(83342) パターン認識特論

張　潮
zhang@u-fukui.ac.jp
0776-27-8477

本日の内容

第1回　パターン認識のための線形代数
第2回　パターン認識のための確率論
第3回　凸最適化の概念
第4回　線形回帰とLMSアルゴリズム
第5回　分類とロジスティック回帰
第6回　一般化線形モデル
第7回　ガウシアン判別分析と単純ベイズ分類器
第8回　サポートベクターマシン（1）
第9回　サポートベクターマシン（2）
第10回　正則化とモデル選択（ハイパーパラメータのチューニング）
第11回　k平均法
第12回　混合ガウスモデルとEMアルゴリズム
第13回　因子分析
第14回　主成分分析

第15回　未定

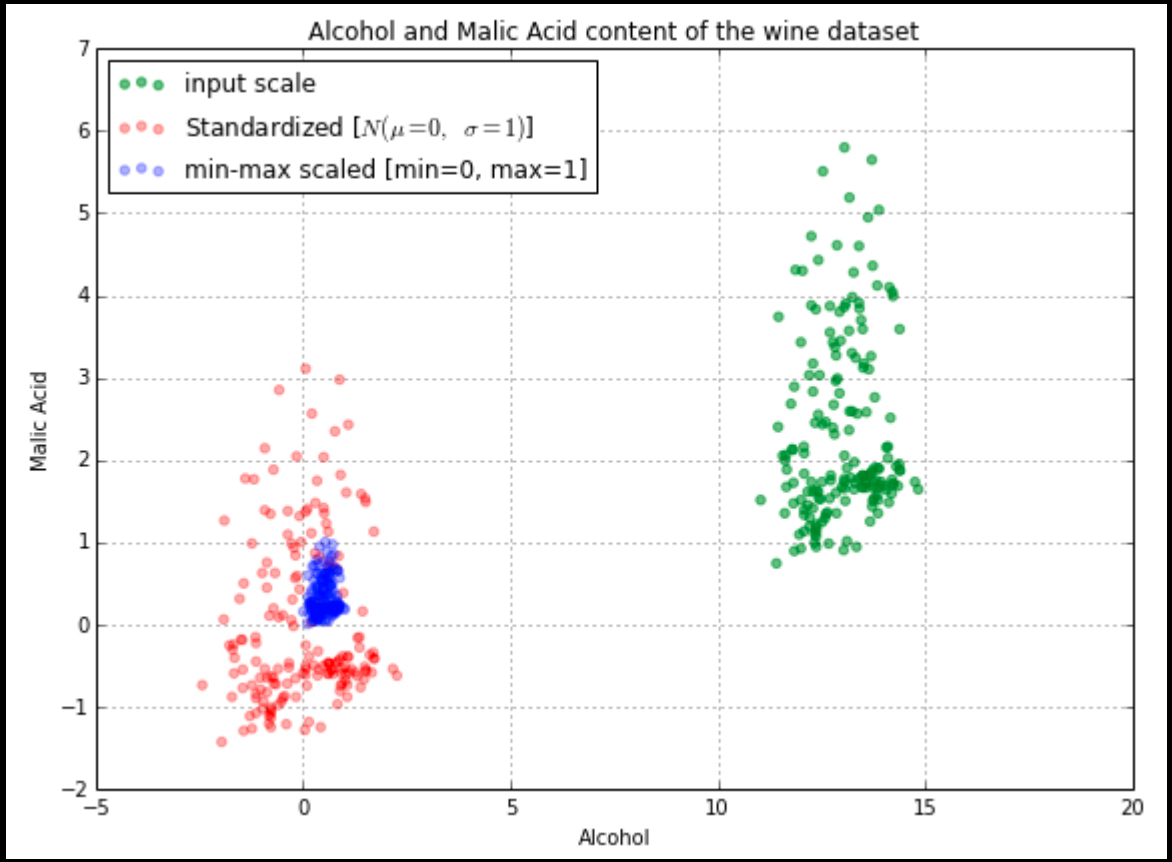
clarification

- Normalization or Min-Max scaling（正規化）
  - １次元：範囲を[0, 1]にするスケーリング処理
  - 学習データ：各属性（特徴）の範囲を[0, 1]にする
- Standardization（標準化）
  - １次元：平均が0かつ分散が1となるように入力データのシフティングとスケーリングを行う
  - 学習データ：各属性（特徴）の平均を0かつ分散を1にする
- Regularization（正則化）

データの前処理ではない

Feature scaling (normalization or standardization)

	Class label	Alcohol	Malic acid
0	1	14.23	1.71
1	1	13.20	1.78
2	1	13.16	2.36
3	1	14.37	1.95
4	1	13.24	2.59



“Standardization or normalization?”

