クレジット:
UTokyo Online Educationデータマイニング入門 2018 森　純一郎

東京大学
THE UNIVERSITY OF TOKYO

CC BY NC ND

課題10

In [ ]:
```
import pandas as pd
import numpy as  np
import matplotlib.pyplot  as plt
%matplotlib inline
import seaborn as sns
from sklearn.preprocessing import StandardScaler
from sklearn.preprocessing import PolynomialFeatures
from sklearn.linear_model import LinearRegression
from sklearn.linear_model import Ridge
from sklearn.pipeline import make_pipeline
from sklearn.model_selection import train_test_split
from sklearn.model_selection import cross_val_score
from sklearn.model_selection import StratifiedKFold
from sklearn.model_selection import GridSearchCV
from sklearn.model_selection import learning_curve
from sklearn.model_selection import validation_curve
from sklearn.metrics import mean_squared_error
```

Question

データ分析の実践のノートブックで使用したデータセットについて、特徴量の選択、特徴量の変換や組み合わせによる新たな特徴量の生成、などを行い、それらの特徴量を用いたモデルの学習と評価を行ってください。モデルは `LinearRegression` を使ってください。

どのような着想の元で特徴量の選択や作成を行ったのかについても報告してください。各自の報告を次回の授業内で紹介することがあるかもしれませんが、憚られる方はその旨も記載しておいてください。

以下では例として、 `alcohol` と `sulphates` の特徴量に加えて、それらの特徴量を掛け合わせた交差項となる特徴量を新たに作成し、これらの特徴量を用いてモデルの学習と評価を行っています。

In [ ]:
```
wine = pd.read_csv("winequality-red.csv", sep=";")
```

In [ ]:
```
X=wine[ ['sulphates', 'alcohol']].values
y=wine[['quality']].values

# 標準化と線形回帰モデルのパイプライン
pipe=make_pipeline(StandardScaler(), LinearRegression())

# 交差検証
scores = cross_val_score(pipe, X, y,
                         scoring='neg_mean_squared_error', cv=10)
print( -scores.mean())
```

In [ ]:
```
X=wine[ ['sulphates', 'alcohol']].values
y=wine[['quality']].values

# 'sulphates'と'alcohol'の交差項となる特徴量
new_feature=X[:,0]*X[:,1]

#  特徴量の追加
X=np.insert(X, 0, new_feature, axis=1)

# 標準化と線形回帰モデルのパイプライン
pipe=make_pipeline(StandardScaler(), LinearRegression())

# 交差検証
scores = cross_val_score(pipe, X, y,
                         scoring='neg_mean_squared_error', cv=10)
print( -scores.mean())
```

In [ ]:
```
### このセルにコードを記入してください ###
```

このマークダウンセルに説明を記入してください（セルをダブルクリックすると編集できます）