統計的機械学習論
3. 最尤推定法

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

最尤推定法の学習を通して観測値に含まれる誤差について考慮することを学んでいく。「誤差」を見積もることで、予測の精度を評価できることを理解する。

達成目標

- 最尤推定法の考え方を理解する
- 観測値に含まれている「誤差」を考慮して問題を考えられる。
- 最尤推定法が数学的には関数の最大値を求める問題であることを理解する。

クラスの進行

3 最尤推定法

パラメトリックモデルの３つのステップをガイドラインとして解説を進めていく

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

3.1 確率モデルの利用

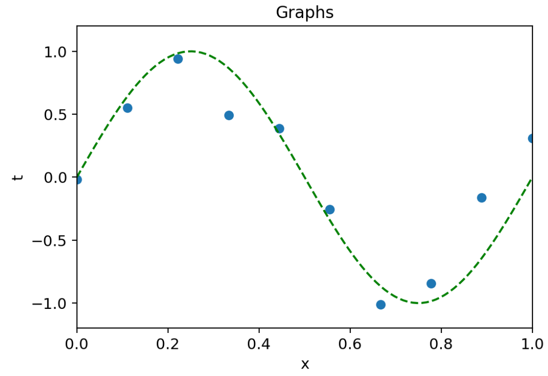
最尤推定法では「あるデータが得られる確率」を設定しパラメータを決定する

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

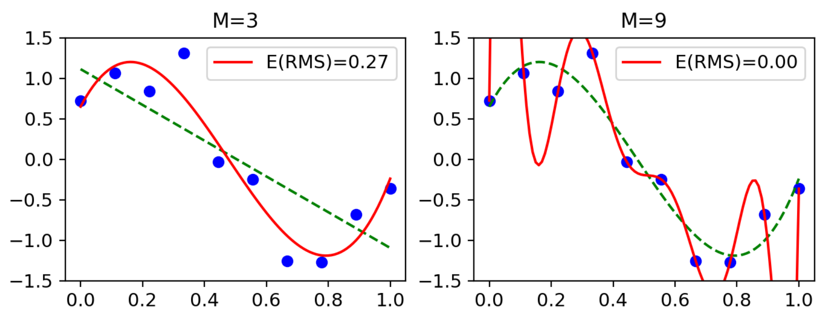
3.1.1 「データの発生確率」の設定

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。このデータの「本質的な性質」を考えてみよう。



3.1.1 「データの発生確率」の設定

最小二乗法では説明変数と目的変数の関数関係を推測したが、「すべての点を通る関数」を発見できても、それは「未来を正確に予測する」関数ではなかった。



3.1.1 「データの発生確率」の設定

ここで最小二乗法の考え方に加えて、誤差を考慮しよう。

観測値は本質的に何らかの誤差を含む

誤差を見積もることができれば、予測の精度を評価できる。

観測値＝多項式＋誤差

多項式

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

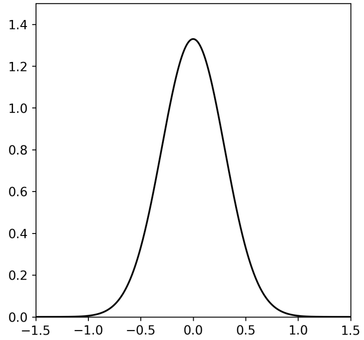
3.1.1 「データの発生確率」の設定

誤差がどのようにして発生するか既存のモデルで仮定する

正規分布に従う乱数で表現

正規分布

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

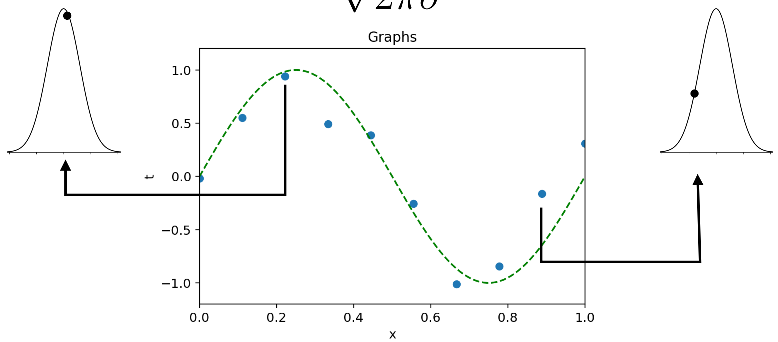


例：平均(μ)0, 標準偏差(σ)0.3の正規分布

3.1.1 「データの発生確率」の設定

観測値は多項式を中心に誤差によって散らばって分布する（平均値μをf(x)に置き換える）

$$N(t\,|\,f(x),\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(t-f(x))^2}$$



3.1.1 「データの発生確率」の設定

観測値＝多項式＋誤差

$$N(t\,|\,f(x),\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(t-f(x))^2}$$

多項式

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

係数

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

このモデルに含まれるパラメータは係数w_mとσである

3.1.1 「データの発生確率」の設定

なぜ誤差を正規分布で表現するのか？

他の確率分布：

- t分布
- ロジスティック分布
- 一様分布

etc, …あらゆる可能性にこだわるとキリがない。何か１つ仮説を立て、有用なら使う、そうでないならその理由を分析し新たな仮説を立て検証する。仮説と検証の繰り返しで最適なモデルを見つける。

