機械学習
2-6

工学院情報通信系
中原 啓貴

今日の内容

- 教師なし学習
  - k-means法
- データの取捨取得について
  - (Down|Up)サンプリング
  - Bug of Feature (BoF)
  - ブートストラップ
- k-meansを用いた応用事例
  - 異常値検出
  - 画像圧縮

＋上記内容の演習

k-means法

- 代表的なクラスタリング法
  - 教師なし学習の分類問題でもある
- 各クラスタの散らばり具合が最小になるようなクラスタラベル $\{y_i | y_i \in \{1, \ldots, c\}\}_{i=1}^{n}$ を標本 $\{x_i\}_{i=1}^{n}$ に割り当てる問題
- 散らばり具合

$$\sum_{i:y_i=y} \|x_i - \mu_y\|^2,$$

ここで，$\sum_{i:y_i=y}$ は $y_i = y$ を満たす i に関する和であり，

$$\mu_y = \frac{1}{n_y} \sum_{i:y_i=y} \|x_i - \mu_y\|^2,$$

はクラスタ y の中心，n_y はクラスタ y に属する標本数

k-means法 (続)

- すべてのクラスタ$y = 1, \ldots, c$に関して足し合わせたものを
  最小にするクラスタラベルを決定する

$$\sum_{y=1}^{c} \sum_{i:y_i=y} \|x_i - \mu_y\|^2$$

- 最適化の計算時間は標本数nに関して指数関数的に増加
  →厳密解を得るのは困難
- 局所最適解を求めるアルゴリズムが使われている

k-means法のアルゴリズム

- 標本を１つずつ一番近いクラスタに割り当てる操作を繰返す

1. クラスタ中心$\boldsymbol{\mu_1, \ldots, \mu_c}$を適当（通常は乱数）に決める
2. 標本$\boldsymbol{x_1, \ldots, x_n}$に対するクラスタラベルを下記の式に従って更新

$$y_i \leftarrow \underset{y \in \{1,\ldots,c\}}{\operatorname{argmin}} \left\| x_i - \mu_y \right\|^2, i = 1, \ldots, n$$

3. クラスタ中心$\boldsymbol{\mu_1, \ldots, \mu_c}$を更新

$$\mu_y \leftarrow \frac{1}{n_y} \sum_{i:y_i=y} \left\| x_i - \mu_y \right\|^2, y = 1, \ldots, c$$

4. クラスタラベルが収束するまで 2.～3.を繰り返す

カーネルk-means法

- k-means法はユークリッド距離でクラスタを決定
- 線形分離可能なクラスタのみ
- SVMと同様にカーネルトリックを適用

  →カーネルk-means法
  - 非線形なクラスタ分離が可能に

カーネルk-means法(続)

- ユークリッド距離の二乗$\left\|x_i - \mu_y\right\|^2$を標本同士の内積$\langle x, x' \rangle$を使って表現

$$\left\|x_i - \mu_y\right\|^2 = \left\|x - \sum_{i:y_i=y} x_i\right\|^2$$

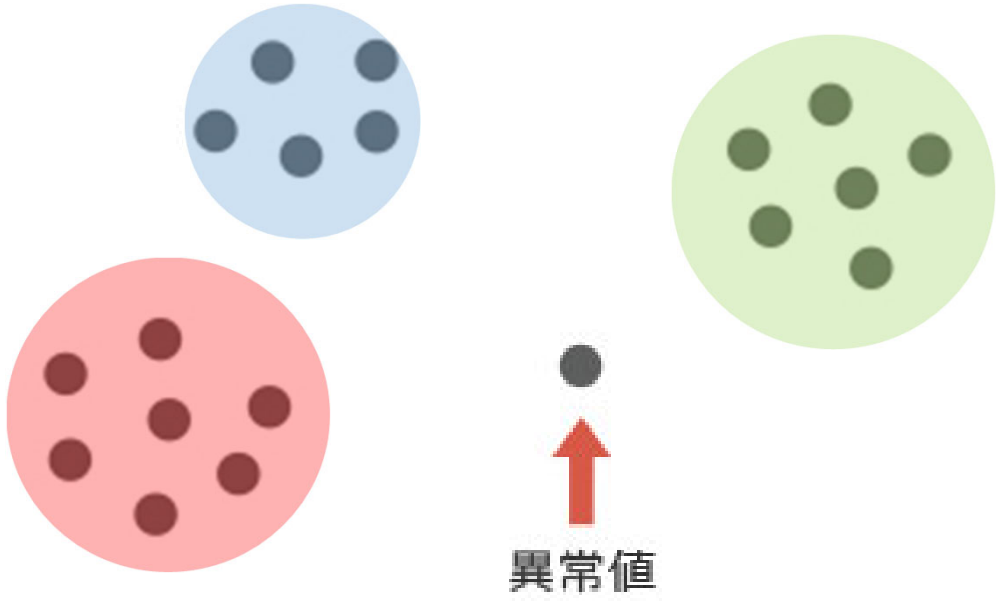
$$= \langle \boldsymbol{x}, \boldsymbol{x} \rangle - \frac{2}{n_y} \sum_{i:y_i=y} \langle \boldsymbol{x}, \boldsymbol{x_i} \rangle + \frac{1}{n_y^2} \sum_{i,i':y_i=y_{i'}=y} \langle x_i, \boldsymbol{x_i}' \rangle$$

