|
| pclass | 舱等 | 1 头等舱，2 二等舱，3 三等舱 |
| survival | 是否生存 | 0 否，1 是 |
| name | 姓名 | |
| sex | 性别 | Female 女性，male 男 |
| age | 年龄 | |
| sibsp | 兄弟姐妹或者配偶也在船上的数量 | |
| parch | 双亲或者子女也在船上的数量 | |
| ticked | 船票号码 | |
| fare | 船票费用 | |
| cabin | 舱位号码 | |
| embarked | 登船港口 | C=Cherbourg, Q=Queenstown, S=Southampton |

浙江大学城市学院
ZHEJIANG UNIVERSITY CITY COLLEGE

# 泰坦尼克号数据处理

# 下载泰坦尼克号上旅客的数据集



下载旅客数据集

```
import urllib.request
import os

data_url="http://biostat.mc.vanderbilt.edu/wiki/pub/Main/DataSets/titanic3.xls"

data_file_path="data/titanic3.xls"

if not os.path.isfile(data_file_path):
    result=urllib.request.urlretrieve(data_url,data_file_path)
    print('downloaded:',result)
else:
    print(data_file_path,'data file already exists.')
```

```
downloaded: ('data/titanic3.xls', <http.client.HTTPMessage object at 0x000001F384057F60>)
```

# 查看数据集

浙江大学城市学院
ZHEJIANG UNIVERSITY CITY COLLEGE

| | A | B | C | D | E | F | G | H | I | J | K | L | M | N |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | **pclass** | **survived** | **name** | **sex** | **age** | **sibsp** | **parch** | **ticket** | **fare** | **cabin** | **embarked** | **boat** | **body** | **home.dest** |
| 2 | 1 | 1 | Allen, Miss. Elisabeth Walton | female | 29 | 0 | 0 | 24160 | 211.3375 | B5 | S | 2 | | St Louis, MO |
| 3 | 1 | 1 | Allison, Master. Hudson Trevor | male | 0.917 | 1 | 2 | 113781 | 151.5500 | C22 C26 | S | 11 | | Montreal, PQ / Chesterville, ON |
| 4 | 1 | 0 | Allison, Miss. Helen Loraine | female | 2 | 1 | 2 | 113781 | 151.5500 | C22 C26 | S | | | Montreal, PQ / Chesterville, ON |
| 5 | 1 | 0 | Allison, Mr. Hudson Joshua Creighton | male | 30 | 1 | 2 | 113781 | 151.5500 | C22 C26 | S | | 135 | Montreal, PQ / Chesterville, ON |
| 6 | 1 | 0 | Allison, Mrs. Hudson J C (Bessie Waldo Daniels) | female | 25 | 1 | 2 | 113781 | 151.5500 | C22 C26 | S | | | Montreal, PQ / Chesterville, ON |
| 7 | 1 | 1 | Anderson, Mr. Harry | male | 48 | 0 | 0 | 19952 | 26.5500 | E12 | S | 3 | | New York, NY |
| 8 | 1 | 1 | Andrews, Miss. Kornelia Theodosia | female | 63 | 1 | 0 | 13502 | 77.9583 | D7 | S | 10 | | Hudson, NY |
| 9 | 1 | 0 | Andrews, Mr. Thomas Jr | male | 39 | 0 | 0 | 112050 | 0.0000 | A36 | S | | | Belfast, NI |
| 10 | 1 | 1 | Appleton, Mrs. Edward Dale (Charlotte Lamson) | female | 53 | 2 | 0 | 11769 | 51.4792 | C101 | S | D | | Bayside, Queens, NY |
| 11 | 1 | 0 | Artagaveytia, Mr. Ramon | male | 71 | 0 | 0 | PC 17609 | 49.5042 | | C | | 22 | Montevideo, Uruguay |
| 12 | 1 | 0 | Astor, Col. John Jacob | male | 47 | 1 | 0 | PC 17757 | 227.5250 | C62 C64 | C | | 124 | New York, NY |
| 1302 | 3 | 1 | Yasbeck, Mrs. Antoni (Selini Alexander) | female | 15 | 1 | 0 | 2659 | 14.4542 | | C | | | |
| 1303 | 3 | 0 | Youseff, Mr. Gerious | male | 45.5 | 0 | 0 | 2628 | 7.2250 | | C | | 312 | |
| 1304 | 3 | 0 | Yousif, Mr. Wazli | male | | 0 | 0 | 2647 | 7.2250 | | C | | | |
| 1305 | 3 | 0 | Youssef, Mr. Gerious | male | | 0 | 0 | 2627 | 14.4583 | | C | | | |
| 1306 | 3 | 0 | Zabour, Miss. Hileni | female | 14.5 | 1 | 0 | 2665 | 14.4542 | | C | | 328 | |
| 1307 | 3 | 0 | Zabour, Miss. Thamine | female | | 1 | 0 | 2665 | 14.4542 | | C | | | |
| 1308 | 3 | 0 | Zakarian, Mr. Mapriededer | male | 26.5 | 0 | 0 | 2656 | 7.2250 | | C | | 304 | |
| 1309 | 3 | 0 | Zakarian, Mr. Ortin | male | 27 | 0 | 0 | 2670 | 7.2250 | | C | | | |
| 1310 | 3 | 0 | Zimmerman, Mr. Leo | male | 29 | 0 | 0 | 315082 | 7.8750 | | S | | | |

