最小二乗法
回帰分析の基礎

機械学習　輪講

14t4015x　井邑映斗

最小二乗法

- 回帰分析の基礎

↓

  - 与えられたデータがどのような関数から生み出されているか推測する
  - 仮定：多項式の関数関係がある

    多項式から得られる予測値と実際の観測データの誤差を最小にするよう多項式の係数を決定する

１．多項式近似と最小二乗法による推定

- 最小二乗法の解法
  - 求めるもの：ｘの多項式
  - 条件：観測データに対する二乗誤差を最小にする

1-1.
トレーニングセットの特徴変数と目的変数

- 例題：、0≦x≦1の範囲を等分する10個所の観測点xに対して、それぞれに観測値tが与えられる。
- Goal：xとt間の関数関係を推測する。

$$\{(x_n, t_n)\}_{n=1}^{10}$$

分析対象のデータ⇒トレーニングセット

表現の違い

- 統計学的には
  - x ⇒ 説明変数
  - t ⇒ 目的変数
- 機械学習的には
  - x ⇒ 特徴変数（分析対象の性質を特徴づける変数）
  - t ⇒ 目的変数

1-2.多項式近似と誤差関数の設定

- 多項式を想定
- $f(x)=w_0+w_1x+\cdots+w_mx_m$
  $=\sum_{m=0}^{M} w_m x^m$ （1）

  ＊実際にはMの値を具体的な値を1つに固定する
  Mの値が決まっているなら
  未知のパラメータはM+1個の係数$\{w_m\}_{m=0}^{M}$

正確の定義

- x_1～x_{10}の10個所の観測点について、(1)で計算される tの値と実際に観測された値t_nを比較すること
- それぞれの差の2乗を合計したものをこの推定における「誤差」と定義
- $\{f(x_1)-t_1\}^2+\{f(x_2)-t_2\}^2+\cdots+\{f(x_{10})-t_{10}\}^2$　（2）
- この値が大きいほど（1）から求められるtの値は実際の観測値とは一致しなくなる

誤差$\boldsymbol{E_D}$

- 資料に合わせ（2)の値を1/2倍したものを「誤差E_D」として定義
- $\mathrm{E}_D = \frac{1}{2}\sum_{n=1}^{N}\{f(x_n) - t_n\}^2$ …（3）
- N=10(観測点の数)
- （3）に（1）を代入
- $\mathrm{E}_D = \frac{1}{2}\sum_{n=1}^{N}(\sum_{m=0}^{M} w_m\, x_n^m - t_n)^2$ …（4）

$$\mathrm{E}_D = \frac{1}{2}\sum_{n=1}^{N}\left(\sum_{m=0}^{M} w_m\, x_n^m - t_n\right)^2$$

- $\{(x_n, t_n)\}_{n=1}^{10}$のトレーニングセットを用いる
- 未確定値は$\{w_m\}_{m=0}^{M}$
- 一般に（3）で求められる誤差を最小誤差、これを最小にする手法を最小二乗法と呼ばれる

1-3.誤差関数を最小にする条件

- $\mathrm{E}_D = \frac{1}{2}\sum_{n=1}^{N}\left(\sum_{m=0}^{M} w_m\, x_n^m - t_n\right)^2$ (4)
- 多項式の係数$\{w_m\}_{m=0}^{M}$の関数とみなす
- （4）を誤差関数と呼ぶことがある。

（4）を最小にする$\{w_m\}_{m=0}^{M}$の決定

- 条件：$\{w_m\}_{m=0}^{M}$の関数とみなした際の偏微分係数が０になる。

$$\frac{\partial \mathrm{E}_v}{\partial w_n}=0 \qquad \{\mathrm{m}=0,\cdots \mathrm{M}\} \qquad (5)$$

