機械学習 第５回 識別（3）

立命館大学 情報理工学部

福森 隆寛

Beyond Borders

講義担当者

高校生の時から立命館に通っているので

立命館
19年目！

- 福森 隆寛（ふくもり たかひろ）
  - 2015年3月　立命館大学 大学院情報理工学研究科 博士課程後期課程 修了
  - 2015年4月　立命館大学 情報理工学部 助教
  - 2020年4月　立命館大学 情報理工学部 講師
  - 専門分野：音声・音響情報処理の研究
    - 音声認識，雑音抑圧，叫び声検出など

大連理工大学の学生向けの講義を担当

私が現地（大連理工大学）で講義していた頃の写真

大連理工大学で授業

大連理工大学の食堂で昼食

休日観光@旅順＆棒棰島

講義スケジュール

（第1～4回、第14回）　（第5～13回、第15回）

- 担当教員：村上 陽平先生・福森 隆寛

1	機械学習とは、機械学習の分類
2	機械学習の基本的な手順
3	識別（1）
4	識別（2）
5	**識別（3）**
6	回帰
7	サポートベクトルマシン
8	ニューラルネットワーク

9	深層学習
10	アンサンブル学習
11	モデル推定
12	パターンマイニング
13	系列データの識別
14	強化学習
15	半教師あり学習

- 担当教員：叶 昕辰先生（第16回の講義を担当）

今回の講義内容

- 取り扱う問題の定義
- 統計的識別
- ベイズの定理（ていり）
- 最尤（さいゆう）推定法
- ナイーブベイズ識別
- 演習問題

取り扱う問題の定義：教師あり・識別問題

- カテゴリデータ、または数値データからなる特徴ベクトルを入力して、それをクラス分けする識別器を作る

※ 教師あり学習の識別問題での学習データは、以下のペアで構成される

入力データの特徴ベクトル（カテゴリデータ/数値データ） ← $\{\underline{\boldsymbol{x}_i}, \underline{y_i}\}, \quad i = 1,2, \dots, \underline{N}$ → 学習データの総数

カテゴリ形式の正解情報 → 「クラス」と呼ぶ

機械学習

- 教師あり学習
  - 識別
  - 回帰
- 中間的学習
- 教師なし学習

※ 第３回と同じ問題を考えます

統計的識別

□ 統計的識別

- 学習データの統計的性質に基づいて
  ある入力が、あるクラスに分類される確率（かくりつ）を求めて、
  最も確率が高いクラスを出力する方法

- 統計的識別を理解する上で必要な確率の知識（ちしき）
  - 事前確率、事後（じご）確率
  - 最大事後確率則
  - 尤度（ゆうど）
  - 最尤推定法

統計的識別：事前確率

□ 事前確率（prior probability）

- 入力を観測する前に持っている、それぞれのクラスの起(お)こりやすさ
- クラス数がcである場合、クラスω_iの事前確率は $P(\omega_i)$ $(i = 1, \ldots, c)$ と表す

※ 補足(ほそく)：$P(A)$と$p(x)$の違い

- $P(A)$：離散(りさん)的な事象(じしょう)Aが起こる確率
- $p(x)$：連続値xに対して定義される確率密度関数

どちらも確率なので…

- Aが全事象の場合：$P(A) = 1$
- $p(x)$をxの全区間で積分(せきぶん)すると1

統計的識別：事後確率、最大事後確率則

◻ 事後確率（posterior probability）

- 入力$\boldsymbol{x}$が観測されたとき、結果がクラスω_iである確率
- 条件付き確率 $P(\omega_i|\boldsymbol{x})$ で表現
  - 入力が観測された後（あと）で計算される確率

◻ 最大事後確率則

（maximum a posterior probability rule; MAP）

- 事後確率が最大となるクラスを識別結果とする手法
  - 統計的識別手法の代表的な方法
- 事後確率最大のクラスC_{MAP}の算出式

$$C_{MAP} = \underset{i}{\operatorname{argmax}}\, P(\omega_i|\boldsymbol{x})$$

統計的識別：事後確率の求め方

- 事後確率を学習データから求める方法を考える
  - 条件付き確率 $P(\omega_i$: 結論部 | $\boldsymbol{x}$: 条件部)
    - 条件部（特徴ベクトル$\boldsymbol{x}$）が一致するデータを集めて、結論部（クラスω_i）の頻度を求めることで推定
  - 事後確率を**<u>直接的</u>**に求めることは困難
    - 基本的に特徴ベクトル$\boldsymbol{x}$の組み合わせは膨大であり、それぞれのω_iに対して全ての組み合わせの$\boldsymbol{x}$を観測するのは困難
    - 観測されない組み合わせが表れると条件付き確率を計算できない
  - 事後確率を**<u>間接的</u>**に求めることを考える