# 查看数据集

浙江大学城市学院 ZHEJIANG UNIVERSITY CITY COLLEGE

| | A | B | C | D | E | F | G | H | I | J | K | L | M | N |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | pclass | survived | name | sex | age | sibsp | parch | ticket | fare | cabin | embarked | boat | body | home.dest |
| 2 | 1 | 1 | Allen, Miss. Elisabeth Walton | female | 29 | 0 | 0 | 24160 | 211.3375 | B5 | S | 2 | | St Louis, MO |
| 3 | 1 | 1 | Allison, Master. Hudson Trevor | | | | | | | | | | | Montreal, PQ / Chesterville, ON |
| 4 | 1 | 0 | Allison, Miss. Helen Loraine | | | | | | | | | | | Montreal, PQ / Chesterville, ON |
| 5 | 1 | 0 | Allison, Mr. Hudson Joshua Creighton | | | | | | | | | | 35 | Montreal, PQ / Chesterville, ON |
| 6 | 1 | 0 | Allison, Mrs. Hudson J C (Bessie Wal | | | | | | | | | | | Montreal, PQ / Chesterville, ON |
| 7 | 1 | 1 | Anderson, Mr. Harry | | | | | | | | | | | New York, NY |
| 8 | 1 | 1 | Andrews, Miss. Kornelia Theodosia | | | | | | | | | | | Hudson, NY |
| 9 | 1 | 0 | Andrews, Mr. Thomas Jr | | | | | | | | | | | Belfast, NI |
| 10 | 1 | 1 | Appleton, Mrs. Edward Dale (Charlot | | | | | | | | | | | Bayside, Queens, NY |
| 11 | 1 | 0 | Artagaveytia, Mr. Ramon | | | | | | | | | | 2 | Montevideo, Uruguay |
| 12 | 1 | 0 | Astor, Col. John Jacob | | | | | | | | | | 24 | New York, NY |
| 1302 | 3 | 1 | Yasbeck, Mrs. Antoni (Selini Alexand | | | | | | | | | | | |
| 1303 | 3 | 0 | Youseff, Mr. Gerious | | | | | | | | | | 12 | |
| 1304 | 3 | 0 | Yousif, Mr. Wazli | | | | | | | | | | | |
| 1305 | 3 | 0 | Youseff, Mr. Gerious | | | | | | | | | | | |
| 1306 | 3 | 0 | Zabour, Miss. Hileni | | | | | | | | | | 28 | |
| 1307 | 3 | 0 | Zabour, Miss. Thamine | | | | | | | | | | | |
| 1308 | 3 | 0 | Zakarian, Mr. Mapriededer | | | | | | | | | | 04 | |
| 1309 | 3 | 0 | Zakarian, Mr. Ortin | | | | | | | | | | | |
| 1310 | 3 | 0 | Zimmerman, Mr. Leo | | | | | | | | | | | |

| 字段 | 字段说明 | 数据说明 |
|---|---|---|
| pclass | 舱等 | 1 头等舱，2 二等舱，3 三等舱 |
| survival | 是否生存 | 0 否，1 是 |
| name | 姓名 | |
| sex | 性别 | Female 女性，male 男 |
| age | 年龄 | |
| sibsp | 兄弟姐妹或者配偶也在船上的数量 | |
| parch | 双亲或者子女也在船上的数量 | |
| ticked | 船票号码 | |
| fare | 船票费用 | |
| cabin | 舱位号码 | |
| embarked | 登船港口 | C=Cherbourg, Q=Queenstown, S=Southampton |