- これらの内積をカーネル関数$\mathbf{K}(\boldsymbol{x}, \boldsymbol{x}')$に置換

$$y_i \leftarrow \underset{y \in \{1,\ldots,c\}}{\operatorname{argmin}} \left[- \frac{2}{n_y} \sum_{i:y_i=y} \mathbf{K}(\boldsymbol{x}, \boldsymbol{x_i}) + \frac{1}{n_y^2} \sum_{i,i':y_i=y_{i'}=y} \mathbf{K}(\boldsymbol{x_i}, \boldsymbol{x_{i'}}) \right]$$

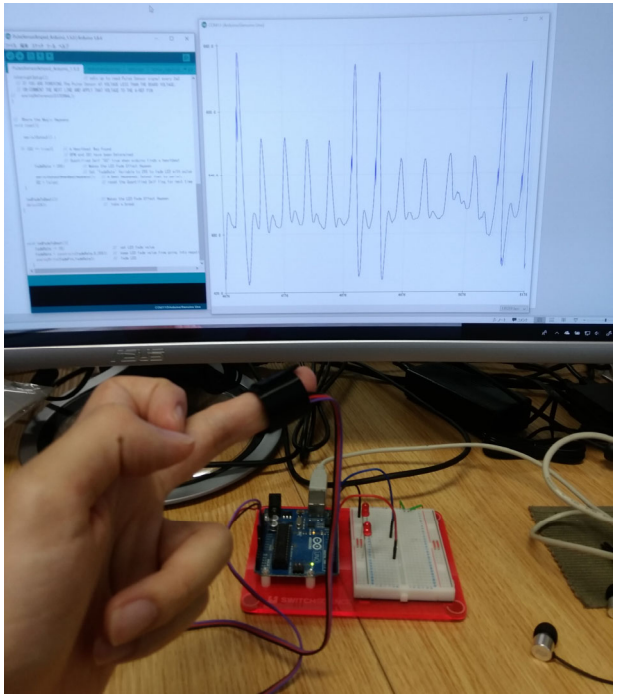
外れ値検出

- 大多数のデータとは振る舞いが異なるデータを検出
  - クレジットカードの不正使用検出
  - システムの故障予知
  - 異常行動検知
- 正常データのみで学習(k-means法など)しておき,
  クラスタから離れたデータを異常とみなす



演習：k-meansによる外れ値検出

- 時系列データの外れ値検知
  - 統計データに基づく手法→静的データ
    （枠毎に適用すれば時系列も使える）
    - ホテリング理論：データが十分に多いとき、自由度１のカイ２乗分布に従うことが証明済み



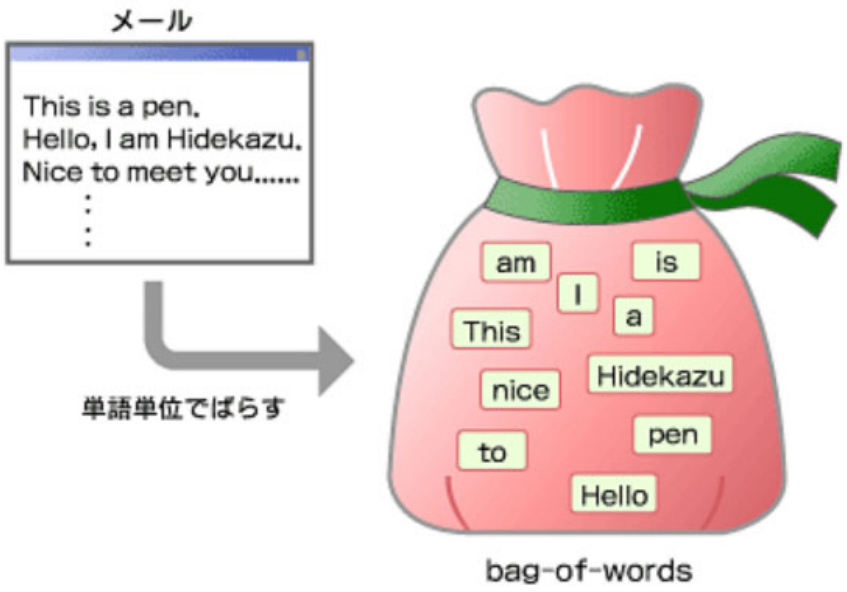
演習：画像圧縮

(中井著, “ITエンジニアのための機械学習入門: 6章”を一部修正)

