機械学習プロセス徹底解説
ステップ5　特徴量エンジニアリング
アイデア２：料金の階級

16T4063F 結城洸太

アイデア２：料金の階級

リスト10

- 「Fare」(乗船料金)を6つに分類

```
train_set['Fare_Cat'] = pd.qcut(train_set['Fare'],6)
```

- 分類ごとに生存率が異なるかを表示

```
FarePlot = train_set['Survived'].groupby(train_set['Fare_Cat']).mean()
FarePlot
```

実行結果

```
#リスト10
train_set['Fare_Cat'] = pd.qcut(train_set['Fare'],6)
FarePlot = train_set['Survived'].groupby(train_set['Fare_Cat']).mean()
FarePlot
```

```
Fare_Cat
(-0.001, 7.775]      0.205128
(7.775, 8.662]       0.190789
(8.662, 14.454]      0.366906
(14.454, 26.0]       0.436242
(26.0, 52.369]       0.417808
(52.369, 512.329]    0.697987
Name: Survived, dtype: float64
```

乗船料金が高ければ高いほど、
生存確率が大幅に高まっていることがわかる

料金階級の新しい特徴量を訓練データに追加する

- 各階級に分ける
- ‘Fare Group 1’～’Fare Group 6’のラベルを付ける

```
train_set['Fare_Cat'] = pd.qcut(train_set['Fare'],6,labels = ['Fare Group 1',
                                                              'Fare Group 2',
                                                              'Fare Group 3',
                                                              'Fare Group 4',
                                                              'Fare Group 5',
                                                              'Fare Group 6',])
```

料金階級の新しい特徴量を訓練データに追加する

リスト17,リスト18からFare(料金)の部分を抜粋

- Fare(料金)をダミー変数へ変換

```
Fare = pd.get_dummies(train_set['Fare_Cat'], drop_first=True)
```

- (訓練データ)へFare(料金)のダミー変数を連結

```
train_set = pd.concat([train_set, Fare],axis=1)
```

- 不要なカラムを訓練データから削除

```
train_set = train_set.drop(columns=[‘Fare‘, 'Fare_Cat'])
```

テストデータも同様に処理する

リスト19からFare(料金)の部分を抜粋

```
test_set['Fare'].fillna(test_set['Fare'].median(), inplace=True)

test_set['Fare_Cat'] = pd.qcut(test_set['Fare'],6)
Fare_cat = pd.get_dummies(test_set['Fare_Cat'], drop_first=True)
Fare_cat.columns = ['Fare Group 2', 'Fare Group 3', 'Fare Group 4','Fare Group 5', 'Fare Group 6’]

test_set = pd.concat([test_set, Fare_cat],axis=1)
PassengerID = test_set['PassengerId']
test_set = test_set.drop(columns=['Fare','Fare_Cat'])
```

まとめ

- 「Fare」(乗船料金)を分類し、(訓練データ)と(テストデータ)に新しい特徴量を追加した。

削除した特徴量	※1	新しい特徴量				
Fare(乗船料金)	Fare Group 1	Fare Group 2	Fare Group 3	Fare Group 4	Fare Group 5	Fare Group 6
-0.001~7.775	1	0	0	0	0	0
7.775~8.662	0	1	0	0	0	0
8.662~14.454	0	0	1	0	0	0
14.454~26.0	0	0	0	1	0	0
26.0~52.369	0	0	0	0	1	0
52.369~512.329	0	0	0	0	0	1

※1　‘Fare Group2’～’Fare Group 6’がすべて0ならば’Fare Group 1’であることがわかるため、この特徴量は追加しない。