# 使用Pandas读取处理数据

```
import numpy
import pandas as pd

# 读取数据文件，结果为DataFrame格式
df_data = pd.read_excel(data_file_path)
```

# 使用Pandas读取处理数据

```
#查看数据摘要
df_data.describe()
```

| | pclass | survived | age | sibsp | parch | fare | body |
|---|---|---|---|---|---|---|---|
| count | 1309.000000 | 1309.000000 | 1046.000000 | 1309.000000 | 1309.000000 | 1308.000000 | 121.000000 |
| mean | 2.294882 | 0.381971 | 29.881135 | 0.498854 | 0.385027 | 33.295479 | 160.809917 |
| std | 0.837836 | 0.486055 | 14.413500 | 1.041658 | 0.865560 | 51.758668 | 97.696922 |
| min | 1.000000 | 0.000000 | 0.166700 | 0.000000 | 0.000000 | 0.000000 | 1.000000 |
| 25% | 2.000000 | 0.000000 | 21.000000 | 0.000000 | 0.000000 | 7.895800 | 72.000000 |
| 50% | 3.000000 | 0.000000 | 28.000000 | 0.000000 | 0.000000 | 14.454200 | 155.000000 |
| 75% | 3.000000 | 1.000000 | 39.000000 | 1.000000 | 0.000000 | 31.275000 | 256.000000 |
| max | 3.000000 | 1.000000 | 80.000000 | 8.000000 | 9.000000 | 512.329200 | 328.000000 |

# 使用Pandas读取处理数据

浙江大学城市学院 ZHEJIANG UNIVERSITY CITY COLLEGE

```
df_data
```

| | pclass | survived | name | sex | age | sibsp | parch | ticket | fare | cabin | embarked | boat | body | home.dest |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 1 | 1 | Allen, Miss. Elisabeth Walton | female | 29.0000 | 0 | 0 | 24160 | 211.3375 | B5 | S | 2 | NaN | St Louis, MO |
| **1** | 1 | 1 | Allison, Master. Hudson Trevor | male | 0.9167 | 1 | 2 | 113781 | 151.5500 | C22 C26 | S | 11 | NaN | Montreal, PQ / Chesterville, ON |
| **2** | 1 | 0 | Allison, Miss. Helen Loraine | female | 2.0000 | 1 | 2 | 113781 | 151.5500 | C22 C26 | S | NaN | NaN | Montreal, PQ / Chesterville, ON |
| **3** | 1 | 0 | Allison, Mr. Hudson Joshua Creighton | male | 30.0000 | 1 | 2 | 113781 | 151.5500 | C22 C26 | S | NaN | 135.0 | Montreal, PQ / Chesterville, ON |
| **4** | 1 | 0 | Allison, Mrs. Hudson J C (Bessie Waldo Daniels) | female | 25.0000 | 1 | 2 | 113781 | 151.5500 | C22 C26 | S | NaN | NaN | Montreal, PQ / Chesterville, |

# 筛选提取字段

survival（是否生存）是标签字段，其他是候选特征字段

筛选提取需要的特征字段，去掉ticket，cabin等

```
#筛选提取需要的特征字段，去掉ticket，cabin等
selected_cols=['survived','name','pclass' ,'sex', 'age', 'sibsp', 'parch', 'fare', 'embarked']
selected_df_data=df_data[selected_cols]
```

```
selected_df_data
```

| | survived | name | pclass | sex | age | sibsp | parch | fare | embarked |
|---|---|---|---|---|---|---|---|---|---|
| **0** | 1 | Allen, Miss. Elisabeth Walton | 1 | female | 29.0000 | 0 | 0 | 211.3375 | S |
| **1** | 1 | Allison, Master. Hudson Trevor | 1 | male | 0.9167 | 1 | 2 | 151.5500 | S |
| **2** | 0 | Allison, Miss. Helen Loraine | 1 | female | 2.0000 | 1 | 2 | 151.5500 | S |
| **3** | 0 | Allison, Mr. Hudson Joshua Creighton | 1 | male | 30.0000 | 1 | 2 | 151.5500 | S |

