統計的機械学習論
2. 最小二乗法

夏季集中講義(2019/09/02-13)
@九州工業大学
(飯塚キャンパス 大学院セミナー室７F)

このクラスのねらいと達成目標

ねらい

最小二乗法の学習を通して機械学習の理論的基礎となっている「統計モデル」の考え方について慣れていく。また、統計学や機会学習を学ぶ時に出てくるキーワード（「説明変数」「目的変数」「特徴変数」etc…）について自分なりに具体的にイメージし説明できるようになる。

達成目標

- 最小二乗法の考え方を理解する
- 統計で用いられる手法の大枠を知る
- 最小二乗法が数学的には関数の最小値を求める問題であることを知る

おしらせ

講義資料と演習用に使うPythonスクリプトを以下の公開リポジトリに上げてます。

授業中にとったノートの復習やテスト前の勉強に使ってください。

https://github.com/SaitaLab/StatsML

クラスの進行

2. 最小二乗法
    - 2.1.多項式近似と最小二乗法による推定
        - 2.1.1.トレーニングセットの特徴変数と目的変数
        - 2.1.2.多項式近似と誤差関数の設定
        - 2.1.3.誤差関数を最小にする条件
        - 2.1.4.サンプルコードによる確認
        - 2.1.5.統計モデルとしての最小二乗法
    - 2.2.オーバーフィッティングによる検出
        - 2.2.1.トレーニングセットとテストセット
        - 2.2.2.テストセットによる検証結果
        - 2.2.3.クロスバリデーションによる汎化能力の検証

2.1 多項式近似と最小二乗法による推定

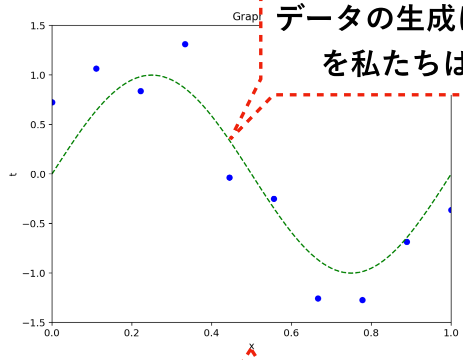
回帰分析の基礎：最小二乗法

データがどのような関数によって生み出されたかを推測する

多項式の関数を仮定して各項の係数を推定
↓
多項式から得られる予測値と観測データを比較する
↓
予測値と観測値の誤差が最小になる多項式の係数を推定
↓
目的の関数を得る

2.1.1 トレーニングセットの特徴変数と目的変数

0 ≦ x ≦ 1の範囲を等分する10個の観測点（トレーニングセット）に対して観測値tが与えられている。



データの生成に使った曲線を私たちは知らない

次にある観測点xを観測した時のtの値を予測したい

2.1.1 トレーニングセットの特徴変数と目的変数

統計学：tがある値をとる理由をxの値によって説明づけている

機械学習：分析対象tの性質をxによって特徴づけている

- トレーニングセット
  機械学習における学習用のデータ
- 説明変数
  観測値がその値をとる理由を説明する変数
  ここではxが該当する→「特徴変数」複数なら「特徴ベクトル」
- 目的変数
  推定することが目的となっている変数
  ここではtが該当

2.1.2 多項式近似と誤差関数の推定

多項式

$$f(x) = w_0 + w_1 x + w_2 x^2 + \ldots + w_M x^M = \sum_{m=0}^{M} w_m x^m$$

上記の多項式で計算された予測値と実際に観測された値を比較する。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」と定義する。

誤差

$$\{f(x_1) - t_1\}^2 + \{f(x_2) - t_2\}^2 + \ldots\{f(x_{10}) - t_{10}\}^2$$

2.1.2 多項式近似と誤差関数の推定

誤差が最小になるように多項式の係数を決定する

係数

$$\{w_m\}_{m=0}^{M}$$

誤差を1/2したものを「二乗誤差」と呼ぶ

二乗誤差

$$\begin{aligned} E_D &= \frac{1}{2}\left[\{f(x_1)-t_1\}^2+\{f(x_2)-t_2\}^2+...+\{f(x_{10})-t_{10}\}^2\right] \\ &= \frac{1}{2}\sum_{n=1}^{N}\{f(x_n)-t_n\}^2 \\ &= \frac{1}{2}\sum_{n=1}^{N}\left(\sum_{m=0}^{M} w_m x_n^m - t_n\right)^2 \end{aligned}$$

