PRML輪講 第一回

平成 29 年 4 月 12 日

1章 序論

- 結局何をしたいのか
  → データに内在するパターンを明らかにしたい
- 例 (手書き文字の認識)(イメージ図 1 参照)

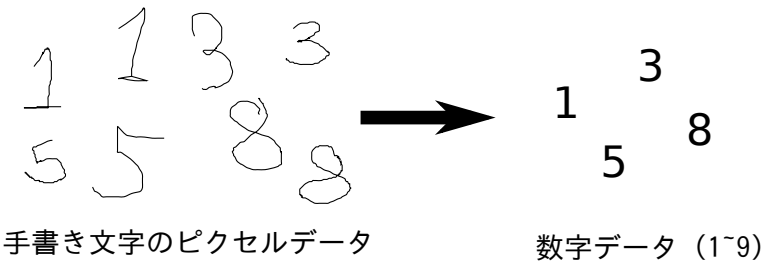

図 1: 文字認識のイメージ図

  - 入力: $\boldsymbol{x}$ (ピクセルを表現するような実数値ベクトル)
    出力: 0 から 9 までの数字
  - $\boldsymbol{x}$ を与えると 0 から 9 までの数字を出してくれるような機械を作りたい
- 機械学習的なアプローチ
  - 調節可能なパラメータを持つモデルを用意
  - N 個の手書き文字データ (**訓練集合**)$\{\boldsymbol{x}_1 \cdots \boldsymbol{x}_N\}$ とそれぞれの文字データに対応する数字 (**目標ベクトル**) $\boldsymbol{t}$ を用いて、パラメータを調整する。
  - 最終的に画像 $\boldsymbol{x}$ を入力すると対応する何らかの数字 $\boldsymbol{y}$ を返してくれる関数 $\boldsymbol{y}(\boldsymbol{x})$ が生成される。
  - 訓練集合にはなかった数字データ $\boldsymbol{x}$ (**テスト集合**) に対しても対応する数字を推測することができるようになる。
- テスト集合に対してどれぐらい正確に数字を判定することができるか → **汎化性能**
  - いくら訓練集合に対して正確に数字を当てることができたとしてもテスト集合でダメダメな結果であれば意味がない。 (過学習)
- **目標: 汎化性能を向上させたい**
- その他のトピック
  - 本来は入力変数を扱いやすい形に変換しておく → 前処理

- 訓練データが入力変数 $\boldsymbol{x}$ のみで目標値がないタイプの問題もある → 教師なし学習 (クラスタリングや密度推定などの問題)
- 与えられた状況で報酬を最大にする行動を学習 → 強化学習

- まずはわかりやすい例 (回帰問題) から初めて概念を説明することにする

1.1 多項式線形フィッティング

- 問題設定
  - 訓練集合として N 個の観測値を並べた入力 $\mathbf{x} = (x_1 \dots x_N)^{\mathrm{T}}$ と出力 $\mathbf{t} = (t_1 \dots t_N)^{\mathrm{T}}$ を用意
  - 出力は $\sin(2\pi x)$ にノイズ (ξ) を加えて生成するが、具体的な関数形 $\sin(2\pi x)$ は事前に知らされていないものとする

$$t_n = \sin(2\pi x_n) + \xi \tag{1}$$

  - これらのデータでモデルを訓練させることで、新たなテスト集合として x が与えられても、それに対する出力 t が正しく推測されるようにしたい (汎化性能の向上)

- まずはナイーブなアプローチ (多項式フィッティング) を考えてみる

$$y(x, \boldsymbol{w}) = \sum_{j=0}^{M} w_j x^j \tag{2}$$

  - 調節可能なパラメータは $\boldsymbol{w}$ と M

- 単純に二乗和誤差 $E(\boldsymbol{w})$ を最小にするようにパラメータ $\boldsymbol{w}$ を求めることを考える (ただの最小二乗法) [1)]

$$E(\boldsymbol{w}) = \frac{1}{2} \sum_{n=1}^{N} \{y(x_n, \boldsymbol{w}) - t_n\}^2 \tag{3}$$

演習 1.1 の解

$E(\boldsymbol{w})$ を最小にするような $\boldsymbol{w}$ を求めるため、誤差関数を w_i で微分することを考えると、

$$\begin{aligned} \frac{\partial}{\partial w_i} E(\boldsymbol{w}) &= \sum_{n=1}^{N} \{y(x_n, \boldsymbol{w}) - t_n\} \cdot x_n^i \\ &= \sum_{n=1}^{N} \sum_{j=0}^{M} w_j x_n^j x_n^i - \sum_{n=1}^{N} t_n x_n^i = 0 \end{aligned}$$