# 数据的进一步处理

# 还需要进一步处理的问题

| 字段 | 处理方式 |
| --- | --- |
| name | 姓名字段训练时不需要，但在预测阶段会使用，先制作一份不含名字的数据集 |
| age | 有一些数据的age字段是null值，把null值改为平均值 |
| fare | 有一些数据的fare字段是null值，把null值改为平均值 |
| sex | 性别字段是文字，需要转换为数字，如：0与1 |
| embarked | 有一些数据的embarked字段是null值，把null值改为某一个值，如：S |
| embarked | 分类特征字段是文字（C、Q、S），需要转换为数字 |

# 找出有null值的字段

Pandas判断缺失值一般采用 isnull()，生成所有数据的True / False矩阵

这是元素级别的判断，把对应的所有元素的位置都列出来，元素为空或者NA就显示True，否则就是False

```
selected_df_data.isnull()
```

| | survived | name | pclass | sex | age | sibsp | parch | fare | embarked |
|---|---|---|---|---|---|---|---|---|---|
| 0 | False | False | False | False | False | False | False | False | False |
| 1 | False | False | False | False | False | False | False | False | False |
| 2 | False | False | False | False | False | False | False | False | False |
| 3 | False | False | False | False | False | False | False | False | False |
| 4 | False | False | False | False | False | False | False | False | False |
| 5 | False | False | False | False | False | False | False | False | False |
| 6 | False | False | False | False | False | False | False | False | False |

# 找出有null值的字段

判断哪些”列”存在缺失值

列级别的判断，只要该列有为空或者NA的元素，就为True，否则False

```
selected_df_data.isnull().any()
```

```
survived    False
name        False
pclass      False
sex         False
age          True
sibsp       False
parch       False
fare         True
embarked     True
dtype: bool
```

# 找出有null值的字段

判断哪些”列”存在缺失值，将列中为空的个数统计出来

```
selected_df_data.isnull().sum()
```

```
survived      0
name          0
pclass        0
sex           0
age         263
sibsp         0
parch         0
fare          1
embarked      2
dtype: int64
```

# 找出有null值的字段

显示存在缺失值的行列，确定缺失值的位置

```
selected_df_data[selected_df_data.isnull().values==True]
```

| | survived | name | pclass | sex | age | sibsp | parch | fare | embarked |
|---|---|---|---|---|---|---|---|---|---|
| **15** | 0 | Baumann, Mr. John D | 1 | male | NaN | 0 | 0 | 25.9250 | S |
| **37** | 1 | Bradley, Mr. George ("George Arthur Brayton") | 1 | male | NaN | 0 | 0 | 26.5500 | S |
| **40** | 0 | Brewe, Dr. Arthur Jackson | 1 | male | NaN | 0 | 0 | 39.6000 | C |
| **46** | 0 | Cairns, Mr. Alexander | 1 | male | NaN | 0 | 0 | 31.0000 | S |
| **59** | 1 | Cassebeer, Mrs. Henry Arthur Jr (Eleanor Genev... | 1 | female | NaN | 0 | 0 | 27.7208 | C |
| **69** | 1 | Chibnall, Mrs. (Edith Martha Bowerman) | 1 | female | NaN | 0 | 1 | 55.0000 | S |
| **70** | 0 | Chisholm, Mr. Roderick Robert Crispin | 1 | male | NaN | 0 | 0 | 0.0000 | S |
| **74** | 0 | Clifford, Mr. George Quincy | 1 | male | NaN | 0 | 0 | 52.0000 | S |
| **80** | 0 | Crafton, Mr. John Bertram | 1 | male | NaN | 0 | 0 | 26.5500 | S |

# 填充null值

显示存在缺失值的行列，确定缺失值的位置

```
#为缺失age记录填充值 设置为平均值
age_mean_value = selected_df_data['age'].mean()
selected_df_data['age'] = selected_df_data['age'].fillna(age_mean_value)

#为缺失fare记录填充值
fare_mean_value = selected_df_data['fare'].mean()
selected_df_data['fare'] = selected_df_data['fare'].fillna(fare_mean_value)

# #为缺失embarked记录填充值
selected_df_data['embarked'] = selected_df_data['embarked'].fillna('S')
```