2.1.3 誤差関数を最小にする条件

二乗誤差を多項式の係数の関数とみなし、係数についての偏微分係数が0になる条件で決定する。

二乗誤差を係数で偏微分する

$$\frac{\partial E_D}{\partial w_m} = 0 \qquad (m = 0, 1, ..., M)$$

係数をベクトルとすると勾配ベクトルが0になることと同じ。

２乗誤差の勾配ベクトル

$$\nabla E_D(\mathbf{w}) = \mathbf{0}$$

2.1.3 誤差関数を最小にする条件

二乗誤差を多項式の係数で偏微分する

２乗誤差を係数で偏微分する

$$\frac{\partial E_D}{\partial w_m} = \sum_{n=1}^{N}\left(\sum_{m'=0}^{M} w_{m'} x_n^{m'} - t_n\right) x_n^m = 0$$

式変形し以下の形にする

$$\Downarrow$$

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

2.1.3 誤差関数を最小にする条件

式をベクトルの形に

２乗誤差の偏微分をベクトルの形に書き直す

$$\mathbf{w}^T \boldsymbol{\Phi}^T \boldsymbol{\Phi} - \mathbf{t}^T \boldsymbol{\Phi} = \mathbf{0}$$

$$\mathbf{w} = (w_0, ..., w_M)^T$$

$$\mathbf{t} = (t_1, ..., t_N)^T$$

$$\boldsymbol{\Phi} = \begin{pmatrix} x_1^0 & x_1^1 & ... & x_1^M \\ x_2^0 & x_2^1 & ... & x_2^M \\ ... & ... & ... & ... \\ x_N^0 & x_N^1 & ... & x_N^M \end{pmatrix}$$

2.1.3 誤差関数を最小にする条件

転置行列の性質（$(A^T)^{-1}=(A^{-1})^T$,（$(AB)^T = B^T A^T$）を使うと、
２乗誤差の偏微分の式は係数ベクトルについての式にできる

$$\mathbf{w}^T \mathbf{\Phi}^T \mathbf{\Phi} - \mathbf{t}^T \mathbf{\Phi} = \mathbf{0}$$

↓

多項式の係数ベクトルについての式

$$\mathbf{w} = \left(\mathbf{\Phi}^T \mathbf{\Phi}\right)^{-1} \mathbf{\Phi}^T \mathbf{t}$$