- 係数をまとめてw=$\{w_0,\ldots w_M\}^T$とみなすと勾配ベクトルは０

$$\nabla \mathrm{E_D}(w)=0 \qquad (6)$$

（4）を最小にする$\{w_m\}_{m=0}^{M}$の決定

- (5)に（4）を代入して偏微分

$$\sum_{n=1}^{N}\left(\sum_{m'=0}^{M} w_{m'} x_n^{m'} - t_n\right) x_n^m = 0 \qquad (7)$$

- 変形すると

$$\sum_{m'=0}^{M} w_{m'} \sum_{n=1}^{N} x_n^{m'} x_n^m - \sum_{n=1}^{N} t_n x_n^m = 0 \qquad (8)$$

- x_n^mを(m,n)成分とするN×（M+1)行列φを用いると

$$w^T \varphi^T \varphi - t^T \varphi = 0$$

wは求める係数を並べたベクトル
tは目的関数の観測値を並べたベクトル

（4）を最小にする$\{w_m\}_{m=0}^{M}$の決定

- φの成分はN個の観測点$\{x_n\}_{n=1}^{N}$についてそれぞれ０～M乗した値を並べた行列

$$\varphi=\begin{bmatrix} x_1^0 & \cdots & x_1^M \\ \vdots & \ddots & \vdots \\ x_N^0 & \cdots & x_N^M \end{bmatrix} \qquad (9)$$

x_n^mを(m,n)成分とするN×（M+1）行列φを用いると行列形式で書き直すことができる

$$\mathrm{w}=(\varphi^T\varphi)^{-1}\varphi^T\mathrm{t} \qquad (10)$$

（4）を最小にするときの条件

- （10）は与えられたトレーニングセットを用 いて、多項式の係数wを決定する公式になっている。
- ここまでの条件：$\{w_m\}_{m=0}^{M}$の関数とみなした際の偏微分係数が 0 になる。
- 追加条件：$(\varphi^T\varphi)^{-1}$が存在する
- E_Dの2階偏微分係数を表すヘッセ行列を用いる

ヘッセ行列

- 次の成分を持つ(M+1)×(M+1)の正方行列

$$\mathrm{H}_{mm\prime}=\frac{\partial^2 Ev}{\partial w_m \partial w_{m\prime}}$$

