In [ ]:

```
【ハイパーパラメタ・チューニング】
問2
① タイタニックの'train.csv'を読み込み、予測精度に影響が少ない特徴量を
   外し、データを学習用:テスト用=7:3に分割し、SVMの学習モデルを作成し、
   予測精度を求めてください。(C,gammaの値はデフォルトで)
```

In [21]:

```
# ①
import pandas as pd
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns

df = pd.read_csv('train.csv')
df.head()
```

Out[21]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Far
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.250
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.283
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.925
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.100
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.050

In [22]:

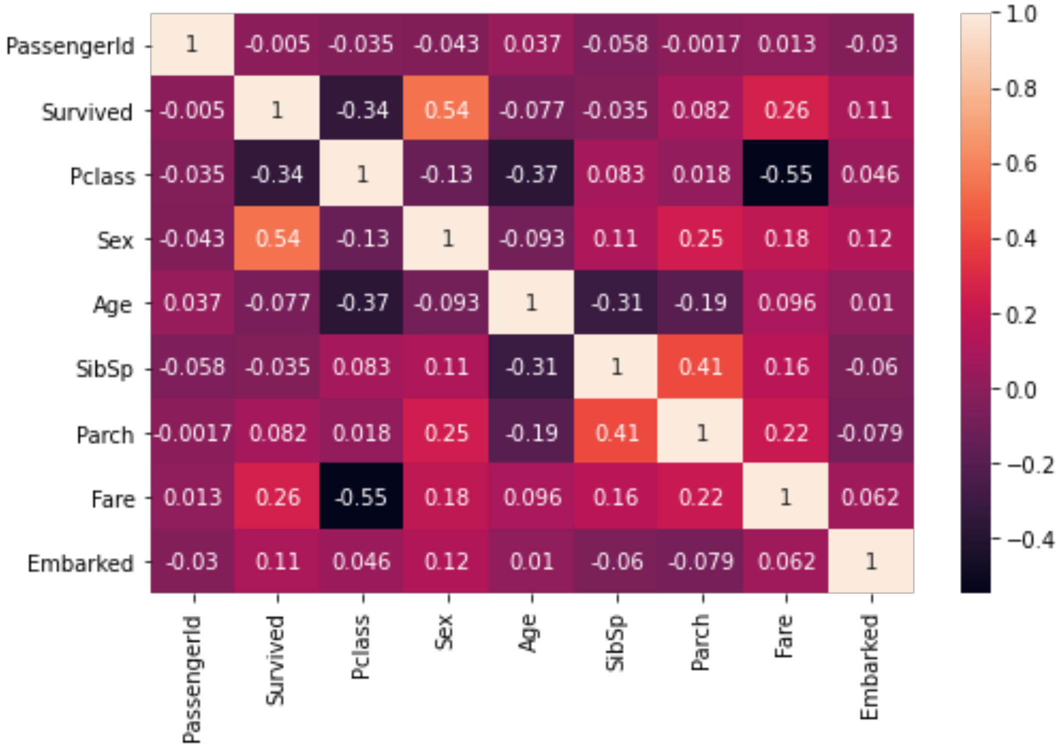
```
# カテゴリカル変数の'Sex'と'Embarked'のラベルを数値化します
df['Sex'] = df['Sex'].apply(lambda x: 0 if x=='male' else 1)
df['Embarked'] = df['Embarked'].fillna('S')
df['Embarked'] = df['Embarked'].map({'S': 0, 'C': 1, 'Q': 2})
df.head()
```

Out[22]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare
0	1	0	3	Braund, Mr. Owen Harris	0	22.0	1	0	A/5 21171	7.2500
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	1	38.0	1	0	PC 17599	71.2833
2	3	1	3	Heikkinen, Miss. Laina	1	26.0	0	0	STON/O2. 3101282	7.9250
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	1	35.0	1	0	113803	53.1000
4	5	0	3	Allen, Mr. William Henry	0	35.0	0	0	373450	8.0500

```
In [23]: plt.figure(figsize=(8,5))
         sns.heatmap(df.corr(), annot=True)

Out[23]: <matplotlib.axes._subplots.AxesSubplot at 0x10fe98518>
```



```
In [24]: # 'Survvied'との相関が低い'PassengerId','Age','SibSp','Parch'と
         # 欠損値が多い'Ticket','Cabin'、明らかにユニークな'Name'をデータフレ
         # ームから落とします
         df = df.drop(['PassengerId','Age','SibSp','Parch','Ticket',
                       'Cabin','Name'], axis=1)
         df.head()
```

Out[24]:

	Survived	Pclass	Sex	Fare	Embarked
0	0	3	0	7.2500	0
1	1	1	1	71.2833	1
2	1	3	1	7.9250	0
3	1	1	1	53.1000	0
4	0	3	0	8.0500	0

In [25]:

```
from sklearn.svm import SVC
from sklearn.metrics import accuracy_score
from sklearn.model_selection import train_test_split

x = df.drop('Survived', axis=1)
y = df['Survived']
x_train, x_test, y_train, y_test = train_test_split(x, y,
                                                    test_size=0.3,
                                                    random_state=0)


clf = SVC()
clf.fit(x_train, y_train)
```

Out[25]:

```
SVC()
```

In [26]:

```
pred = clf.predict(x_test)
acc = accuracy_score(pred, y_test)
print('accuracy : {:.5f}' .format(acc))
```

```
accuracy : 0.70522
```

In [27]:

```
# ちなみに、↑の計算で使用されたハイパーパラメーターの値は次のように求められます。
clf.get_params()
```

Out[27]:

```
{'C': 1.0,
 'break_ties': False,
 'cache_size': 200,
 'class_weight': None,
 'coef0': 0.0,
 'decision_function_shape': 'ovr',
 'degree': 3,
 'gamma': 'scale',
 'kernel': 'rbf',
 'max_iter': -1,
 'probability': False,
 'random_state': None,
 'shrinking': True,
 'tol': 0.001,
 'verbose': False}
```