「機械学習プロセス徹底理解」PART3-Step3

河野 慎司

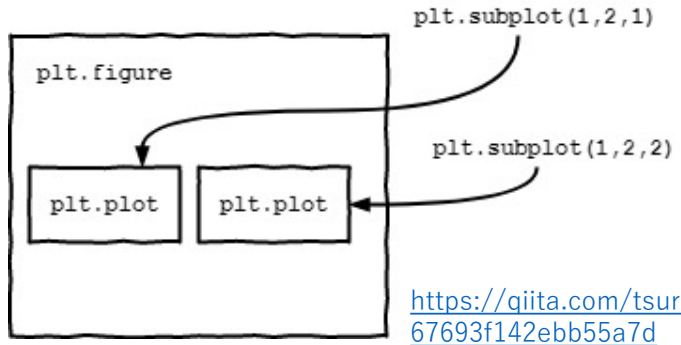
探索的データ解析(EDA)と可視化

探索的データ解析(EDA)とは,
 データの特徴を探求し, 構造を理解すること.

4つの特徴量に絞って構造を理解できるように可視化していく.

旅客等級と性別

```
fig = plt.figure(figsize=(12, 4))
ax1 = fig.add_subplot(121)
ax2 = fig.add_subplot(122)
```



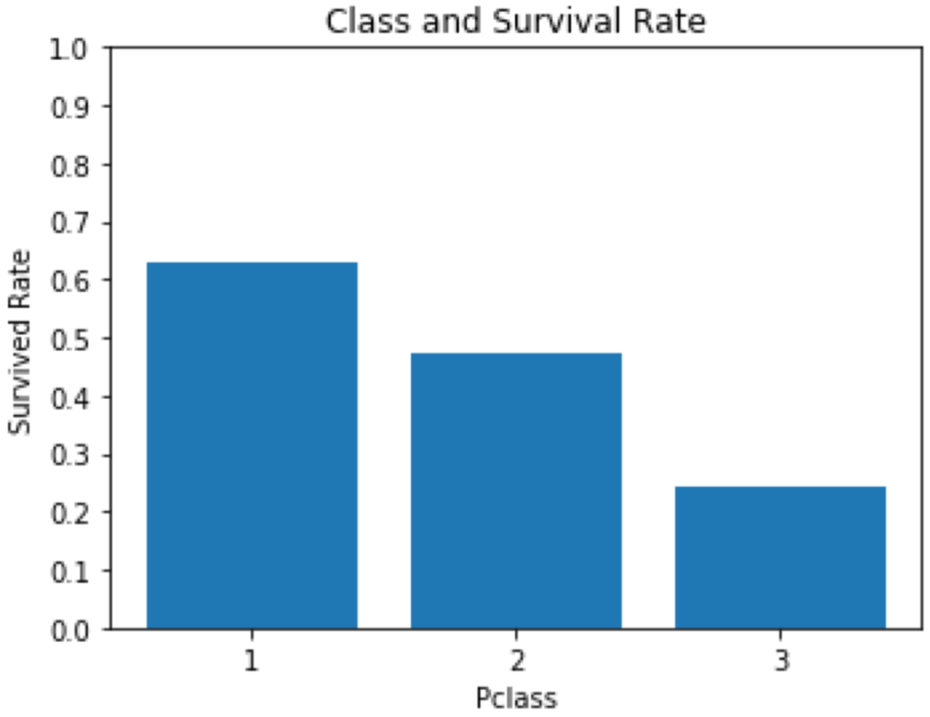
https://qiita.com/tsuruokax/items/90167693f142ebb55a7d