# 转换编码

性别和港口数据转换编码

```
# 性别sex由字符串转换为数字编码
selected_df_data['sex']= selected_df_data['sex'].map({'female':0, 'male': 1}).astype(int)

# 港口embarked由字母表示转换为数字编码
selected_df_data['embarked']= selected_df_data['embarked'].map({'C':0, 'Q': 1,'S': 2}).astype(int)
```

```
selected_df_data[:3]
```

| | survived | name | pclass | sex | age | sibsp | parch | fare | embarked |
|---|---|---|---|---|---|---|---|---|---|
| **0** | 1 | Allen, Miss. Elisabeth Walton | 1 | 0 | 29.0000 | 0 | 0 | 211.3375 | 2 |
| **1** | 1 | Allison, Master. Hudson Trevor | 1 | 1 | 0.9167 | 1 | 2 | 151.5500 | 2 |
| **2** | 0 | Allison, Miss. Helen Loraine | 1 | 0 | 2.0000 | 1 | 2 | 151.5500 | 2 |

# 删除name字段

drop不改变原有的df中的数据，而是返回另一个DataFrame来存放删除后的数据

axis = 1 表示删除列

```
selected_df_data=selected_df_data.drop(['name'], axis=1)
```

```
selected_df_data[:3]
```

| | survived | pclass | sex | age | sibsp | parch | fare | embarked |
|---|---|---|---|---|---|---|---|---|
| **0** | 1 | 1 | 0 | 29.0000 | 0 | 0 | 211.3375 | 2 |
| **1** | 1 | 1 | 1 | 0.9167 | 1 | 2 | 151.5500 | 2 |
| **2** | 0 | 1 | 0 | 2.0000 | 1 | 2 | 151.5500 | 2 |

# 分离特征值和标签值

```
# 转换为ndarray数组
ndarray_data = selected_df_data.values

# 后7列是特征值
features = ndarray_data[:,1:]

# 第0列是标签值
label = ndarray_data[:,0]
```

```
features[:3]
```

```
array([[ 1.    ,  0.    , 29.    ,  0.    ,  0.    , 211.3375,
         2.    ],
       [ 1.    ,  1.    ,  0.9167,  1.    ,  2.    , 151.55  ,
         2.    ],
       [ 1.    ,  0.    ,  2.    ,  1.    ,  2.    , 151.55  ,
         2.    ]])
```

```
label[:3]
```

```
array([1., 1., 0.])
```

# 特征值标准化处理

```
from sklearn import preprocessing

# 特征值标准化
minmax_scale = preprocessing.MinMaxScaler(feature_range=(0, 1))
norm_features=minmax_scale.fit_transform(features)
```

```
norm_features[:3]
```

```
array([[0.        , 0.        , 0.36116884, 0.        , 0.        ,
        0.41250333, 1.        ],
       [0.        , 1.        , 0.00939458, 0.125     , 0.22222222,
        0.2958059 , 1.        ],
       [0.        , 0.        , 0.0229641 , 0.125     , 0.22222222,
        0.2958059 , 1.        ]])
```

# 定义数据预处理函数



```
def prepare_data(df_data):
    df=df_data.drop(['name'], axis=1) #删除姓名列
    age_mean = df['age'].mean()
    df['age'] = df['age'].fillna(age_mean)  #为缺失age记录填充值
    fare_mean = df['fare'].mean()
    df['fare'] = df['fare'].fillna(fare_mean)  #为缺失fare记录填充值
    df['sex']= df['sex'].map({'female':0, 'male': 1}).astype(int)  #把sex值由字符串转换为数值
    df['embarked'] = df['embarked'].fillna('S')  #为缺失embarked记录填充值
    df['embarked']=df['embarked'].map({'C':0, 'Q': 1,'S': 2}).astype(int)  #把embarked值由字符串转换为数值

    ndarray_data = df.values #转换为ndarray数组

    features = ndarray_data[:,1:] #后7列是特征值
    label = ndarray_data[:,0]     #第0列是标签值

    # 特征值标准化
    minmax_scale = preprocessing.MinMaxScaler(feature_range=(0, 1))
    norm_features=minmax_scale.fit_transform(features)

    return norm_features,label
```