- アプリケーションによる
- 例
  - PCA →standardization?
  - Neural network→normalization?

Regularization（正則化）

- ざっくり言うと．．．
- データの前処理ではない
- 最適化問題の目的関数において，追加されたペナルティ（罰則）項のイメージが強い
  - 目的関数を最小化することでモデルを最適化
  - 学習データにノイズなどイレギュラーなものがあるときに、モデルの最適化に使うと同時に、ペナルティを与える→過学習を防ぐ（ノイズまで学習してしまうとモデルが細かすぎる．．．）
  - 罰則項の例
    - モデルの複雑さ
    - 滑らかさ
    - ノルムの大きさ

Regularization（正則化）

- ペナルティ項にハイパーパラメータ（hyper parameter）がある

学習を行う際に人間が予め設定し
ておかなければいけないパラメータ

重み付き線形回帰の場合：

$$\text{minimize} \sum_i w^{(i)}(y^{(i)} - \theta^T x^{(i)})^2$$

$$w^{(i)} = \exp\left(-\frac{(x^{(i)} - x)^2}{2\tau^2}\right)$$

帯域幅 (bandwidth)

SVMの場合：

$$\begin{aligned} \min_{\gamma,w,b} \quad & \frac{1}{2}||w||^2 + C\sum_{i=1}^{m}\xi_i \\ \text{s.t.} \quad & y^{(i)}(w^T x^{(i)} + b) \geq 1 - \xi_i, \quad i = 1, \ldots, m \\ & \xi_i \geq 0, \quad i = 1, \ldots, m. \end{aligned}$$

hyper parameterの決め方

- 良いhyper parameterの選び方
- 多項式回帰

$$h_\theta(x) = g(\theta_0 + \theta_1 x + \theta_2 x^2 + \cdots + \theta_k x^k)$$

  - *k*=?
  - *k*が大きくなるとモデルが細かくなる

$$\mathcal{M} = \{M_1, \ldots, M_d\}$$

モデル選択

hyper parameterの決め方

- 学習データセットS
- 自然に考えられる方法

1. Sに対してそれぞれのモデルM_iで学習してみる
2. 学習誤差（目的関数値）が一番小さくなるM_iを選ぶ

残念ながらこの方法ではうまくいきません☹

hyper parameterの決め方

- M_iが細かくなるほど、よりSにフィッティングできる

学習誤差がより小さくなる

- 大きいkを選んでしまう☹
- ホールドアウト交差検証（hold-out cross validation）

hyper parameterの決め方

- ホールドアウト交差検証（hold-out cross validation）

1. ランダムにSを2つのセットに分ける（70%のデータ→S_{train}，30%のデータ→ S_{cv} ）
   S_{train} : 学習用セット
   S_{cv} : テスト用セット（ホールドアウト交差セット）

2. 学習で目的関数値を最小化することで各M_iを最適化
3. S_{cv}を使って各M_iをテストする
4. テスト誤差が一番小さいなM_iを選んで、全データを使った学習を行う

S_{cv} で、より正確な汎化誤差（generalization error）を最小化するようなM_iを選ぶ

※学習に使った訓練集合をテスト集合にして求めた誤差→カンニング→過小評価してしまう

hyper parameterの決め方

- 汎化誤差
  - サンプルの母集団に対する誤差
  - 期待損失 (expected risk) ともいう
  - 学習に使った訓練集合は母集団ではなく，標本データである
    - 標本データ：集めたデータ
    - 母集団：本来存在するすべてのデータ
  - 学習データをテストデータとして計算した誤差は汎化誤差より一般に過小評価になる

学習理論といえば汎化誤差の理論

hyper parameterの決め方

- ホールドアウト交差検証の欠点
  - 学習に使うデータを30%“浪費”してしまった
  - 後から全データを使って学習し直すことができるが、70%データを使って選んだhyper parameterがそもそも悪い可能性がある

***K*-分割交差検証（k-fold cross validation）**

hyper parameterの決め方

- **K-分割交差検証（k-fold cross validation）**

（1）SをK個に分割する（$S_1, S_2, \ldots, S_k$）.