統計的識別：事後確率の求め方

- **ベイズの定理** $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$ を用いても事後確率最大のクラスを計算できる

$$C_{MAP} = \underset{i}{\operatorname{argmax}} P(\omega_i|\boldsymbol{x}) = \underset{i}{\operatorname{argmax}} \frac{P(\boldsymbol{x}|\omega_i)P(\omega_i)}{P(\boldsymbol{x})}$$

ここで、$P(\boldsymbol{x})$はクラス番号(ばんごう)iを変化させても一定(いってい)なので、下式に置き換えられる

$$C_{MAP} = \underset{i}{\operatorname{argmax}} \underbrace{P(\boldsymbol{x}|\omega_i)}_{\text{尤度}} P(\omega_i)$$

尤度（likelihood）
クラスω_iから特徴ベクトル$\boldsymbol{x}$が出現する尤(もっと)もらしさ

「**尤度と事前確率の積**」を最大とするクラスを求めることによって事後確率が最大となるクラスを間接的に計算できる

演習問題5-1（10分間）

- 外見では区別のつかない箱Aと箱Bがある
  - 箱A：白玉が1個、黒玉が3個入っている
  - 箱B：白玉が4個、黒玉が1個入っている

1. 箱A、箱Bのいずれかから白玉を取り出す確率を求めよ。
   それぞれの箱の選び方は等確率とする。
2. 箱A、箱Bのいずれかから玉を取り出すと、白玉であった。
   この白玉が箱A、箱Bから取り出された確率をそれぞれ求めよ。
   それぞれの箱の選び方は等確率とする。
3. 箱A、箱Bの選ばれる確率が 9 : 1 であったとき、
   1. と 2. で求めた確率は、どのように変化するか？

学習データの生成確率

- 一般の問題では、事前確率や尤度はわからない
  - 事前確率や尤度を計算する確率モデルを仮定して、そのパラメータを学習データに合うように調整
  - それぞれのクラスは、特徴ベクトルを生成する何らかの確率分布をもっていて、学習データはその確率分布から、事例ごとに生成されたものと仮定
  - 学習データ全体Dが生成される確率$P(D)$の算出式

$$P(D) = \prod_{i=1}^{N} P(\boldsymbol{x}_i)$$

- この式は、個々の事例$\{\boldsymbol{x}_1, \dots, \boldsymbol{x}_N\}$の独立性を仮定している。

対数尤度

- $P(\cdot)$は、何らかのパラメータ$\boldsymbol{\theta}$に基づいて
データの生成確率を計算するモデルなので
下記(かき)のように$\boldsymbol{\theta}$を明示しながら$P(D)$を書き直せる

$$P(D|\boldsymbol{\theta}) = \prod_{i=1}^{N} P(\boldsymbol{x}_i|\boldsymbol{\theta})$$

- 確率は1以下なので、全学習データの複数回の積は
とても小さな数(かず)になるので、$P(D|\boldsymbol{\theta})$を対数化

対数尤度(たいすう) → $\mathcal{L}(D) = \log P(D|\boldsymbol{\theta}) = \sum_{i=1}^{N} \log P(\boldsymbol{x}_i|\boldsymbol{\theta})$

対数尤度が高いほど、そのモデルから学習データが生成された確率が高い

最尤推定法

□ 最尤推定法（maximum likelihood estimation）

- 対数尤度$\mathcal{L}(D)$が最大となるパラメータ$\boldsymbol{\theta}$を推定する手法

- モデルパラメータが1次元θの場合、
  対数尤度$\mathcal{L}(D)$を最大にするパラメータ$\hat{\theta}$は

$$\frac{d\mathcal{L}(D)}{d\theta} = 0$$

  を満たすθを計算する

- パラメータが多次元でも考え方は同様（どうよう）
  - 各パラメータで$\mathcal{L}(D)$を偏微分して極値（きょくち）を計算すれば良い

尤度の計算

- 事後確率最大則

$$C_{MAP} = \underset{i}{\operatorname{argmax}} P(\boldsymbol{x}|\omega_i)P(\omega_i)$$

の尤度$P(\boldsymbol{x}|\omega_i)$を求める方法を考える

- $P(\boldsymbol{x}|\omega_i) = P(x_1, \dots, x_d|\omega_i)$を統計で求めるには
  1. 学習データにあるクラスω_iに属するデータを取り出す
  2. そのデータに対して、**全ての特徴値の組み合わせ**の出現回数を数える
  - 学習データが小規模であるほど、統計が取れるほどのデータを揃えられる保証はない

ナイーブベイズ識別

- 各特徴値が、他の特徴とは独立に決定すると仮定
  - 特定の特徴について、特定の値（あたい）をとるデータを集めるので、すべての特徴値の組み合わせに対するデータよりは、かなり少ない量のデータで学習できる

- **ナイーブベイズ識別法**（naive Bayes classifier）
  - 上記の特徴の独立性を仮定した識別法
  - 識別結果をC_{NB}、特徴値の種類数をdとすると

$$C_{\mathrm{NB}} = \operatorname*{argmax}_{i} P(\omega_i) \prod_{j=1}^{d} P\left(x_j \middle| \omega_i\right)$$