3.1.2 尤度関数によるパラメータの評価

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

パラメトリックモデルの２つ目のステップに進む。

3.1.2 尤度関数によるパラメータの評価

ある観測点x_nについて、観測値t_nが得られる確率は：

観測値=多項式＋誤差

$$N(t \mid f(x), \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(t-f(x))^2}$$

トレーニングセット全ての観測値を得られる確率

=各データを得られる確率を掛け合わせたもの

$$\begin{aligned} P &= N(t_1 \mid f(x_1), \sigma^2) \times \dots \times N(t_N \mid f(x_N), \sigma^2) \\ &= \prod_{n=1}^{N} N(t_n \mid f(x_n), \sigma^2) \end{aligned}$$

3.1.2 尤度関数によるパラメータの評価

Pは多項式の係数w_m(m=0, 1, ..., M)とσの関数：尤度関数

尤度関数

$$\begin{aligned} P &= N(t_1 \mid f(x_1), \sigma^2) \times \dots \times N(t_N \mid f(x_N), \sigma^2) \\ &= \prod_{n=1}^{N} N(t_n \mid f(x_n), \sigma^2) \end{aligned}$$

ここで仮説を立てる：

観測されたデータは最も発生確率が高い

↓

Pが最大になるようにパラメータを決定する関数の最大値問題

3.1.2 尤度関数によるパラメータの評価

Pについては、式変形によって答えを求めることができる

尤度関数

$$\begin{aligned} P &= \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(t-f(x))^2} \\ &= \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{N}{2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{n=1}^{N}\{t_n - f(x_n)\}^2\right] \end{aligned}$$

この式には２乗誤差が含まれている

２乗誤差

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

3.1.2 尤度関数によるパラメータの評価

E_Dを使ってPを書き換える

尤度関数

$$P = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{N}{2}} e^{-\frac{1}{\sigma^2}E_D}$$

Pが依存するパラメータは σ^2 とE_D。

（β=(1/σ^2)、E_Dはベクトルw=(w_1, w_2, ⋯ , w_M)^{T}に依存）

Pの依存性を明示すると

尤度関数

$$P(\beta, \mathbf{w}) = \left(\frac{\beta}{2\pi}\right)^{\frac{N}{2}} e^{-\beta E_D(\mathbf{w})}$$

3.1.2 尤度関数によるパラメータの評価

Pを最大にする（β, w）を求めていく。

このとき、計算を簡単にするためにPの対数lnPを求める

尤度関数

$$P(\beta, \mathbf{w}) = \left(\frac{\beta}{2\pi}\right)^{\frac{N}{2}} e^{-\beta E_D(\mathbf{w})}$$

