Data Science and Machine Learning

照屋 佑喜仁

May 29, 2025

2.1 Introduction

この章では，statistical learning(統計的学習) の概要とテーマについて導入する

- supervised learning(教師あり学習) と unsupervised learning(教師なし学習) の違い
- 教師あり学習の予測性能の評価

2.1 Introduction

- データサイエンスの主な課題はデータの数学的分析
- モデルを解釈し、データの不確実性を定量化するような分析は statistical learning(統計的学習) と呼ばれる
- 大規模データで予測を行うことに重点が置かれている場合は machine Learning(機械学習) や data mining(データマイニング) と呼ばれるのが一般的

2.1 Introduction

- データのモデリングには 2 つの主要な目標
  - 観測されたデータに基づいて、関心のある将来の値を正確に予測する
  - データ内の異常なパターン、または興味深いパターンを発見する
- これを達成するために数理科学の重要な 3 つ知識を用いる
  - *Function approximation*(関数近似)
  - *Optimization*(最適化)
  - *Probability and Statistics*(確率と統計)

2.1 Introduction

- Function approximation(関数近似)
  - データの数学モデルを構築するとは、ある変数が別の変数にどのように依存するかを理解すること
  - 変数間の関係を表現するために、関数または写像を用いる
  - 通常、この関数は完全にはわからないと仮定するが，十分な計算能力とデータがあれば良く近似できる
  - 最小限のコンピューターとメモリで関数を近似しうる最良の方法を理解する必要がある

2.1 Introduction

- Optimization(最適化)
  - モデルのクラスが与えられたとき、クラス内のどのモデルが最良か？
  - 何らかの効率的な探索または最適化手順が必要
  - 最適化アルゴリズムとコンピューターコーディングまたはプログラミングの知識が必要

2.1 Introduction

- Probability and Statistics(確率と統計)
  - 一般に，モデルを適合させるために使われるデータは不規則仮定 (確率過程？ random process) または数値ベクトルの実現 (結果？ realization) とみなされる
  - その確率法則が将来の予測できる精度を決定する
  - 将来に関する予測に内在する不確実性と、誤差の原因を定量化するために、確率論と統計的推論の知識が必要

2.2 Supervised Learning

機械学習の目標の一つとして，入力または特徴ベクトル $\boldsymbol{x}$ が与えられた場合に，出力または応答変数 y を予測することがある．
例えば

- $\boldsymbol{x}$ は署名のデジタル画像，y は署名が本物か偽物かを表す 2 値変数
- $\boldsymbol{x}$ は妊婦の体重と喫煙習慣，y は赤ちゃんの出生時体重

この予測を行う試みは，関数 g(予測関数) として符号化され，入力 $\boldsymbol{x}$ を受け取り推定値 $g(\boldsymbol{x})$ $(= \hat{y})$ を出力する．

2.2 Supervised Learning

- regression problems(回帰問題) では，応答変数 y は実数値をとることができる.
- y が有限集合 (例えば $y \in \{0, 1, \ldots, c-1\}$) に限定される場合，$y$ を予測することは入力 $\boldsymbol{x}$ に対して y の値を c 個のカテゴリーいずれかに分類することと同じであり，classification problems(分類問題) となる.

2.2 Supervised Learning

- 予測 $\hat{y}$ の精度を，与えられている応答 y に対して Loss function(損失関数)$\mathrm{Loss}(y, \hat{y})$ と呼ばれる関数を用いて測定することができる.
- 回帰の場合，通常 2 乗誤差損失 $(y - \hat{y})^2$ が用いられる.
- 分類の場合，通常は 0-1 損失関数
  $\mathrm{Loss}(y, \hat{y}) = \mathbf{1}\{y \neq \hat{y}\} \begin{cases} 0 & \text{if } y = \hat{y} \\ 1 & \text{if } y \neq \hat{y} \end{cases}$ が用いられる.
- この本の後半では cross-entropy loss(交差エントロピー損失) や hinge loss function(ヒンジ損失関数) などの他の損失関数も紹介される.