整理して

$$\sum_{j=0}^{M} \left(\sum_{n=1}^{N} x_n^{i+j} \right) w_j = \sum_{n=1}^{N} t_n x_n^j \tag{4}$$

となるから、これより $\boldsymbol{w}$ が一意に決まることがわかる。

[1)] 一見単純なことを延々とやっているようにも見えるが、後半でこのアプローチが確率を用いた枠組みで説明できることがわかる

- まだパラメーターとして多項式の次数 M が残っている
- M を増やせば増やすほどモデルを複雑にでき、誤差関数を 0 に近づけることができる
  - 特に、データ数と同じ数に M を設定すると誤差関数は必ず 0 となる
- **誤差関数を小さくしさえすればそれで OK なのか？**
  - 次数 M を大きくしすぎると与えられたデータに無理やり合わせたような結果になり、明らかに $\sin(2\pi x)$ を表現していない $\rightarrow$ **過学習**
  - 過学習が起きないように (まずは場当たり的に？) 工夫する必要がある
- 過学習の回避方法
  - 訓練集合を増やすことで過学習は起こりにくくなる (教科書図 1.6)
    - (訓練集合に対応してモデルの複雑さを変えるのはいかがなものか)
  - 過学習が起きるとき、$\boldsymbol{w}$ のノルムが大きくなる傾向があるため、これをペナルティ項を用いて抑える $\rightarrow$ **正則化**[2)]

$$\tilde{E}(\boldsymbol{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, \boldsymbol{w}) - t_n\}^2 + \frac{\lambda}{2}|\boldsymbol{w}|^2 \tag{5}$$

  - この場合の誤差関数を最小にする $\boldsymbol{w}$ も、前と同じように 1 つに定まる。

演習 1.2 の解

同じように誤差関数を w_i で微分すると、

$$\frac{\partial}{\partial w_i}\tilde{E}(\boldsymbol{w}) = \sum_{n=1}^{N}\{y(x_n, \boldsymbol{w}) - t_n\} \cdot x_n^i + \lambda w_i = 0$$

整理して

$$\sum_{j=0}^{M}\left(\sum_{n=1}^{N} x_n^{i+j} + \lambda\delta_{ij}\right) w_j = \sum_{n=1}^{N} t_n x_n^j \tag{6}$$

となる。

- M や λ についてはどう考慮すべきか
  - 得られているデータを訓練集合と、それとは別にチェックを行う用の**確認集合**に分ける
  - 訓練に使えるデータが減ってしまうデメリットがある
- 次の節から今までの概念を**確率**の観点で説明することを考える

[2)] この手法も後で確率を用いた定式化ができる。

1.2 確率論

- 確率の基本法則 (そこまで詳しく行う必要はないと思うので簡単に紹介)
  - 加法定理

$$p(X) = \sum_{Y} p(X, Y) \tag{7}$$

  - 乗法定理

$$p(X, Y) = p(X|Y)p(Y) \tag{8}$$

- 同時確率の対称性 $P(X, Y) = P(Y, X)$ を用いることで、

$$P(Y|X)P(X) = P(X|Y)P(Y) \tag{9}$$

であることから、**ベイズの定理**

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)} \tag{10}$$

が成り立つ。

  - 条件付き確率の変数 (ここでは X と Y) をひっくり返すことができる。

- **頻度主義** (ざっくりいうと従来の確率の扱い) と**ベイズ主義** (ざっくりいうと新しい確率の見方) の違い
  - 頻度主義
    観測されたデータは**「神のみぞ知る唯一無二の真のモデル」から発生したものの 1 つ**であると考える。つまり真のモデルはただ 1 つだけであり、データは確率的に変動すると考える。データをいっぱい取りまくれば (観測しまくれば) それだけ真値に近づくことができると考える。
  - ベイズ主義
    唯一無二の真のモデルなど存在しない。データが全てである。データを取ることにより (観測することにより) 真値の”確率分布”をベイズの定理を用いて逐次更新していく。ここで言う”確率”とは”信念の度合い”のようなものを指す。[3)]
    - ベイズ主義の適用例 (教科書 p15 参照)
    - オレンジを得たという情報を得ることにより、箱の色がどっちであるかという確率がベイズの定理により更新される。P(赤い色の箱を選んだ確率) → ベイズの定理 → P(**箱からオレンジを取り出**赤い色の箱を選んだ確率)
  - 要するに、**頻度主義では真値が定数でデータが確率変数、ベイズ主義ではデータが定数で真値が確率変数**
  - 前の多項式フィッティングの例で考えてみると、ベイズの定理で次のように表せる、

$$P(\boldsymbol{w}|\mathbf{t}) = \frac{P(\mathbf{t}|\boldsymbol{w})P(\boldsymbol{w})}{P(\mathbf{t})} \tag{11}$$