2.1.4 サンプルコードによる確認

トレーニングセット

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

多項式

$$f(x) = \sum_{m=0}^{M} w_m x^m$$

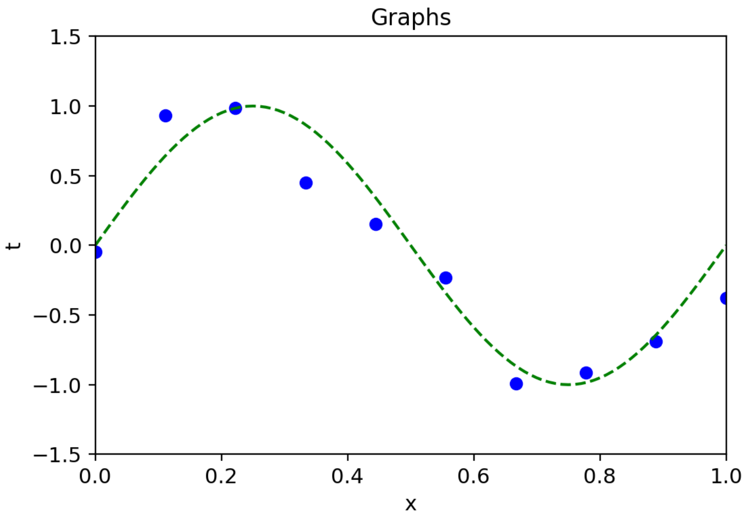
２乗誤差

$$E_D = \frac{1}{2} \sum_{n=1}^{N} \{f(x_n) - t_n\}^2$$

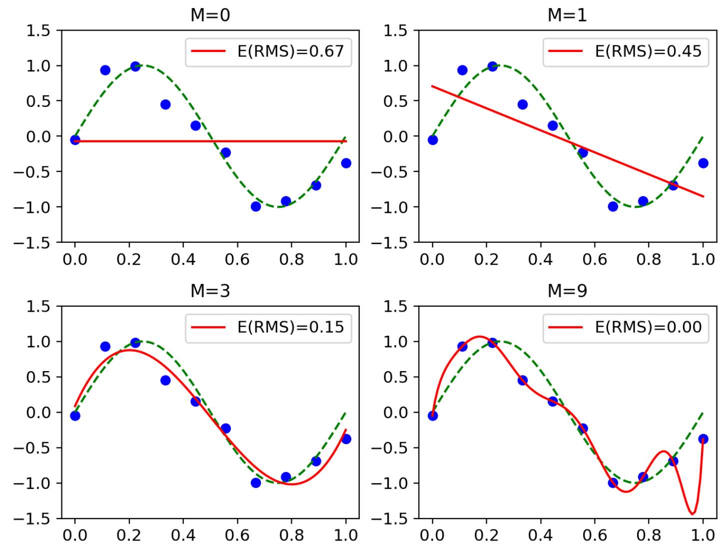
多項式の係数ベクトルについての式

$$\mathbf{w} = \left(\mathbf{\Phi}^T \mathbf{\Phi}\right)^{-1} \mathbf{\Phi}^T \mathbf{t}$$