・(4)を代入

$$\mathrm{H}_{mm\prime}=\sum_{n=1}^{N} x_n^{m\prime} x_n^m \qquad (11)$$

以上より逆行列をとっているところがヘッセ行列に一致する。

M+1≦Nの場合

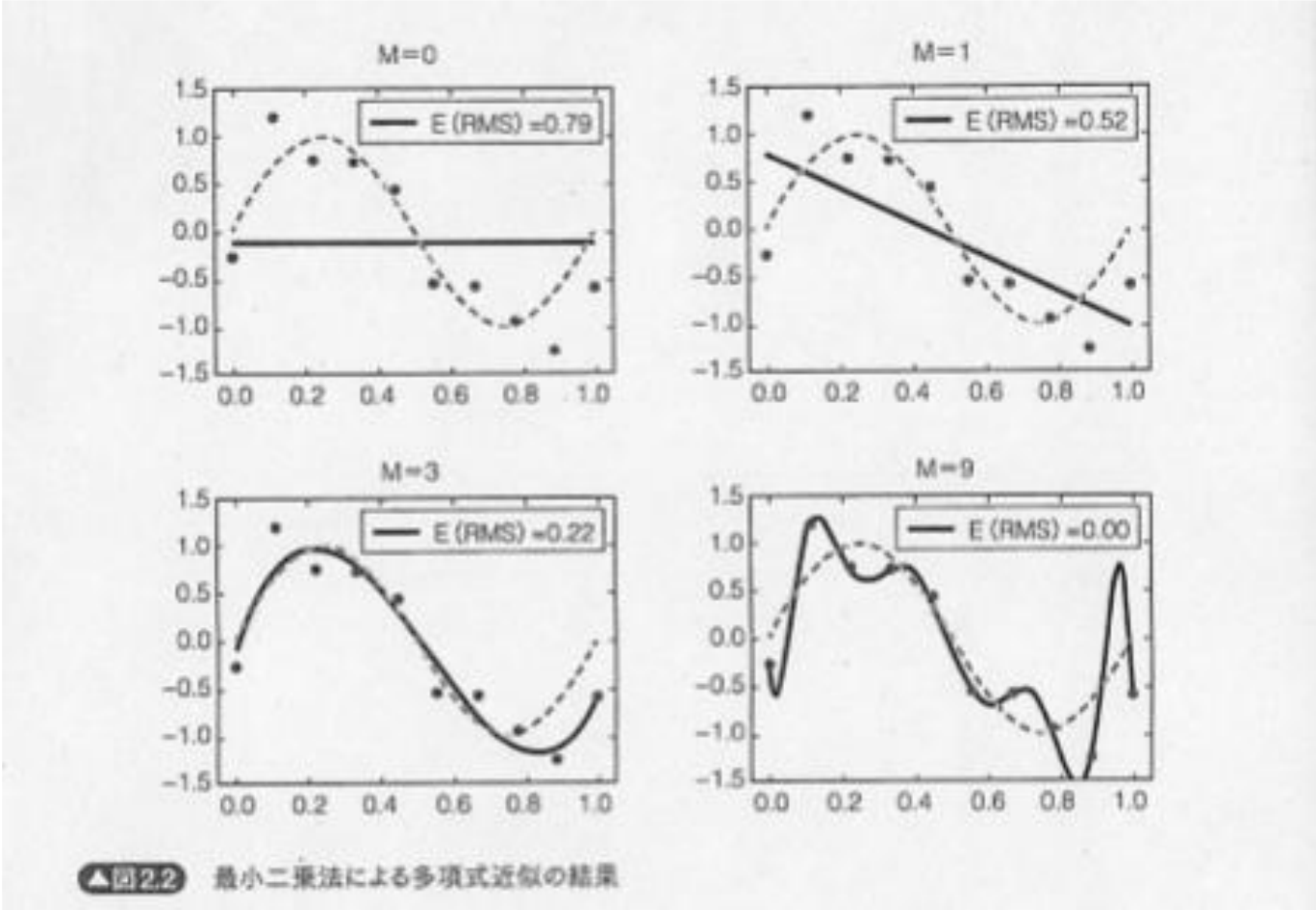
- M+1がトレーニングセットのデータ数N以下ならば
  ヘッセ行列は正定値
- 任意のu≠ 0に対して　u^tHu>0
- 正定値な行列は逆行列を持つので$\left(\varphi^{\mathrm{T}}\varphi\right)^{-1}$が存在する
  この時極小値$\mathrm{E_D}$

M+1＞Nの場合

- ヘッセ行列は半正定値をとる
- E_Dを最小にするwは複数存在して一意に決定できない

1-4.サンプルコードによる確認

- Mの値が大きくなるほどグラフの形状がいびつになりやすい
- Mの値は後から考える（この場合ならM=3）
- 客観的基準で次数Mを決定する



▲図2.2 最小二乗法による多項式近似の結果

ERMS　平方根平均二乗誤差

- 「多項式から予想される値とトレーニングセットの値が、平均的にどの 程度異なっているか」を示す

$$\mathrm{E_{RMS}}=\sqrt{\frac{2E_D}{N}}$$

- $\mathrm{E}_D=\frac{1}{2}\sum_{n=1}^{N}(\sum_{m=0}^{M} w_m\, x_n^m - t_n)^2$
- トレーニングセットに含まれるN個のデータについて「多項式 から予測される値とトレーニングセットの値の差の2乗を合計したものの半分」になっている

