第 7 章

データサイエンス

担当者 : 小山聡（知能ソフトウェア研究室）
連絡先 : 部屋：情報エレクトロニクス棟 8-02　内線：6814
E-mail: oyama@ist.hokudai.ac.jp

主題とねらい

1. 多項式回帰について学ぶ。
2. 訓練誤差とテスト誤差の違い、過学習について学ぶ。
3. 交差検定とモデル選択の方法について学ぶ。
4. 正則化の方法について学ぶ。

キーワード：多項式回帰、訓練誤差、テスト誤差、過学習、交差検定、モデル選択、正則化

1　事前の準備等

講義「データサイエンス」のスライドの該当部分を復習しておくこと。参考書を一読しておくこと。余裕があれば Python のドキュメンテーションに目を通しておくこと。

2　実験の流れ（6 日間）

1 日目： Python の基本的な機能を理解し、実験に用いるデータを作成する。
2 日目： 多項式回帰の基本的なプログラムを作る。
3 日目： 様々な条件で、訓練誤差とテスト誤差の違いを評価し、過学習が生じる状況を観察する。
4 日目： 交差検定の方法を実装し、モデル選択を行う。
5 日目： 正則化多項式回帰のプログラムを作成する。
6 日目： プログラム、実験結果を整理しレポートにまとめる。

3　レポートの評価基準

本テーマではプログラム作成の際に特に以下を重視する。

1. 作成したプログラムについて、その内容を説明する適切なコメントが付与されていること。
2. 実験の条件が明記され、体系的に実験が行われていること。
3. 実験の結果について、適切な考察が行われていること。

4 注意

最初の方の課題で作成したプログラムが後で必要になる場合がある。プログラムは上書きせずに、古いバージョンのものも残しておくこと。また、関数を作成するなどして、作業の効率化を図ること。

特に、課題 7.21、課題 7.27 で作成したプログラムと課題 7.9、7.15、7.16、7.20、7.22、7.25、7.26、7.28 で作成したグラフはレポートに含める必要があるため、保存しておくこと。(他の課題の結果はレポートに含める必要はない。)

5 参考となる本やプログラム等

C. M. ビショップ: パターン認識と機械学習上, 丸善出版, 2012.

C. M. Bishop: Pattern Recognition and Machine Learning, Springer, 2006. (上記の原著)

Python 公式ドキュメンテーション `https://docs.python.org/ja/3/`

Python 演習ドキュメンテーション `https://github.com/islab-hokudai/csit-exercise4`

6 レポートの提出方法

作成したプログラムやグラフにコメントを加え、実験結果に考察を加えたものを PDF ファイルにして Moodle に提出すること。(詳細は 6 日目のページを参照のこと。)

（1日目）乱数を用いた実験データの作成

Python の基本的な機能を理解し、次回以降の実験に用いるデータを作成する。作成したデータは後で繰り返し使用できるようにファイルに保存する。

課題 7.1

まず Python でベクトルや行列の作成、行列とベクトルの間の基本演算に関する基本的な操作を試してみる。具体的には以下のドキュメントにある実行例を実行してみよ。

`https://github.com/islab-hokudai/csit-exercise4`

課題 7.2

回帰分析の正解となる関数 $y = \sin(2\pi x)$ を区間 $[0, 1]$ でグラフ上にプロットせよ。グラフのプロットの方法は以下に記載がある。

`https://matplotlib.org/gallery/lines_bars_and_markers/simple_plot.html#sphx-glr-gallery-lines-bars-and-markers-simple-plot-py`

ただし、上の例では区間 $[0, 2]$ となっている。ここでは、区間 $[0, 1]$ で 0.01 刻みの値 101 個を作成し配列 `x` に格納し、それを関数 `sin(2*numpy.pi*x)` への入力とせよ。

課題 7.3

入力データとして区間 $(0, 1)$ の一様乱数を 10 個生成する。`numpy.random` 関数を用いて一様乱数を要素に持つ 10 行 1 列の行列を作成せよ。

課題 7.4

一つ前の課題で作成した行列を `x` とし、`y=sin(2*numpy.pi*x)` として計算した `y` の各要素に、さらに乱数を使ってノイズを加えたデータを作成せよ。ノイズは平均 0、標準偏差 0.1 の正規分布に従うとし、`numpy.random.normal` 関数を用いて生成せよ。データ点 (x, y) の集合を平面上にプロットしてみよ。`matplotlib` モジュールをインポートして `matplotlib.pyplot.plot(x,y,'o')` を使うと良い。

課題 7.5

データ点 (x, y) の数 N を 10,20,50,100 と変化させて、ノイズを加えたサイズの異なる複数のデータセットを生成して行列に格納せよ。これらの行列を `numpy.savetxt` を用いてファイルに保存せよ。ファイルに保存するとき、後の課題で個別に使用するためデータの個数ごとに分けて保存することを推奨する。