（2）for j=1,2,…k

S_jをテストセットとし，残る $K - 1$ 個を学習セットとする．すべてのM_iに対して学習を行う.

（3）（2）で得られたj個のテスト誤差の平均を最小化するM_iを選ぶ

よく使うのが：K=10

hyper parameterの決め方

- *K*=*m*(サンプル総数)

leave-one-out 交差検証(leave-one-out cross validation)
(mが小さいときによく使われる)

特徴選択（Feature Selection）

- 特徴量（属性）の次元数n>>サンプル数m
- そのまま学習してしまうと過学習が起こりやすい
- 一部の次元しか学習タスクに影響する→不要で冗長な特徴量を除去したい

- n次元の特徴量の場合，2^n個のsubsetが存在する

nが大きいとき，全探索がexpensive

ヒューリスティックの手法がよく用いられる

探索アプローチの例
- 総当たり（×）
- 最良優先探索
- 焼きなまし法
- 遺伝的アルゴリズム
- 貪欲前向き選択
- 貪欲後ろ向き選択

特徴選択（Feature Selection）

- 貪欲前向き選択（forward search）

1. Initialize $\mathcal{F} = \emptyset$.
2. Repeat {

   繰り返し回数もしくは閾値で制御できる

   (a) For $i = 1, \ldots, n$ if $i \notin \mathcal{F}$, let $\mathcal{F}_i = \mathcal{F} \cup \{i\}$, and use some version of cross validation to evaluate features $\mathcal{F}_i$. (I.e., train your learning algorithm using only the features in $\mathcal{F}_i$, and estimate its generalization error.)

   (b) Set $\mathcal{F}$ to be the best feature subset found on step (a).

   }

3. Select and output the best feature subset that was evaluated during the entire search procedure.

特徴選択（Feature Selection）

- Wrapper model feature selectionの欠点

1. Initialize $\mathcal{F} = \emptyset$.
2. Repeat {
   (a) For $i = 1, \ldots, n$ if $i \notin \mathcal{F}$, let $\mathcal{F}_i = \mathcal{F} \cup \{i\}$, and use some version of cross validation to evaluate features $\mathcal{F}_i$. (I.e., train your learning algorithm using only the features in $\mathcal{F}_i$, and estimate its generalization error.)
   (b) Set $\mathcal{F}$ to be the best feature subset found on step (a).
   }
3. Select and output the best feature subset that was evaluated during the entire search procedure.

必要な学習回数が多い．最大n^2回が必要．

特徴選択（Feature Selection）

- Filter feature selection
- 何かしらの評価関数*f*(i)を作って，第*i*次元の特徴がどのくらい情報量あるかをランク付ける
- 上位k個の特徴を選ぶ
  - kも交差検証で決めることが可能
- 相互情報量（mutual information）
  - 2つの確率変数の相互依存の尺度を表す量
  - ある特徴x_iとラベルの依存尺度を計算

$$\mathrm{MI}(x_i, y) = \sum_{x_i \in \{0,1\}} \sum_{y \in \{0,1\}} p(x_i, y) \log \frac{p(x_i, y)}{p(x_i)p(y)}.$$

特徴選択（Feature Selection）

- **カルバック・ライブラー情報量**
  - *Kullback–Leibler divergence*
  - 確率論と情報理論における2つの確率分布の差異を計る尺度
  - 考え方：2つの確率変数が独立しているなら関連性が低い

$$D_{\mathrm{KL}}(P\|Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}$$

$$p(x_i, y) = p(x_i)p(y)$$

$$\mathrm{MI}(x_i, y) = \sum_{x_i \in \{0,1\}} \sum_{y \in \{0,1\}} p(x_i, y) \log \frac{p(x_i, y)}{p(x_i)p(y)}.$$