ERMS　平方根平均二乗誤差

- $\mathbf{E_{RMS}}$=0.79　　⇒平均的に0.79離れている
- $\mathbf{E_{RMS}}$=0　　⇒すべてのデータを正確に通過
- M=9の場合、多項式の係数は$\boldsymbol{w_1}$～$\boldsymbol{w_{10}}$の10個⇒10個のパラメーターを調整すれば、任意の10個の点を通過する曲線を作り出すことができるので、必ず誤差を0にすることが可能
- M≧10の場合は、データ数よりもパラメーターの個数の方が多くなる⇒すべてのデータを通過する係数は無数に存在する

1-5.統計モデルとしての最小二乗法

- 統計モデルの定義

  「何らかの現象について、統計学的な手法を用いて、それを説明、あるいは、予測するモデル(数式）を作り出すこと

- パラメトリックモデルと呼ばれる手法

  (1)パラメーターを含むモデル(数式）を設定する

  (2)パラメーターを評価する基準を定める

  (3)最良の評価を与えるパラメーターを決定する

パラメトリックモデルと呼ばれる手法

- （1）について
  目的関数tを予想する数式としてM次多項式を仮定する
- （2）について
  二乗誤差$\mathbf{E}_D$が小さいほどいいモデルであるという基準
- （3）について
  以上の基準に従って計算する
  w=$(\boldsymbol{\varphi}^T\boldsymbol{\varphi})^{-1}\boldsymbol{\varphi}^T$tが与えられている

2.オーバーフイッテイングの検出

- 機械学習
  トレーニングセット として与えられたデータに基づいて、最適なパラメーターを決定する
- データサイエンティスト
  得られた結果が「未来の値を予測する」ことに役立つ