- 各色毎にクラスタに分類
- 中心画素色にクラスタリング

データの取捨取得について

- ダウンサンプリング・アップサンプリング
  - データ数の偏りを調整する手法
  - 通常はダウンサンプリングが使われる
- Bag-of-Words (BoW)
  - Bag-of-Visual Words, Bag-of-Features

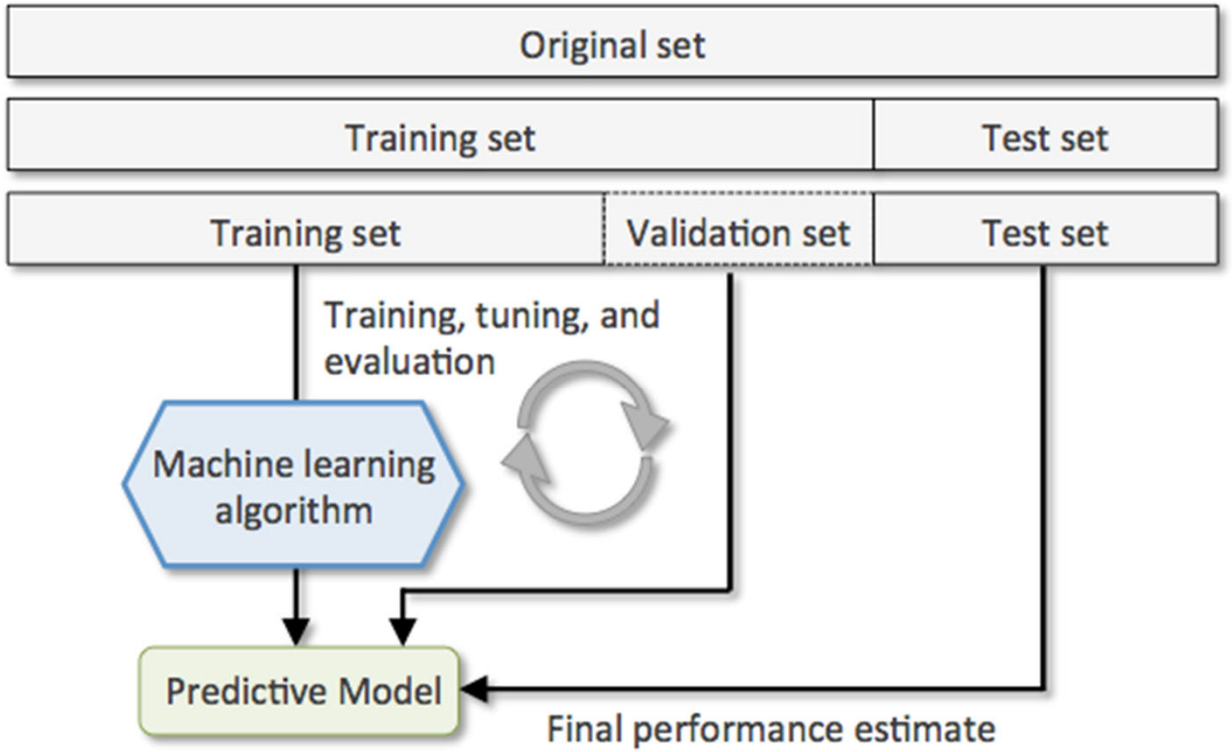


学習後の評価について

- 汎化性能を担保するため未学習データによる評価
- 基本的にはデータを訓練用とテスト用に分割
- ハイパーパラメータの設定やモデルを選択
- ホールドアウト法(holdout method)
- k分割交差検証(k-fold cross-validation)