2.2 Supervised Learning

- 関数 g が，正確に対 $(\boldsymbol{x}, y)$ を予測することはほとんどありえない
- なぜなら，同じ入力 $\boldsymbol{x}$ でも偶然の状況やランダム性によってことなる出力 y が得られる可能性があるから
- したがって確率論のアプローチを採用．各対 $(\boldsymbol{x}, y)$ に対して何らかの joint probability density(結合確率密度 同時確率密度のこと？)$f(\boldsymbol{x}, y)$ を持つランダムな対 $(\boldsymbol{X}, Y)$ の結果であると仮定する
- 予測性能は，期待損失 (通常 risk と呼ばれる) によって評価する．

$$\ell(g) = \mathbb{E}\,\mathrm{Loss}(Y, g(\boldsymbol{X}))$$

2.2 Supervised Learning

- 分類問題なら，0-1 損失関数を用いて，リスクは $\ell(g) = \mathbb{P}[Y \neq g(\boldsymbol{X})]$ となる．
- この い文脈で，予測関数 g は classifier(分類器) と呼ばれる．
- $(\boldsymbol{X}, Y)$ の分布が与えられたとき，原理的に最良の $g^* = \arg\min_g \mathbb{E}[\mathrm{Loss}(Y, g(\boldsymbol{X}))]$ を見つけることができる．(これはすなわち g^* が最小リスクを与えるということ)
- 第 7 章では，分類問題において $g^*(\boldsymbol{x}) = \arg\max_{y \in \{0,\ldots,c-1\}} f(y|\boldsymbol{x})$ について調べる．
  - ここで $f(y|\boldsymbol{x}) = \mathbb{P}[Y = y|\boldsymbol{X} = \boldsymbol{x}]$ は $\boldsymbol{X} = \boldsymbol{x}$ が与えられたときの $Y = y$ の条件付き確率である

2.2 Supervised Learning

- 回帰で最も広く使われる損失関数は 2 乗誤差損失である（前述）
- この状況で，最適予測関数 g^* はしばしば regression function(回帰関数) と呼ばれる

Theorem (Optimal Prediction Function for Squared-Error Loss)

2 乗誤差損失 $\mathrm{Loss}(y, \hat{y}) = (y - \hat{y})^2$ に対して，最適予測関数 g^* は Y の $\boldsymbol{X} = \boldsymbol{x}$ が与えられたときの条件付き期待値に等しい.

$$g^*(\boldsymbol{x}) = \mathbb{E}[Y|\boldsymbol{X} = \boldsymbol{x}]$$

2.2 Supervised Learning

証明の前に，条件付き期待値の定義や propertiy を確認する．

Definition (conditional pdf)

X, Y は pdf$f$ に従う離散また連続確率変数とし，$f_X(x) \geq 0$ とする．
$X = x$ が与えられたときの Y の conditional pdf (条件付き確率密度関数) は

$$f_{Y|X}(y|x) = \frac{f(x,y)}{f_X(x)}$$

2.2 Supervised Learning

Definition (conditional expectation (条件付き期待値))

$X = x$ が与えられたときの Y の conditional expectation (条件付き期待値) は

$$\mathbb{E}[Y|X=x] = \begin{cases} \int_{\mathbb{R}} y f_{Y|X}(y|x)dy & \text{discrete case(連続)} \\ \sum_{y\in\mathbb{R}} y f_{Y|X}(y|x) & \text{continuous case(離散)} \end{cases}$$

$\mathbb{E}[Y|X=x]$ は x の関数であることに注意．有用な性質として，以下の tower property などがある．それについて示す．

2.2 Supervised Learning

Remark (tower property, taking out what is known)

tower property:

$$\mathbb{E}[\mathbb{E}[Y|X]] = \mathbb{E}[Y]$$

taking out what is known:

$$\mathbb{E}[XY|X] = X\mathbb{E}[Y|X]$$

2.2 Supervised Learning

Proof.