2-1.トレーニングセットとテストセット

- トレーニングセットに含まれるデータ
  ⇒たまたま得られた値
  次に得られる値はこの限りではない。

  「M＝3を採用するのが良さそう」と判断したのは、
  このような考え方に基づく

最適な次数は
グラフを描いて判断するか？

- 複雑なデータを扱う場合はグラフで描くのが困難
- 正解のグラフは誰もわからない

⇒必要なのは「検証/仮説」の科学的思考

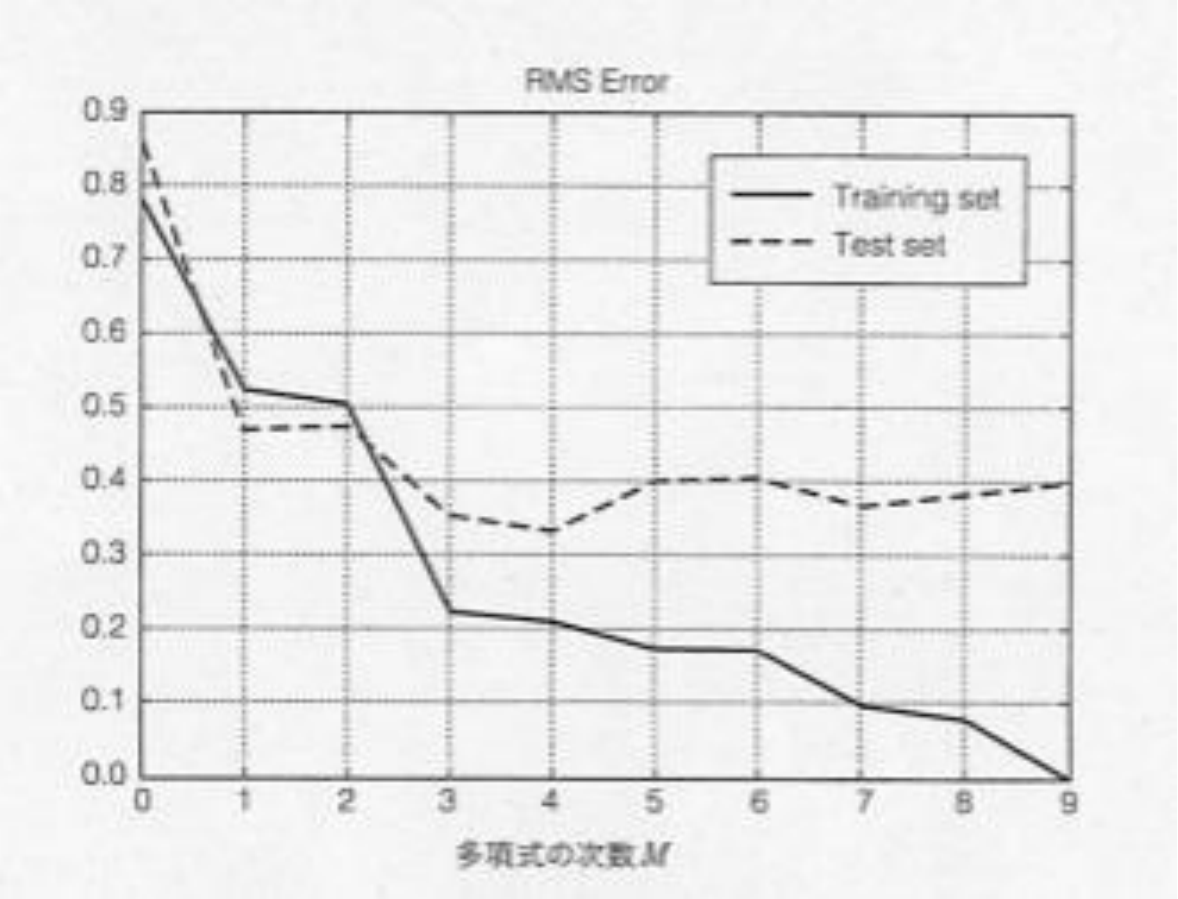
テストセット

- 実際に未来の値を予測して、どこまで正しく予測できるのかを検証する
- トレーニングセットとは別に、もう一度テスト用のデータを生成して、そちらに対し て、それぞれの多項式がどの程度よくあてはまるかを確認する
- 実際には、利用可能なデータを事前にトレーニング用とテスト用に分けておく

⇒トレーニングセットのデータを用いて係数を決定した後テストセットに対する平方根二乗誤差を計算

2-2.テストセットによる検証結果

- M=4以上に多項式の次数を大きくしてもテストセットに対する予測力は高くならない



モデルの汎化能力

- 未知のデータに対する予測能力
- M＝4を越えると、テストセットに対する誤差は減少せず、トレーニングセットに対する誤差だけが減少する
- トレーニングセットだけが持つ特徴に合わせて、過剰なチューニングをしている

⇒オーバーフィッティング（過学習）

2-3.
クロスバリデーションによる汎化能力の検証

- 機械学習のために収集したデータをトレーニングセットとテストセットに分割して使用

⇒テストセットのデータを 増やすということ

＝その分だけトレーニングセットのデータを減らすこと

検証の注意点

- 機械学習に使用したトレーニングセットに
  含まれるデータをテストセットに混ぜてはいけない

- テストセットの目的は、未知のデータに対する予測力、
  すなわちモデルの汎化能力を検証すること

⇒トレーニングセットのデータで検証することに意味がない

クロスバリデーション（交差検証）

- 利用可能なデータをパート1～パート5の5つのグループに分割して、どれか1つをテストセットとして用いる
- 異なるトレーニングセットを使用しているため5種類の検証結果が得られる。

⇒5種類の検証結果に基づいて、オーバーフィッティングが発生する（汎化能力が増加しなくなる）次数Mを決定できる

2-4.データ数による
オーバーフイッテイングの変化

- トレーニングセットのデータが10個の場合、パラメーター数が10個以上あれば、すべてのデータを正確に再現できる

⇒データ数が十分に多くあれ ば、多項式の次数が上がってもすべてのデータを再現することは難しくなる。

データ数を多くした場合（N=100）

- 多項式の次数を上げても
グラフの形状は歪まなくな
平方根平均二乗誤差は
ほぼ同じになる。
⇒M=3より次数を上げても
汎化能力は向上しない