2.1.4 サンプルコードによる確認



2.1.4 サンプルコードによる確認



2.1.4 サンプルコードによる確認

平方根平均二乗誤差（E_{RMS}）とは多項式から予測される値が平均的にどの程度異なっているか確認するために使う

2 乗誤差

$$E_D = \frac{1}{2}\sum_{n=1}^{N}\{f(x_n) - t_n\}^2$$

E_{RMS}はE_Dを含めた形になる

平方根平均二乗誤差

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

2.1.4 サンプルコードによる確認

多項式の係数が観測値の値に近づくとE_{RMS}=0にはなる

M≧Nのとき、E_{RMS}=0となる組み合わせは無数に存在する

↓

係数が一意に定まらない

2.1.5 統計モデルとしての最小二乗法

統計モデルとは
「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の３つのステップ

パラメータを含む数式モデルを設定する
↓
パラメータを評価する基準を定める
↓
最良の評価を与えるパラメータを決定する

2.2 オーバーフィッティングの検出

最適な次数Mを決定する方法

仮説で構築したモデルは未来の予測に役立つか？

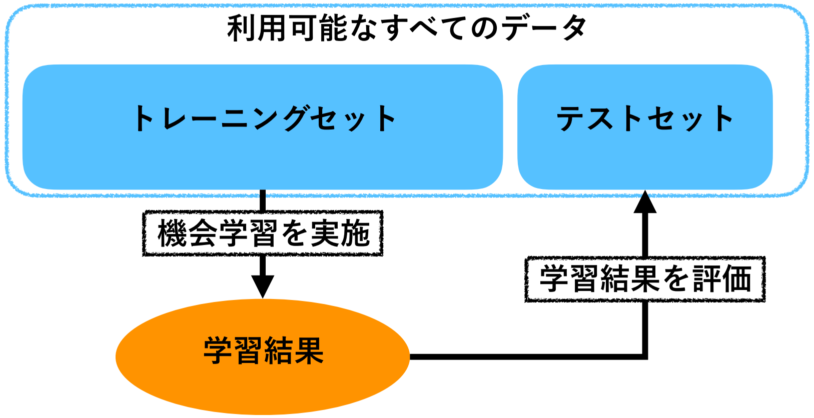
Mが大きければトレーニングセットを正確に再現可能だが…

↓

次に取得するデータは正確に再現できる？
（いや、できない（反語））

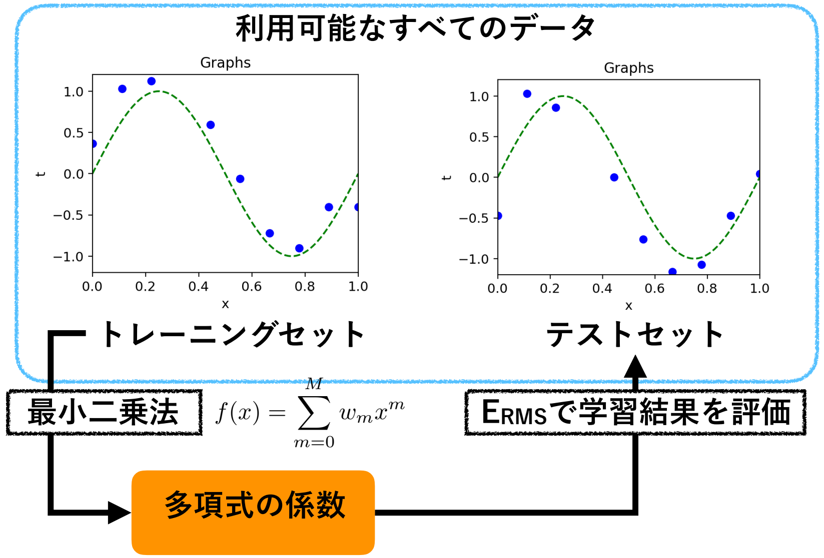
2.2.1 トレーニングセットとテストセット

利用可能なすべてのデータをトレーニング用とモデルの性能テスト用に分ける