事後確率最大則の尤度$P(\boldsymbol{x}|\omega_i)$がこれに置き換わる

ナイーブベイズ識別

$$C_{\mathrm{NB}} = \underset{i}{\operatorname{argmax}} P(\omega_i) \prod_{j=1}^{d} P(x_j|\omega_i)$$

□ $P(x_j|\omega_i)$の算出方法

$$P(x_j|\omega_i) = \frac{n_j}{n_i}$$

n_i：学習データの中でクラスω_iに属するデータ数
n_j：クラスω_iのデータの中で、ある特徴がx_jをとるデータ数

□ ゼロ頻度問題

- n_jが0の場合、この特徴に対する尤度 $P(x_j|\omega_i)$ も0となり、この特徴を含む全ての事例の確率が0となる
- **m推定**：ゼロ頻度問題の解決（かいけつ）手段
  - すでにm個の仮想（かそう）的なデータがあり、それらによって各特徴値も出現している状況を設定する考え方

$$P(x_j|\omega_i) = \frac{n_j + mp}{n_i + m}$$

m：標本数（ひょうほんすう）
p：各特徴値の出現割合（$p = 1/d$ が多い）

ナイーブベイズ識別

- 数値特徴に対するナイーブベイズ識別
  - カテゴリ特徴では、尤度が**離散事象**に対する確率分布 $P(x_j|\omega_i)$ であるのに対して、数値特徴では、数値に対する**確率密度関数** $p(x_j|\omega_i)$ になる

$$C_{\mathrm{NB}} = \operatorname*{argmax}_i P(\omega_i) \prod_{j=1}^{d} P(x_j|\omega_i)$$

カテゴリ特徴に対するナイーブベイズ識別

$$C_{\mathrm{NB}} = \operatorname*{argmax}_i P(\omega_i) \prod_{j=1}^{d} p(x_j|\omega_i)$$

数値特徴に対するナイーブベイズ識別

事前確率 $P(\omega_i)$
学習データ中のクラスω_iに属するデータを数える最尤推定で計算する

尤度 $p(x_j|\omega_i)$ ※ x_jが連続値なので、頻度を数える方法を利用できない
尤度を計算する確率密度関数に適切な**統計モデル**を当てはめて
そのモデルのパラメータを学習データから推定する

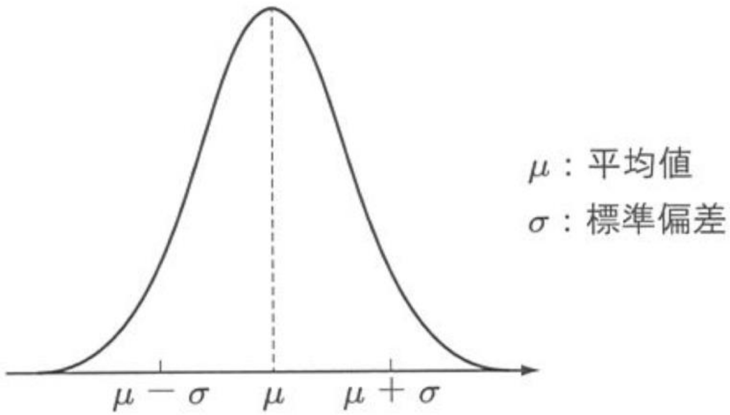
統計モデルの例

□ 正規(せいき)分布

- 数値データに対して多く用いられる統計モデル
- 釣(つ)り鐘型(がねがた)の分布
- 正規分布のパラメータ：平均値 μ、標準偏差 σ
  - これらのパラメータの値が定まると、分布の形状(けいじょう)が決まる

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

1次元データの正規分布



正規分布の形

※ 正規分布の平均値と標準偏差の最尤推定値 → 学習データの平均値と標準偏差

生成モデル・識別モデル

□ 生成モデル

- 「あるクラス」の「ある特徴ベクトル」が「ある確率」に基づいて生成されるという考え方
- ベイズの定理を使って事後確率を考える

$$P(\omega_i|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\omega_i)P(\omega_i)}{p(\boldsymbol{x})}$$

> $p(\boldsymbol{x}|\omega_i)P(\omega_i)$：生成モデルと解釈可能
> あるクラスω_iが確率$P(\omega_i)$で選ばれ、そのクラスから特徴ベクトル$\boldsymbol{x}$が確率$p(\boldsymbol{x}|\omega_i)$に基づいて生成

□ 識別モデル

- $P(\omega_i|\boldsymbol{x})$を直接推定する考え方
- **識別関数法**（識別モデルに近い考え方）
  - 入力に対して正しい値（真値に近い値）を出力する関数を確率分布などの制約を一切考えずにデータだけに注目して構成

演習問題5-2（10分間）

- 以下の識別法は、生成モデルと識別モデルのどちらに分類されるか？理由も含めて答えよ。
  - ナイーブベイズ識別
  - パーセプトロンの学習規則に基づく識別関数