対数尤度関数

$$\begin{aligned}\ln P(\beta, \mathbf{w}) &= \ln\left[\left(\frac{\beta}{2\pi}\right)^{\frac{N}{2}} e^{-\beta E_D(\mathbf{w})}\right]\\ &= \ln\beta^{\frac{N}{2}} - \ln(2\pi)^{\frac{N}{2}} - \beta E_D(\mathbf{w})\\ &= \frac{N}{2}\ln\beta - \frac{N}{2}\ln(2\pi) - \beta E_D(\mathbf{w})\end{aligned}$$

3.1.2 尤度関数によるパラメータの評価

P（尤度関数）が最大 = lnP（対数尤度関数）が最大

lnPを最大にする(β, w)の条件を考える。

尤度関数

$$P(\beta, \mathbf{w}) = \left(\frac{\beta}{2\pi}\right)^{\frac{N}{2}} e^{-\beta E_D(\mathbf{w})}$$

対数尤度関数

$$\begin{aligned}\ln P(\beta, \mathbf{w}) &= \ln\left[\left(\frac{\beta}{2\pi}\right)^{\frac{N}{2}} e^{-\beta E_D(\mathbf{w})}\right]\\ &= \ln\beta^{\frac{N}{2}} - \ln(2\pi)^{\frac{N}{2}} - \beta E_D(\mathbf{w})\\ &= \frac{N}{2}\ln\beta - \frac{N}{2}\ln(2\pi) - \beta E_D(\mathbf{w})\end{aligned}$$

3.1.2 尤度関数によるパラメータの評価

lnPを最大にする(β, w)の条件を考える。

対数尤度関数

$$\begin{aligned}\ln P(\beta, \mathbf{w}) &= \ln\left[\left(\frac{\beta}{2\pi}\right)^{\frac{N}{2}} e^{-\beta E_D(\mathbf{w})}\right]\\ &= \ln\beta^{\frac{N}{2}} - \ln(2\pi)^{\frac{N}{2}} - \beta E_D(\mathbf{w})\\ &= \frac{N}{2}\ln\beta - \frac{N}{2}\ln(2\pi) - \beta E_D(\mathbf{w})\end{aligned}$$

lnPを最大にする(β, w)は次の条件で決まる

各変数による偏微分の式

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

$$\frac{\partial \ln P}{\partial \beta} = 0$$

3.1.2 尤度関数によるパラメータの評価

まず、w_mについて微分する

w_mについての微分の式の変形

$$\frac{\partial \ln P}{\partial w_m} = \frac{\partial}{\partial w_m}\left(\frac{N}{2}\ln\beta - \frac{N}{2}(\ln 2\pi)\right) - \beta E_D(\mathbf{w})$$
$$= \frac{\partial E_D}{\partial w_m} = 0$$

これは最小二乗法で２乗誤差を最小にする条件と同じ

wについての式に変形

$$\frac{\partial E_D}{\partial w_m} = 0 \rightarrow \nabla E_D(\mathbf{w}) = 0$$
$$\mathbf{w} = (\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\mathbf{\Phi}^T\mathbf{t}$$

21

3.1.2 尤度関数によるパラメータの評価

一方、βについて微分する

βについての微分の式の変形

$$\frac{\partial \ln P}{\partial \beta} = \frac{\partial}{\partial \beta}\left(\frac{N}{2}\ln\beta\right) - E_D = 0$$
$$\frac{\partial}{\partial \beta}(\ln\beta) = \frac{1}{\beta} = \frac{2}{N}E_D$$

これでσはE_{RMS}と同値であることがわかる

σについての式に変形

$$\sigma = \sqrt{\frac{1}{\beta}} = \sqrt{\frac{2}{N}E_D} = E_{RMS}$$

22

3.1.3 サンプルコードによる確認

トレーニングセット

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

観測値＝多項式＋誤差

$$N(t\,|\,f(x), \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2\sigma^2}(t-f(x))^2}$$