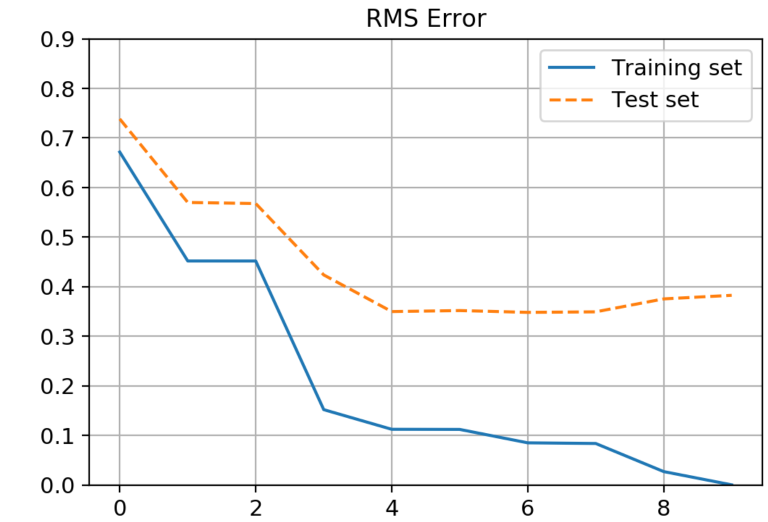
21

2.2.2 テストセットによる検証結果



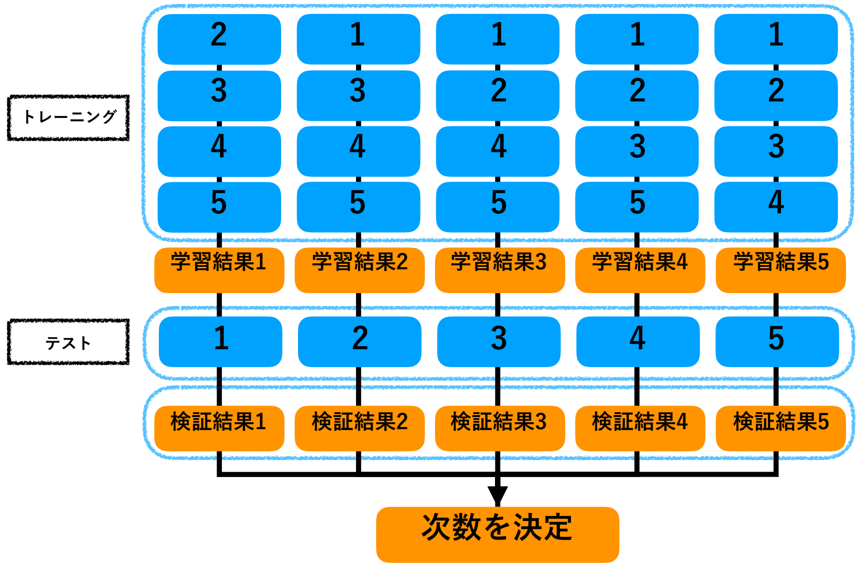
22

2.2.2 テストセットによる検証結果



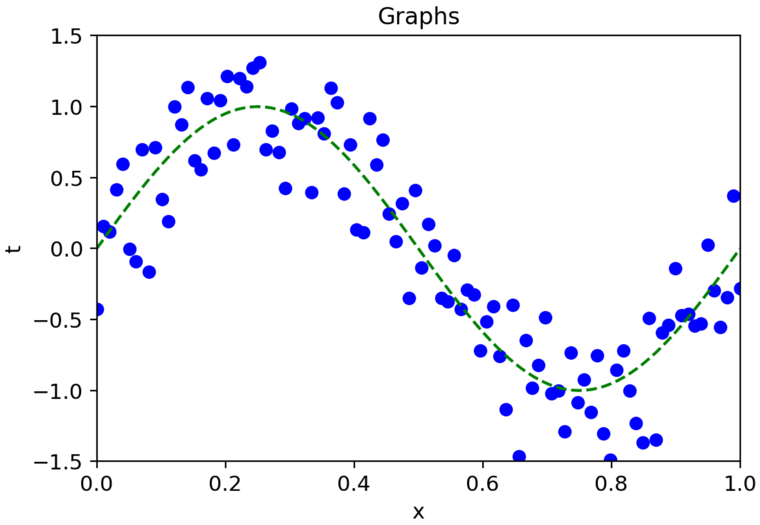
23

2.2.3 クロスバリデーションによる汎化能力の検証

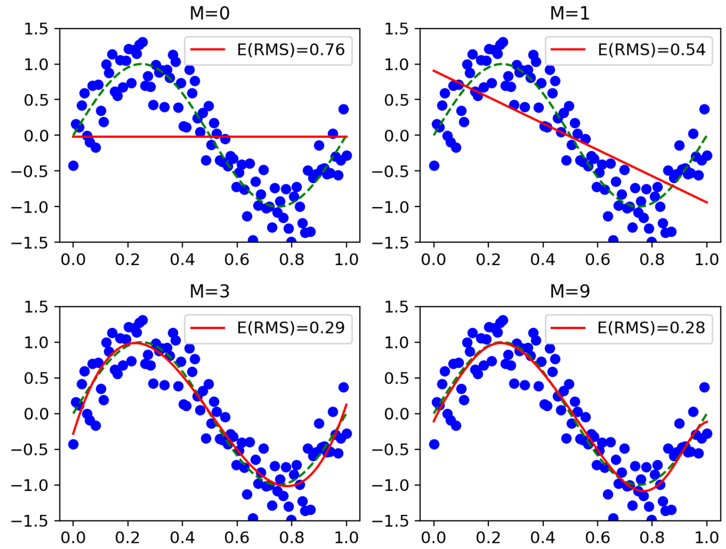


24

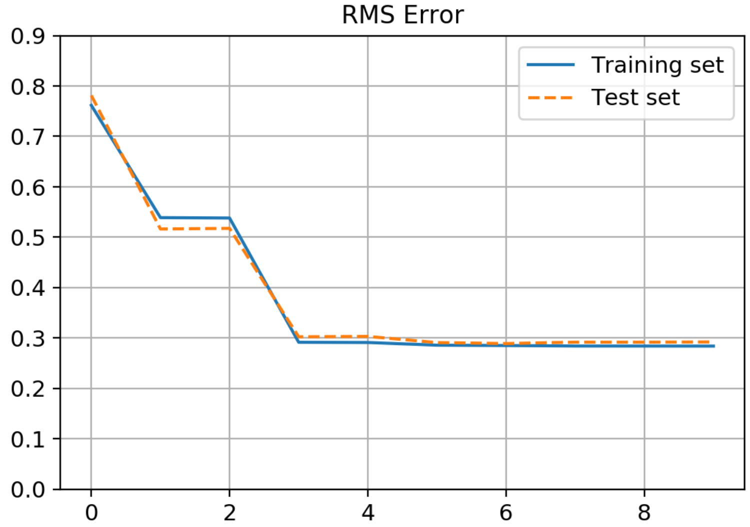
2.2.3 クロスバリデーションによる汎化能力の検証



2.2.3 クロスバリデーションによる汎化能力の検証



2.2.3 クロスバリデーションによる汎化能力の検証



まとめ

- 最小二乗法を理解した
- 機械学習でよく使われる言葉（特徴変数（説明変数）、目的変数、トレーニングセット、テストセット、オーバーフィッティング）について理解し自分の言葉で説明できるようになった
- パラメトリックモデルの３つのステップを把握した
- クロスバリデーションのためにどのようなデータセットを用意しなければならないか理解した

ミニッツペーパー

- 二乗誤差を記述してください。

$$E_D = \frac{1}{2}\sum_{n=1}^{N}\{f(x_n) - t_n\}^2$$

- パラメトリックモデルの３つのステップを記述してください。
  また、最小二乗法でのパラメトリックモデルの３つのステップで用いられている「モデル」「評価基準」「パラメータ」とは何かをそれぞれ説明してください。（スライドの14,19枚目参照）

- オーバーフィッティングとはどのような現象か説明してください。
  モデルが既存のデータに過度に最適化され汎化能力を失うこと