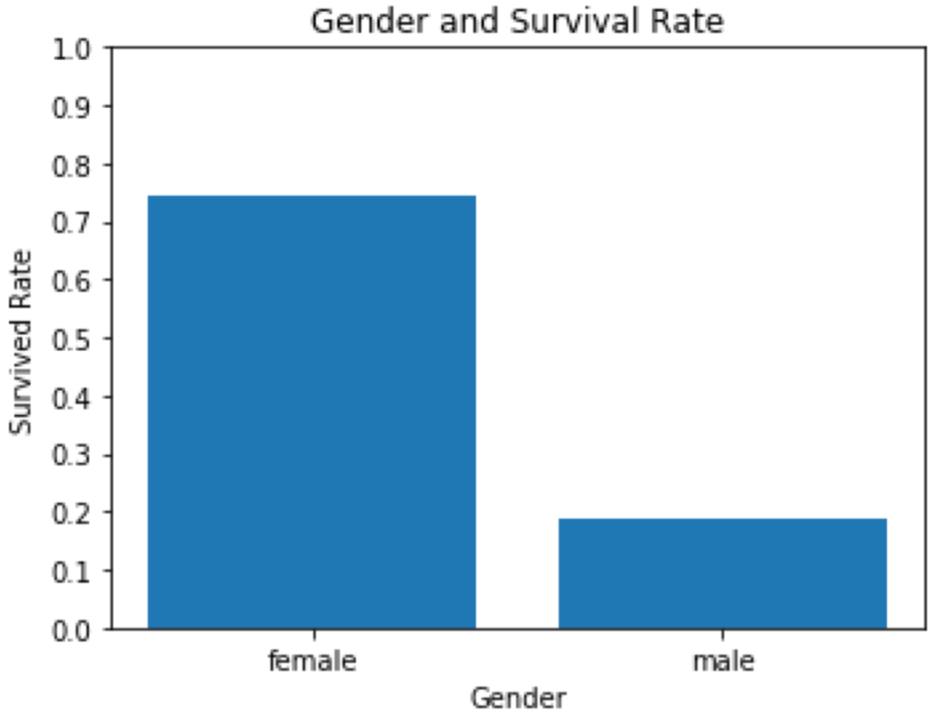
```
# PClass (旅客等級)
PClassPlot = train_set['Survived'].groupby(train_set['Pclass']).mean()
ax1.bar(x=PClassPlot.index, height=PClassPlot.values)
ax1.set_ylabel('Survived Rate')
ax1.set_xlabel('Pclass')
ax1.set_xticks(PClassPlot.index)
ax1.set_yticks(np.arange(0, 1.1, .1))
ax1.set_title("Class and Survival Rate")
```

旅客等級と性別

```
# Sex (性別)
GenderPlot = train_set['Survived'].groupby(train_set['Sex']).mean()
ax2.bar(x=GenderPlot.index, height=GenderPlot.values)
ax2.set_ylabel('Survived Rate')
ax2.set_xlabel('Gender')
ax2.set_xticks(GenderPlot.index)
ax2.set_yticks(np.arange(0, 1.1, .1))
ax2.set_title("Gender and Survival Rate")
```

実行結果





等級が高いほど生存率は高くなる.
男性より女性の方が生存率は高くなる.

