Part3 機械学習プロセスの王道を学ぶ

ステップ4 データの前処理

16T4056S 溝口賢治

2019/4/12

- 欠損値の処理

　データに含まれる欠損値を確認して、欠損がある箇所へ適切な値を代入して補う。

- カテゴリデータの処理

　データセットに含まれる値が文字列の場合、機械学習のアルゴリズムは学習が適切に行えないため、これらの特徴量の値をアルゴリズムが学習しやすい適切な形に処理する。

・欠損値を捜索する関数　isnull()

　データフレームの各列ごとに欠損値を探して、sum()で合計した値を出力する

```
train_set.isnull().sum()
```

・欠損値をデータから探して、指定した値を代入する関数　fillna()

リスト６　欠損値を探して指定した値を代入するコード

```
#欠損値へ適切な値を代入
train_set['Age'].fillna(train_set['Age'].median(), inplace=True)
train_set['Embarked'].fillna('C',inplace=True)
```

カテゴリデータの処理

・乗客乗員名簿にある最初の５行（５人）のSex(性別)を表示

```
train_set['Sex'].head()
```

リスト７　Sex（性別）の処理をするコード

```
#sex（性別）の値を処理
labelencoder=LabelEncoder()
train_set[‘Sex’] = labelencoder.fit_transform(train_set[‘Sex’])

#処理後の最初の５行を表示
train_set[‘Sex].head()
```

・ダミー変数

　カテゴリデータのようにもともと数値ではないデータを、０と１を使って表現した変数

　ダミー変数では元のカテゴリがN種類ある場合、N-1個のダミー変数で対応することができる

リスト８　Embarked（出発港）とPclass（旅客等級）をダミー変数へ変換するコード

```
#Embarked（出発港）をダミー変数へ変換
Embarked = pd.get_dummies(train_set[‘Embarked’], drop_first=True)
Embarked.columns = [‘Embarked-Q’, ‘Embarked-S’]

#Pclass（旅客等級）をダミー変数へ変換
Pclass = pd.get_dummies(train_set[‘Pclass’], drop_first=True)
Pclass.columns = [‘PClass2’, ‘PClass3’]

#Embarkedの確認
Embarked.head()
```