（2日目）多項式回帰

多項式回帰の基本的なプログラムを作る。多項式回帰を線形方程式を解く問題に帰着させる。多項式の次数を増やしたときに生じる現象を観察する。課題 7.7〜7.10 では 1 日目に作成した 10 個のデータを用いる

課題 7.6

1 日目に作成した N 個のデータ (x_n, t_n) における x と y の関係を多項式

$$y(x, \boldsymbol{w}) = w_0 + w_1 x + w_2 x^2 + \cdots + w_M x^M = \sum_{j=0}^{M} w_j x^j$$

で近似することを考える。

関数のデータへの適合度を示す関数（誤差関数）として二乗和誤差関数

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

を用いる。このとき、誤差関数を最小にする $\boldsymbol{w}$ を求めるためには以下の線形方程式を解けばよいことを示せ。

$$\sum_{j=0}^{M} A_{ij} w_j = T_i$$

ここで、$A_{ij} = \sum_{n=1}^{N} (x_n)^{i+j}$、$T_i = \sum_{n=1}^{N} (x_n)^i t_n$ であり、線形方程式を行列形式で書くと

$$\boldsymbol{A}\boldsymbol{w} = \boldsymbol{T}$$

となる。ここで、i, j はそれぞれ 0 から始まることに注意せよ。（本課題は講義「データサイエンス」の内容の復習であり、提出する必要はない。理解が難しい場合は，講義資料で復習すること.）

課題 7.7

先週作成したデータから上の行列 $\boldsymbol{A}$ およびベクトル $\boldsymbol{T}$ を作成せよ。（最初は $M = 1$ から始めるとよい。）$\boldsymbol{A}$ は $M + 1$ 行 $M + 1$ 列の正方行列、$\boldsymbol{w}$ と $\boldsymbol{T}$ は長さ $M + 1$ の列ベクトルであることに注意せよ。

課題 7.8

線形方程式を解き、多項式の係数ベクトル **w** を求めるプログラムを作成せよ。

課題 7.9

多項式の次数 M を 1 から順に増やしていき、上で作成したプログラムで係数ベクトル **w** を求め、各次数の多項式関数を入力区間 $[0, 1]$ でグラフにプロットせよ。（訓練データと、学習で求まった多項式関数を同じグラフにプロットしてみよ。`matplotlib.pyplot.axis([0,1,-1,1])` でグラフの表示範囲を固定すると比較しやすくなる。）

課題 7.10

多項式の次数が大きくなると、どのような現象が起きるか、観察せよ。

（3日目）学習結果の評価

機械学習においては、訓練に用いるデータ（訓練データ）と学習したモデルの評価に用いるデータ（テストデータ）は独立である必要がある。通常は、元のデータをランダムに訓練データとテストデータに分割することで用意する。

課題 7.11

出力結果の平均二乗平方根誤差

$$E_{\mathrm{RMS}} = \sqrt{2E(\boldsymbol{w})/N} = \sqrt{\frac{1}{N}\sum_{n=1}^{N}\{y(x_n, \boldsymbol{w}) - t_n\}^2}$$

を計算するプログラムを作成せよ。（出力結果と正解をそれぞれベクトルに格納しておき、それらから計算すればよい。）

課題 7.12

1 日目に作成した 20 個のデータを同じ大きさの 2 つの集合に分割し、片方を用いて 3 次の多項式関数を学習させよ（二乗和誤差関数を最小にする係数を求めよ。）

課題 7.13

学習に用いたデータ（訓練データ）を学習した関数に入力した際の出力結果の平均二乗平方根誤差（訓練誤差）を計算せよ。

課題 7.14

もう片方のデータ（テストデータ）を学習した関数に入力した際の出力結果の平均二乗平方根誤差（テスト誤差）を計算せよ。

課題 7.15

多項式の次数を 1 から 9 まで変化させ、訓練誤差およびテスト誤差をグラフにプロットせよ。（多項式の次数を横軸に、訓練誤差およびテスト誤差を縦軸に取ること。）

課題 7.16

上の課題を、50 個のデータおよび 100 個のデータに対しても行い、グラフにプロットしてみよ。多項式の次数およびデータの数が訓練誤差とテスト誤差にどのような影響を与えるかを考察せよ。

（4日目）交差検定とモデル選択

限られたデータを有効に利用してモデルの評価を行うための交差検定を行う。データを同じ大きさの複数の（以下の例では 5 つの）部分集合に分割し、そのうち 4 つの部分集合のデータを訓練データとして用いて訓練し、残りの 1 つの部分集合のデータをテストデータとして用いて評価する。これを訓練データとテストデータの組合せを変えて 5 回行い平均をとる。

課題 7.17

1 から 5 までの整数値を取る一様乱数を 20 個発生させよ。（1 から 5 までに一度も現れない数があると以下の実験でエラーとなるため，その場合は再度乱数を発生させよ。）発生させた乱数はファイルに保存し、以下の実験ではファイルから繰り返し読みだして使うと実験の再現性が保たれてよい。