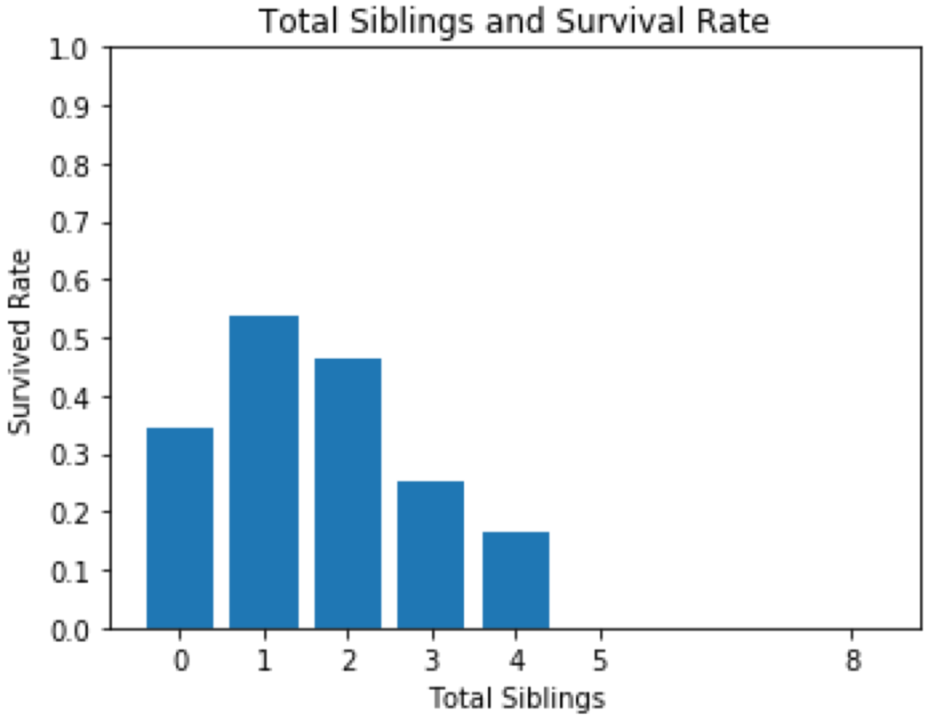
同乗中の兄弟/配偶者の数と同乗中の親/子供の数

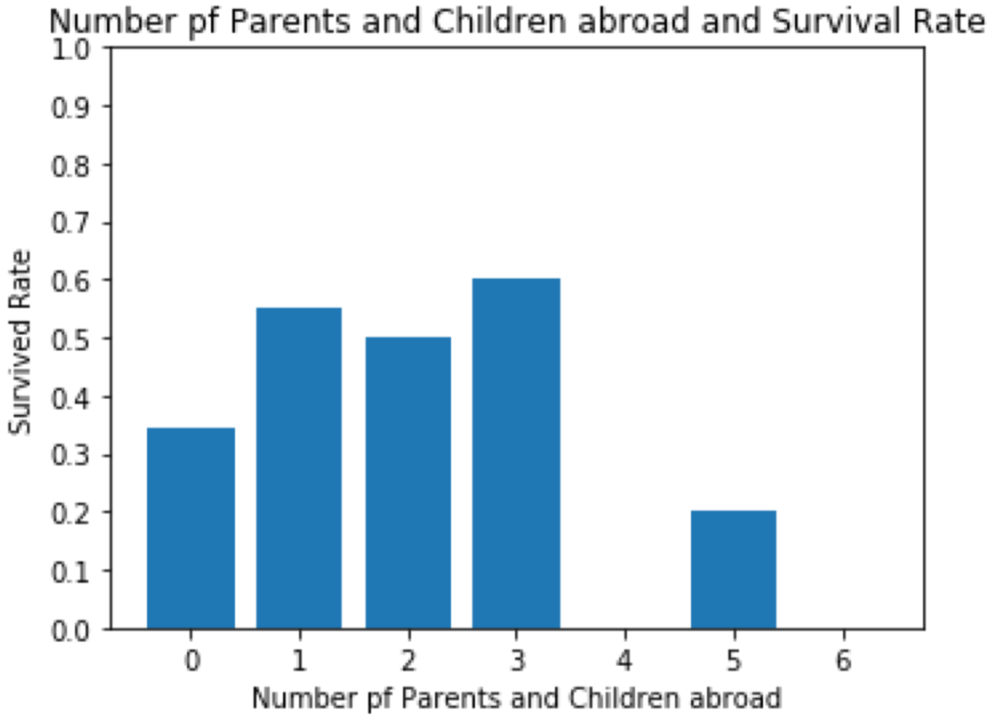
```
fig = plt.figure(figsize=(12, 4))
ax1 = fig.add_subplot(121)
ax2 = fig.add_subplot(122)

# Sibsp (同乗中の兄弟/配偶者の数)
SiblingPlot = train_set['Survived'].groupby(train_set['SibSp']).mean()
ax1.bar(x=SiblingPlot.index, height=SiblingPlot.values)
ax1.set_ylabel('Survived Rate')
ax1.set_xlabel('Total Siblings')
ax1.set_xticks(SiblingPlot.index)
ax1.set_yticks(np.arange(0, 1.1, .1))
ax1.set_title("Total Siblings and Survival Rate")
```

同乗中の兄弟/配偶者の数と同乗中の親/子供の数

```
# Parch (同乗中の親/子供の数)
ParchPlot = train_set['Survived'].groupby(train_set['Parch']).mean()
ax2.bar(x=ParchPlot.index, height=ParchPlot.values)
ax2.set_ylabel('Survived Rate')
ax2.set_xlabel('Number pf Parents and Children abroad')
ax2.set_xticks(ParchPlot.index)
ax2.set_yticks(np.arange(0, 1.1, .1))
ax2.set_title("Number pf Parents and Children abroad and Survival Rate")
```

実行結果





兄弟・配偶者が1,2人が生存率が高い.
子供・親は1,2,3人のときが生存率が高い.