機械学習とは

1　機械学習の基礎

1.1　人工知能と機械学習

1.1.1　人工知能

人工的に知能を模倣しようという試み全般を広く指す必ずしも学習を伴わない機構を含む。

- 最適化
- 記号推論

  ex. 第 5 世代知識を論理で書き下しておいてそれを用いて推論

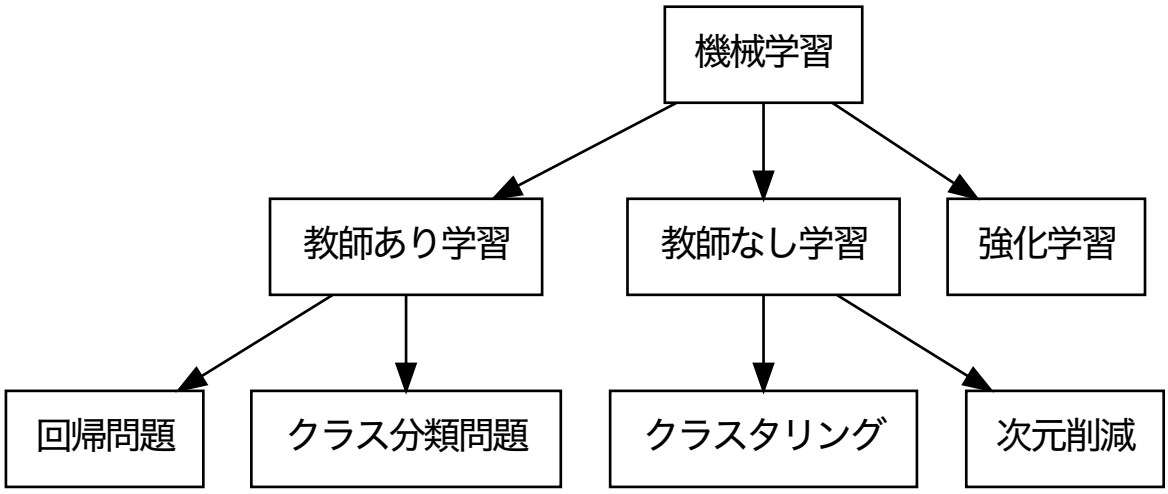
1.1.2　機械学習

「経験からの学習により自動で改善するコンピューターアルゴリズム」(wikipedia)

1.1.3　機械学習の分類

- 教師あり学習 - supervised learning
- 教師なし学習 - unsupervised learning
- 強化学習 - reinforcement learning

[1]:



1.2 機械学習用語

- 訓連セット、テストセット
- 特徴量
- ハイパーパラメータ
- クラス

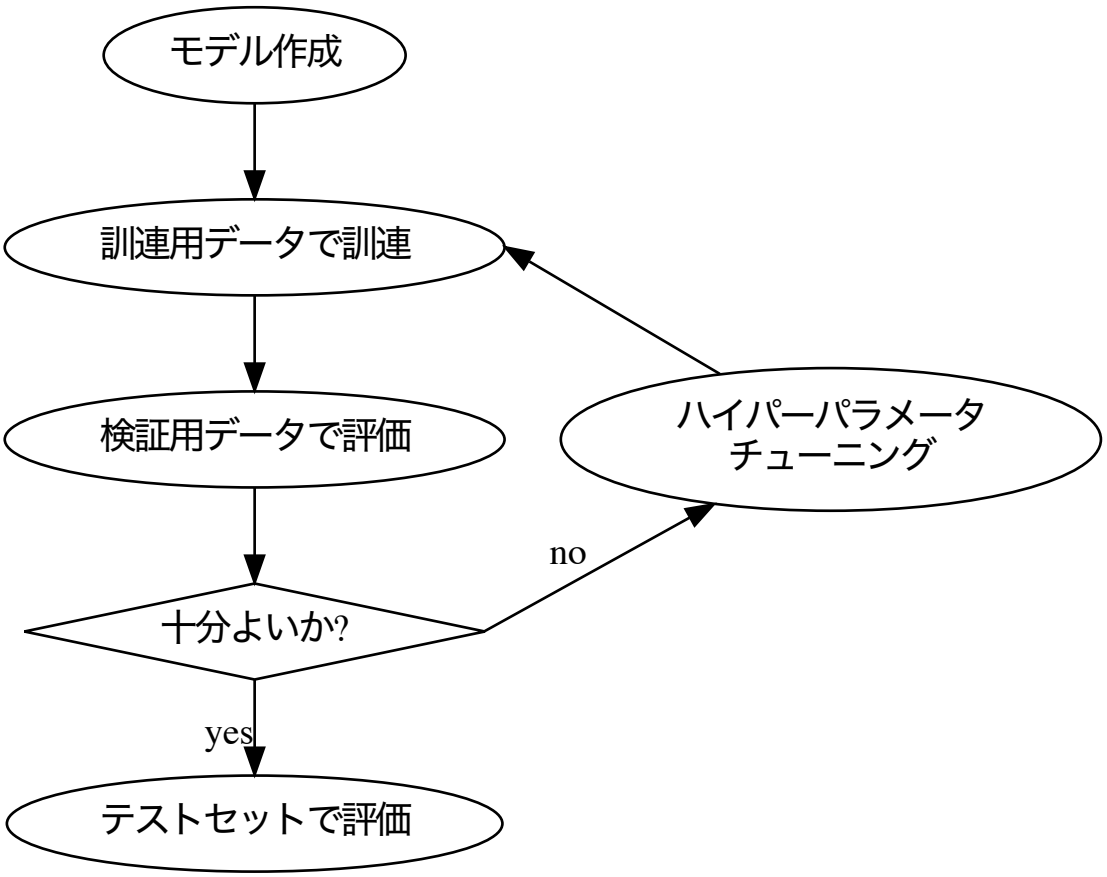
1.3 機械学習のフロー

- 訓連セットとテストセットに分割する
- モデルを構築し、チューニングする
  - 訓連セットを訓連用データと検証用データに分割
  - 訓連用データで訓連したモデルを、検証用データで評価
  - 検証用データでの評価結果が良くなるように、モデルのハイパーパラメータを更新
- テストセットでモデルを評価

1.3.1 注意点

- テストセットは最後のモデル評価以外では使ってはいけない
- テストセットを使ってハイパーパラメータチューニングするのは**絶対に**禁止
  - しりたいのは未見のデータに対する性能。テストセットは未見データの代用
  - テストセットに対してチューニングすると、未見ではなくなるので意味がない

[67]:



1.4 データの前処理

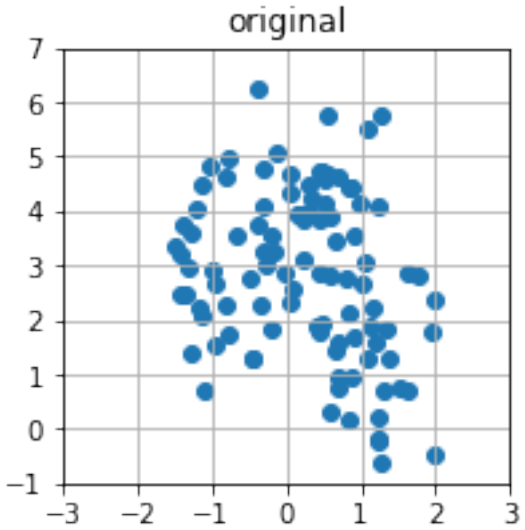
- 特徴量によって値のレンジが大きく異る場合
  - 特徴量 A が 0-1、特徴量 B が 0-100 の値を取るようなケース
- そのまま扱うと値のレンジの大きい特徴量が支配的になり、レンジの小さい特徴量が十分に反映されない
- レンジがだいたい同じになるようにスケール変換する必要がある

1.5 注意点

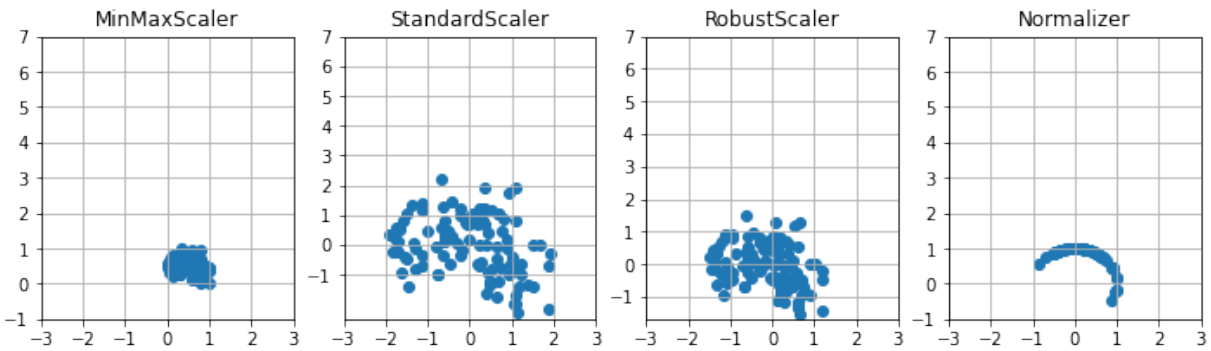
- スケール変換が必要ないアルゴリズムもある
  - 決定木ベースの手法
  - 適当な値で各特徴量空間を分割していくだけなので、レンジが異なっても全く影響を受けない
- スケール変換は分割後の訓連データに対して学習を行い、同じ変換手法を用いてテスト/検証データを変換する
  - スケール変換はある種の教師なし学習

1.6 いくつかの前処理手法

- MinMaxScalar - 最大値を 1、最小値を 0 になるように変換
- StandardScalar - 平均 0、分散 1 になるように変換
- RobustScalar - 中央値と四分位数を用いて変換。外れ値を無視する
- Normalizer - 原点からの方向だけを維持して、超球面上に投射



[115]:
```
trans = [MinMaxScaler(), StandardScaler(), RobustScaler(), Normalizer()]
_, axes = plt.subplots(1, 4, figsize=(12,3))
for i, t in enumerate(trans):
    X_ = t.fit_transform(X)
    plot(axes[i], X_, t.__class__.__name__);
```


MinMaxScaler
StandardScaler
RobustScaler
Normalizer