ここで

$$P(\boldsymbol{w}) \cdots \text{(まだ情報が与えられてない時点での) パラメータの確率 (信念の度合い)} \tag{12}$$

$$P(\mathbf{t}|\boldsymbol{w}) \cdots \text{パラメータ } \boldsymbol{w} \text{ を固定した際の観測データ } \mathbf{t} \text{ の起こりやすさ} \tag{13}$$

$$P(\boldsymbol{w}|\mathbf{t}) \cdots \text{(観測データが与えられたという条件の元での) パラメータの確率 (信念の度合い)} \tag{14}$$

[3)] 確率が高い → おそらくそれっぽいだろう、確率が低い → たぶんこれではないだろう、といった具合。合格率 xx 割みたいな使いかたに近い。

- $P(\boldsymbol{w})$ を事前確率、$P(\mathbf{t}|\boldsymbol{w})$ を尤度関数、$P(\boldsymbol{w}|\mathbf{t})$ を事後確率と呼ぶ。
- ベイズの定理を用いることで、$\boldsymbol{w}$ が与えられたときのデータ $\mathbf{t}$ の分布から、この 2 つをひっくり返した、データ $\mathbf{t}$ が与えられたときの $\boldsymbol{w}$ の分布を評価できるという強みがある。

- そもそもやりたかったことは何か → パラメータ $\boldsymbol{w}$ の推定
- 頻度主義で広く用いられる手法 → **最尤推定**
  - $P(\mathbf{t}|\boldsymbol{w})$ を最大にするような $\boldsymbol{w}$ を選んでくる。
  - 手元にあるデータを最も実現しそうな (最も尤もらしい) パラメータ $\boldsymbol{w}$ を決める操作
- この後、尤度関数をガウス分布と仮定することで (前に見た) 最小二乗法と最尤法とのつながりが見えることを示す。
- ガウス分布 (μ: 平均、σ^2: 分散)

$$\mathcal{N}(x|\mu,\sigma^2)=\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \tag{15}$$

- 特に多変量ガウス分布の場合 (D 次元の場合) ($\boldsymbol{\Sigma}$: 共分散行列) [4)]

$$\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma})=\frac{1}{(2\pi)^{D/2}}\frac{1}{|\boldsymbol{\Sigma}|^{1/2}}\exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\} \tag{16}$$

- N 個の観測データ $\mathbf{x}_{\mathrm{tr}}=(x_1,\ldots,x_N)^{\mathrm{T}}$ がガウス分布から独立に生成されるとすると、(i.i.d)

$$p(\mathbf{x}_{\mathrm{tr}}|\mu,\sigma^2)=\prod_{n=1}^{N}\mathcal{N}(x_n|\mu,\sigma^2) \tag{17}$$

- 最尤法 → $p(\mathbf{x}_{\mathrm{tr}}|\mu,\sigma^2)$ を $\mathbf{x}_{\mathrm{tr}}|$ について最大化するような μ と σ^2 を求める問題
  - 解いてみると

$$\mu_{\mathrm{ML}}=\frac{1}{N}\sum_{n=1}^{N}x_n \tag{18}$$

$$\sigma^2_{\mathrm{ML}}=\frac{1}{N}\sum_{n=1}^{N}(x_n-\mu_{\mathrm{ML}})^2 \tag{19}$$

  - σ^2_{ML} の期待値が σ^2 に一致せず、$(N-1)/N$ 倍に縮小されてしまう。 → 最尤法による過学習の問題と関連
- 曲線フィッティングとの関係
  - 二乗和誤差の最小化 → 尤度関数がガウス分布 (平均 $y(x,\boldsymbol{w})$, 分散 β^{-1}) で与えられたと仮定したときの最尤法と同じ
    - 推定された値 $\boldsymbol{w}_{\mathrm{ML}}$、$\boldsymbol{\beta}_{\mathrm{ML}}$ を用いて新たな x が与えられたときの t の分布を推定できる
  - 正則化も含めた場合の最小化 → 事前分布を導入した際の**事後確率の最大化 (MAP)** と定式化できる
- **ML** を点推定するだけでなく、$\boldsymbol{w}$ のすべての値で積分することにより、よりベイズ的な扱いをして予測分布を求める。

[4)] 規格化因子は共分散行列を対角化することで求められる。