確率・統計（真貝）第 11 回　参考資料
11_PS_MachineLearning.pdf

機械学習に使われる数学

1　機械学習の要素

ビッグデータの時代である．大量のデータから必要な情報を抽出する「機械学習」は，分類・回帰，クラスタリングなどに応用される．

- **分類（Classification）**
  データの類似性に基づいて，データをカテゴリー（あるいはクラス）に分ける機能である．郵便番号に記載された文字の判読，画像中の猫の判定などが代表例である．

  データの中でどこに境界線（境界面）を引くか，という数学になる．
  $\Longrightarrow$ 教科書 §4.6 判別分析

- **クラスタリング（Clustering）**
  類似しているデータどうしを集めて集合（クラスター）をつくり，データを仕分けする機能である．類似性分類の前処理でもある．正解か不正解かが明確ならば，「教師あり学習」と言える．

  データ間の距離を定義する，という数学になる．
  $\Longrightarrow$ 教科書 §4.7 クラスター分析

  データ $(x_1, \cdots, x_n)$ と $(y_1, \cdots, y_n)$ のユークリッド距離 d_2 は，

  $$d_2 = ||\boldsymbol{x}-\boldsymbol{y}|| = \sqrt{(x_1-y_1)^2 + \cdots + (x_n-y_n)^2}$$

  である．一般に，p ノルム d_p を

  $$d_p = ||\boldsymbol{x}-\boldsymbol{y}||_p = \sqrt[p]{(x_1-y_1)^p + \cdots + (x_n-y_n)^p}$$

  とする．また，重みをつけるさまざまな距離が提案されている．

- **回帰（Regression）**
  データの類似性に基づいて，データの数値を予測する機能である．株価予測や製造業における工程管理などに応用される．

  データ $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$ から，全体の傾向を知るための関数 $y = f(x)$ を導く数学になる．
  $\Longrightarrow$ 教科書 §4.3 回帰分析

2　目的関数の最適化

入力するデータが $(x_1, \cdots, x_m)$，出力するデータが $(y_1, \cdots, y_n)$ のとき, 両者の関係が線形であるとすれば,

$$\begin{aligned} y_1 &= a_{11}x_1 + a_{12}x_2 + \cdots + a_{1m}x_m + b_1 \\ y_2 &= a_{21}x_1 + a_{22}x_2 + \cdots + a_{2m}x_m + b_2 \\ \vdots &= \vdots \\ y_n &= a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nm}x_m + b_n \end{aligned}$$

と書ける．これをベクトルと行列を用いて,

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b} \tag{1}$$

と表す．以下のように定義した．

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{pmatrix}, \quad \boldsymbol{b} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix},$$

$$\boldsymbol{A} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nm} \end{pmatrix}.$$

2.1　目的関数

例えば，最小二乗法で，データ $\boldsymbol{x}$ と $\boldsymbol{y}$ の関係式 $y = ax + b$ を求めたいとき，誤差

$$e = \sum_i \{y_i - (ax_i + b)\}^2$$

を最小にするような (a, b) を求める問題になる．

これを一般化して，ある関数 S を最小化（最大化）するような問題を考えよう．例えばノルム

$$S = ||\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}||_2 \tag{2}$$

を最小化する問題を考える．

$$\begin{aligned} S &= (\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x})^T(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}) \\ &= \boldsymbol{y}^T(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}) - \boldsymbol{x}^T\boldsymbol{A}^T(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}) \\ &= \boldsymbol{y}^T\boldsymbol{y} - \boldsymbol{y}^T\boldsymbol{A}\boldsymbol{x} - \boldsymbol{x}^T\boldsymbol{A}^T\boldsymbol{y} + \boldsymbol{x}^T\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{x} \\ &= \boldsymbol{y}^T\boldsymbol{y} - 2\boldsymbol{y}^T\boldsymbol{A}\boldsymbol{x} + \boldsymbol{x}^T\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{x} \end{aligned}$$

S を $\boldsymbol{x}$ で微分して，（ベクトルの微分は成分ごとの微分をまとめたもの）

$$\frac{\partial S}{\partial \boldsymbol{x}} = -2\boldsymbol{A}^T\boldsymbol{y} + 2\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{x}$$

となるので，S を最小にする $\boldsymbol{x}$ は，この微分がゼロとなることから

$$\boldsymbol{x} = (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{y} \tag{3}$$

などと求められる．

2.2 最小二乗解が不偏推定量であること

式 (1) の $\boldsymbol{b}$ が誤差 $\boldsymbol{e}$ とする．$\boldsymbol{e}$ は，平均ゼロ，分散 σ_0 の正規分布にしたがうとする．最小二乗解を導いた式 (3) で，真の値 $\boldsymbol{x}$ を $\boldsymbol{x}_0$ とすると，

$$\begin{aligned}\hat{\boldsymbol{x}} &= (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T(\boldsymbol{A}\boldsymbol{x}_0 + \boldsymbol{e}) \\ &= \boldsymbol{x}_0 + (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{e}. \end{aligned} \tag{4}$$

誤差 $\boldsymbol{e}$ は正規分布にしたがう確率変数とすれば，$\hat{\boldsymbol{x}}$ の各成分も正規分布にしたがう．期待値をとると，

$$E[\hat{\boldsymbol{x}}] = \boldsymbol{x}_0 + (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T E[\boldsymbol{e}] = \boldsymbol{x}_0 \tag{5}$$

最後の部分は，$E[\boldsymbol{e}] = 0$ を用いた．この式は，最小二乗解 $\hat{\boldsymbol{x}}$ の期待値が真の値に一致するという保証（不偏推定量であること）を与えてくれる．

2.3 最小二乗解の分散

最適化された解を得たとしても，その不定性を添えないと不毛な議論になる．最小二乗解 $\hat{\boldsymbol{x}}$ の分散を計算しよう．

$$\begin{aligned} & E[(\hat{\boldsymbol{x}} - \boldsymbol{x}_0)(\hat{\boldsymbol{x}} - \boldsymbol{x}_0)^T] \\ = \; & E[((\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T E[\boldsymbol{e}])((\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T E[\boldsymbol{e}])^T] \\ = \; & (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T E[\boldsymbol{e}\boldsymbol{e}^T]\boldsymbol{A}(\boldsymbol{A}^T\boldsymbol{A})^{-1} \\ \underbrace{=}_{*} \; & (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T(\sigma_0^2\boldsymbol{I})\boldsymbol{A}(\boldsymbol{A}^T\boldsymbol{A})^{-1} = \sigma_0^2(\boldsymbol{A}^T\boldsymbol{A})^{-1} \end{aligned}$$

ここで，*部分では，$V[\boldsymbol{e}] = \sigma_0^2$ を用いた．この式で，$(\boldsymbol{A}^T\boldsymbol{A})^{-1}$ の (i,i) 成分を a_{ii} とする．

σ_0^2 の不偏推定量は，データ数を N，モデルのパラメータ数を K とすれば，

$$\hat{\sigma}^2 = \frac{||\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}||_2^2}{N - K} \tag{6}$$

となるので，パラメータ $\hat{x}_i$ の分散は，$\hat{\sigma}^2 a_{ii}$ となる．この平方根

$$\hat{s}(\hat{x}_i) = \hat{\sigma}\sqrt{a_{ii}} \tag{7}$$

は $\hat{x}_i$ の標準誤差と呼ばれる．