ml

February 11, 2025

[79]:
```
import pandas as pd
```

[108]:
```
train = pd.read_csv("./data/train.tsv", sep="\t", index_col=0)
test = pd.read_csv("./data/test.tsv", sep="\t", index_col=0)
sample_submit = pd.read_csv("./data/sample_submit.tsv", sep="\t", index_col=0,␣
↪header=None)
```

[109]:
```
train["survived"].count()
train["survived"].value_counts()
```

[109]:
```
survived
0    266
1    179
Name: count, dtype: int64
```

[102]:
```
test.head()
```

[102]:

id	pclass	sex	age	sibsp	parch	fare	embarked
0	3	male	22.0	1	0	7.2500	S
1	1	female	38.0	1	0	71.2833	C
2	3	female	26.0	0	0	7.9250	S
5	3	male	NaN	0	0	8.4583	Q
6	1	male	54.0	0	0	51.8625	S

[119]:
```
train = train[["survived", "sibsp", "parch", "fare"]]
test = test[["sibsp", "parch", "fare"]]
```

[120]:
```
y = train["survived"]
X = train.drop(["survived"], axis=1)
```

[121]:
```
X
```

[121]:

id	sibsp	parch	fare
3	1	0	53.1000
4	0	0	8.0500

```
7        3      1  21.0750
9        1      0  30.0708
11       0      0  26.5500
..     …    …      …
873      0      0   9.0000
874      1      0  24.0000
879      0      1  83.1583
884      0      0   7.0500
888      1      2  23.4500

[445 rows x 3 columns]
```

[124]:
```
from sklearn.linear_model import LogisticRegression
model = LogisticRegression()
model.fit(X, y)
```

[124]:
```
LogisticRegression()
```

[138]:
```
pred = model.predict_log_proba(test)[:, 1]
print(pred[:5])
```

```
[-1.30612683 -0.63763617 -1.11730362 -1.11112355 -0.67490419]
```

[166]:
```
sample_submit[1] = pred
sample_submit.to_csv("submit.tsv", header=None, sep="\t")
```