多項式

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

多項式の係数ベクトルについての式

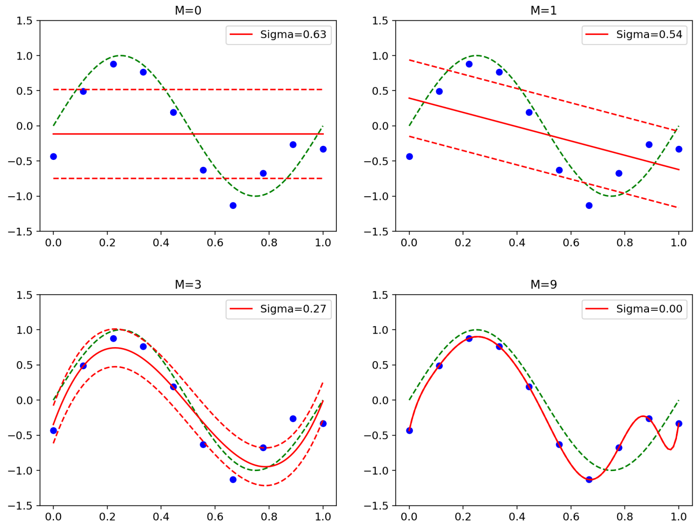
$$\mathbf{w} = \left(\mathbf{\Phi}^T\mathbf{\Phi}\right)^{-1}\mathbf{\Phi}^T\mathbf{t}$$

σ

$$\sigma = E_{RMS}$$

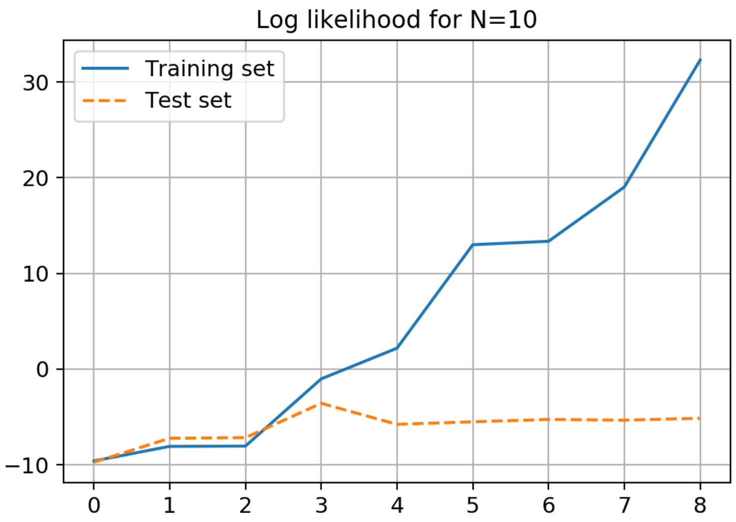
23

3.1.3 サンプルコードによる確認



24

3.1.3 サンプルコードによる確認



まとめ

- 観測値には誤差が必ず含まれていることを理解した
- 最尤推定法を理解した
- 最尤推定法と最小二乗法の類似点と相違点を把握した
- 最尤推定法とは数学的には関数の最大値を求める問題であることが理解できた

ミニッツペーパー

- 対数尤度関数を記述してください。

$$
\begin{aligned}
\ln P(\beta, \mathbf{w}) &= \ln \left[\left(\frac{\beta}{2\pi} \right)^{\frac{N}{2}} e^{-\beta E_D(\mathbf{w})} \right] \\
&= \ln \beta^{\frac{N}{2}} - \ln (2\pi)^{\frac{N}{2}} - \beta E_D(\mathbf{w}) \\
&= \frac{N}{2} \ln \beta - \frac{N}{2} \ln (2\pi) - \beta E_D(\mathbf{w})
\end{aligned}
$$

- 最尤推定法がなぜ尤度関数の最大値をとる問題になるのか説明してください。
  トレーニングセットを得る確率がもっとも高かった前提で考えるから

- 関数と誤差（標準偏差 σ）によって表現される、観測値tについての式を記述してください。

$$
N(t \mid f(x), \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(t-f(x))^2}
$$