課題 7.18

1 日目で作成した 20 個のデータのうち、上で発生させた乱数値 1 から 4 に対応するデータを訓練データ、5 に対応するデータをテストデータとして、学習を行いテスト誤差を計算せよ。たとえば、乱数が格納されたベクトルを `a` とすると、ベクトル `x` から乱数の値が 5 に対応する要素を取り出すには `x[a==5]` とすればよい。

課題 7.19

上の処理を訓練データとテストデータの 5 通りの組合せに対して実行し、テスト誤差の平均を求める（5 分割交差検定）プログラムを作成せよ。

課題 7.20

多項式の次数 M を 1 から 9 まで変化させて交差検定によりテスト誤差の平均を計算し、グラフにプロットせよ。

課題 7.21

多項式の次数 M を自動で 1 から 9 まで変化させて線形回帰を実施し、交差検定を行い最適な次数を選択する（モデル選択）プログラムを作成せよ。（誤差が同じ場合は、小さな次数を選択せよ。）最適な次数はデータを分割する乱数によって異なる場合があるが、どれか一回の結果を使ってよい。

課題 7.22

20 個のデータ全てを訓練データとし、課題 7.21 で求めた最適な次数を用いて学習した関数を課題 7.9 と同じ要領でグラフにプロットせよ。

（5日目）正則化

2 日目に作成した多項式回帰のプログラムに正則化項を追加し，過学習が避けられることを確認する。

課題 7.23

以下の正則化項付きの二乗和誤差関数

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

を用いる。このとき、誤差関数を最小にする $\boldsymbol{w}$ を求めるためには以下の線形方程式を解けばよいことを示せ。

$$\sum_{j=0}^{M} A_{ij} w_j + \lambda w_i = T_i$$

ここで、$A_{ij} = \sum_{n=1}^{N}(x_n)^{i+j}$、$T_i = \sum_{n=1}^{N}(x_n)^i t_n$ であり、線形方程式を行列形式で書くと

$$(\boldsymbol{A} + \lambda \boldsymbol{I})\boldsymbol{w} = \boldsymbol{T}$$

となる。（$\boldsymbol{I}$ は単位行列である。）

課題 7.24

上の線形方程式を解いて正則化多項式回帰を行うプログラムを作成せよ。

課題 7.25

1 日目に作成した 20 個のデータを同じ大きさの訓練データとテストデータに分割し、多項式の次数 M を 9 に固定し、正則化パラメータ λ を 10^{-8} から 10^{4} まで変化させて訓練誤差およびテスト誤差（平均二乗平方根誤差）の変化をグラフにプロットせよ。また、その際に多項式の係数ベクトルの大きさ

$$\sqrt{\sum_{j=0}^{M} w_j^2}$$

の変化もグラフにプロットせよ。（係数ベクトルの大きさと誤差は別のグラフにプロットすること。）

課題 7.26

多項式の次数 M を 9 に固定し、正則化パラメータ λ を 10^{-8} から 10^{4} まで 10 倍ずつ変化させて、5 分割交差検定により各 λ におけるテスト誤差の平均を計算しグラフにプロットせよ。

課題 7.27

パラメータ λ を 10^{-8} から 10^{4} まで 10 倍ずつ自動で変化させて正則化多項式回帰を実施し、交差検定を用いて最適な λ を決定するプログラムを作成せよ。最適なパラメータはデータを分割する乱数によって異なる場合があるが、どれか一回の結果を使ってよい。

課題 7.28

20 個のデータ全てを訓練データとし、課題 7.27 で求めた最適なパラメータ λ を用いた学習した関数を課題 7.9 と同じ要領でグラフにプロットせよ。

（6日目）レポート作成

これまでの課題をレポートにまとめる。

- レポートには課題 7.21、課題 7.27 で作成したプログラムを含めること。プログラムには適切なコメントを付けること。
- レポートには課題 7.9、7.15、7.16、7.20、7.22、7.25、7.26、7.28 で作成したグラフを含めること。グラフには作成された条件（データ数や多項式の次数、正則化パラメータ）を明記すること。各グラフから読み取れることについて、それぞれ説明を記述すること。
- 多項式の次数、訓練データの数、正則化パラメータの大きさと、訓練誤差、テスト誤差の関係について考察を行うこと。
- 課題全体をとおして自分がどのような知識や技術を新しく身に付けることができたかについて述べよ。プログラミングや実験において技術的に工夫した点があれば記述せよ。最後にこの課題に関する感想（課題の難易度や分量、来年度の実験に向けての改善提案など）について述べよ。

締切は別途指定。PDF ファイルを Moodle に提出すること。レポートの表紙には、氏名、学年、コース、学生番号を記載すること。