ホールドアウト法

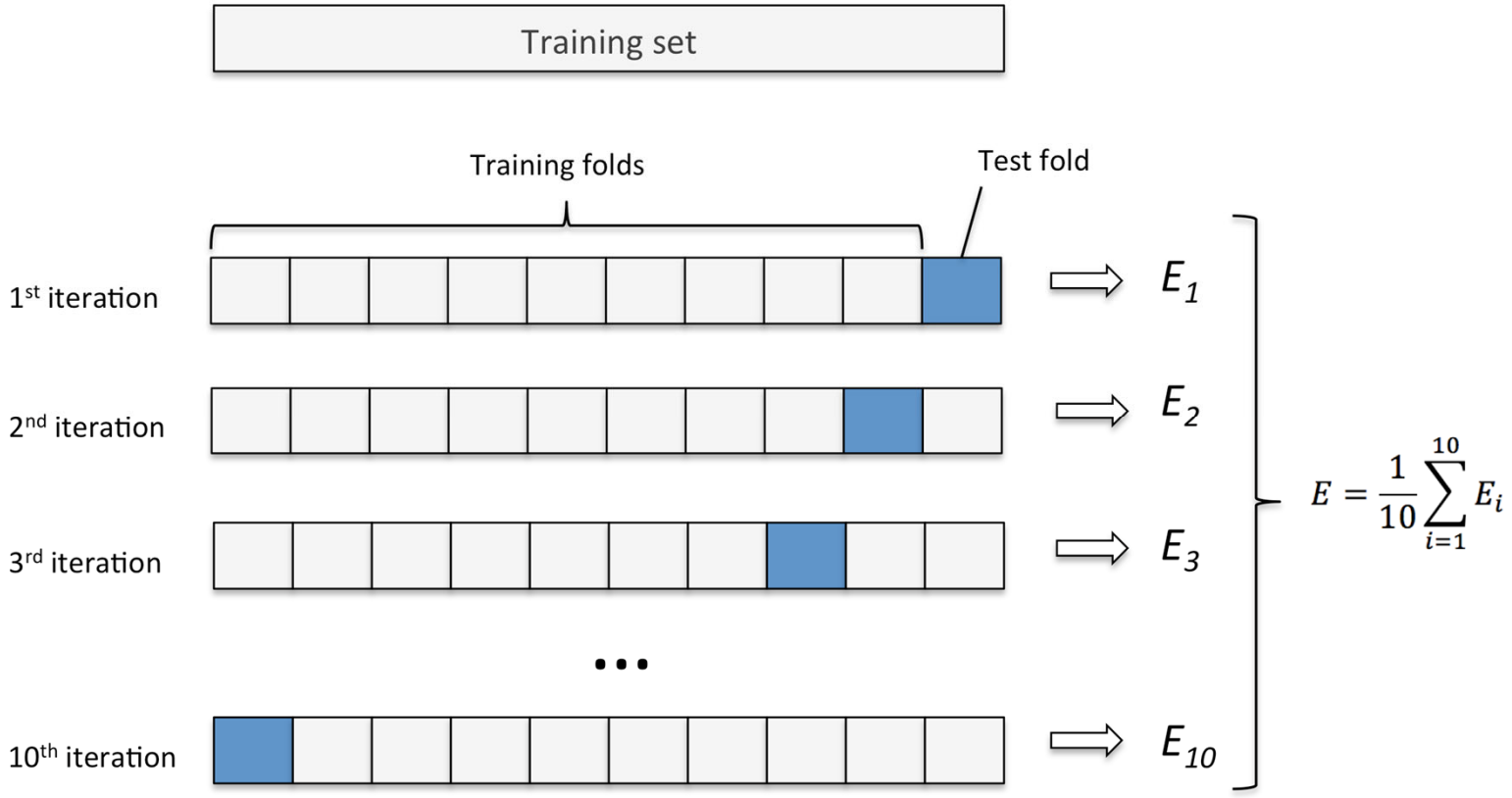
- モデル選択時に同じデータを利用→過学習
- 検証(Validation)データをさらに用意
  - モデル選択（評価）に利用
- 最後にテストデータで汎化性能を評価



Source: S. Raschka, “Python機械学習プログラミング: 6章”, pp.167

k分割交差検証

- ホールドアウト法の欠点はデータの分割に依存すること
- データをランダムにk個に分割し, k-1個を学習, 1個をテストに使用 (c.f. 非復元抽出)



Source: S. Raschka, “Python機械学習プログラミング: 6章”, pp.169

演習: k分割交差検証

- scikit-learn のクラスを利用
- パイプライン処理の方法を紹介
  - マスターすると便利な機能
- グリッドサーチによるパラメータ探索

まとめ

- 教師なし学習の１つであるk-means法
  - 外れ値検出, 画像圧縮
- データの取捨選択について
- 学習後の評価方法について
- 様々なTips
  - パイプライン処理
  - グリッドサーチ
  - (Python3系への対応…)