相互情報量

$$\mathrm{MI}(x_i, y) = \mathrm{KL}\left(p(x_i, y) || p(x_i)p(y)\right)$$

カルバック・ライブラー情報量

余談

- 統計学の世界

ベイズ主義 **VS.** 頻度主義

ベイズの定理

推計統計学

余談

逆確率の理論はある誤謬の上に立脚するのであって，完全に葬り去らなければならない

ベイズの定理を批判

https://www.slideshare.net/KojiKosugi/ss-50740386

余談

- 現状
  - ベイズの定理が広く実用化されているため
  - naive Bayes classifier （スパムメール検出，天気予報．．．）

余談

仮説検定，帰無仮説，対立仮説，統計量の算出，両側検定，片側検定，第1種の誤り，第2種の誤り，t検定，p値，…（品質管理やアンケート分析によく使われる）

	頻度主義	ベイズ主義
母数θ	定数	確率変数
データx,y	確率変数	定数

頻度主義では，たった一つの真値を求めて慎重に議論する

→仮説は真か偽のどちらかである

ベイズ主義では，データから考えられる母数の分布を考える

→データから何がどの程度言えるのかを主張する

https://www.slideshare.net/KojiKosugi/ss-50740386

余談

母数の推定

- 人の身長のデータは正規分布に従うと思われる。この度，20人分の身長を測定したところ次のようになった。

```
 [1] 165.42 164.28 168.10 161.43 156.18 169.60 183.77 181.46 143.44 171.87 177.61 167.81
[13] 159.81 179.76 178.23 160.22 177.72 185.02 162.32 187.67
```

https://www.slideshare.net/KojiKosugi/ss-50740386

余談

頻度主義者ならどう言うか？

$$(\bar{X}) - t_{\alpha/2}(S_{\bar{x}}) < \mu < (\bar{X}) + t_{\alpha/2}(S_{\bar{x}})$$

だから，164.7897～175.3823の範囲に真の値が入っている可能性が95%

・頻度論
現実世界では母集団のパラメータは真に決まっているし、データもどこから取ってくるかで変わる。つまり、世の中的に正しい考え方。

https://www.slideshare.net/KojiKosugi/ss-50740386

余談

ベイジアンならどう言うか？

- 正規分布に基づくデータyがあったとして，ここから考えられる正規分布の形はどうなっているだろう？

・ベイズ論
世の中的には正しい考えではないが、結局母集団のパラメータはわからないし、いま手元にあるのはデータだけである。だったらいまある情報だけで母集団を考えよう、という合理的な考え方。

余談

- 頻度論における最尤推定
  - maximum likelihood (ML)

$$\theta_{\mathrm{ML}} = \arg\max_{\theta} \prod_{i=1}^{m} p(y^{(i)}|x^{(i)};\theta).$$

θは定数である．尤度を最大化するようなθが一番信頼できる．

https://www.slideshare.net/KojiKosugi/ss-50740386

余談

- 結局真のθが分からないから，変数と見なそう
- p(θ)の分布を使ってパラメータの事前信頼度（prior beliefs）を評価

$$\theta_{\mathrm{ML}} = \arg\max_{\theta} \prod_{i=1}^{m} p(y^{(i)}|x^{(i)};\theta).$$

$$\theta_{\mathrm{MAP}} = \arg\max_{\theta} \prod_{i=1}^{m} p(y^{(i)}|x^{(i)},\theta)p(\theta).$$

$$\theta \sim \mathcal{N}(0,\tau^2 I)$$

最大事後確率推定（Maximum a posteriori, MAP）

次回は

第1回　パターン認識のための線形代数
第2回　パターン認識のための確率論
第3回　凸最適化の概念
第4回　線形回帰とLMSアルゴリズム
第5回　分類とロジスティック回帰
第6回　一般化線形モデル
第7回　ガウシアン判別分析と単純ベイズ分類器
第8回　サポートベクターマシン（1）
第9回　サポートベクターマシン（2）
第10回　正則化とモデル選択
第11回　*k*平均法
第12回　混合ガウスモデルとEMアルゴリズム
第13回　因子分析
第14回　主成分分析

第15回　未定