tower property の証明:

$$
\begin{aligned}
\mathbb{E}[\mathbb{E}[Y|X]] &= \int_{\mathbb{R}} \mathbb{E}[Y|X=x] f_X(x) dx \\
&= \int_{\mathbb{R}} \left(\int_{\mathbb{R}} y f_{Y|X}(y|x) dy \right) f_X(x) dx \\
&= \int_{\mathbb{R}} \int_{\mathbb{R}} y f_{Y|X}(y|x) f_X(x) dy dx \quad \because \text{() 内部は } x \text{ に依存しない} \\
&= \int_{\mathbb{R}} y \int_{\mathbb{R}} f(x,y) dx dy \\
&= \int_{\mathbb{R}} y f(y) dy \\
&= \mathbb{E}[Y]
\end{aligned}
$$

□

2.2 Supervised Learning

Proof.

taking out what is known の証明:

$$\begin{aligned}\mathbb{E}[XY|X] &= \int_{\mathbb{R}} xyf_{Y|X}(y|x)dy \\ &= x\int_{\mathbb{R}} yf_{Y|X}(y|x)dy \\ &= X\mathbb{E}[Y|X]\end{aligned}$$

□

2.2 Supervised Learning

Proof.

$g^*(\boldsymbol{x}) = \mathbb{E}[Y|\boldsymbol{X} = \boldsymbol{x}]$ とおく．任意の関数 g に対し，2 乗誤差損失は

$$\begin{aligned}
\mathbb{E}[(Y - g(\boldsymbol{X}))^2] &= \mathbb{E}[(Y - g^*(\boldsymbol{X} + \boldsymbol{X} - g(\boldsymbol{X}))^2] \\
&= \mathbb{E}[(Y - g^*(\boldsymbol{X}))^2] + 2\mathbb{E}[(g^*(\boldsymbol{X}) - g(\boldsymbol{X}))^2] + 2\mathbb{E}[(Y - g^*(\boldsymbol{X}))(g^*(\boldsymbol{X}) - g(\boldsymbol{X}))] \\
&\geq \mathbb{E}[(Y - g^*(\boldsymbol{X}))^2] + 2\mathbb{E}[(Y - g^*(\boldsymbol{X}))(g^*(\boldsymbol{X}) - g(\boldsymbol{X}))] \\
&= \mathbb{E}[(Y - g^*(\boldsymbol{X}))^2] + 2\mathbb{E}\left[\mathbb{E}\left[(Y - g^*(\boldsymbol{X}))(g^*(\boldsymbol{X}) - g(\boldsymbol{X}))|\boldsymbol{X}\right]\right] \\
&= \mathbb{E}[(Y - g^*(\boldsymbol{X}))^2] + 2\mathbb{E}\left[(g^*(\boldsymbol{X}) - g(\boldsymbol{X}))\mathbb{E}[Y - g^*(\boldsymbol{X})|\boldsymbol{X}]\right]
\end{aligned}$$

最後から 2 行の等式には tower property を，最後の行には taking out what is known を用いた．どちらも条件付き期待値の性質である．

2.2 Supervised Learning

Proof.

定義より,

$$\begin{aligned}
\mathbb{E}[Y - g^*(\boldsymbol{X})|\boldsymbol{X}] &= \int_{\mathbb{R}} (y - g^*(\boldsymbol{x})) f_{Y|\boldsymbol{X}}(y|\boldsymbol{x}) dy \\
&= \int_{\mathbb{R}} y f_{Y|\boldsymbol{X}}(y|\boldsymbol{x}) dy - \int_{\mathbb{R}} g^*(\boldsymbol{x}) f_{Y|\boldsymbol{X}}(y|\boldsymbol{x}) dy \\
&= \mathbb{E}[Y|\boldsymbol{X}] - \mathbb{E}[Y|X] \int_{\mathbb{R}} \frac{f(x,y)}{f_X(x)} dy \\
&= 0
\end{aligned}$$

$$\begin{aligned}
\because \int_{\mathbb{R}} \frac{f(x,y)}{f_X(x)} dy &= \frac{1}{f_X(x)} \int_{\mathbb{R}} f(x,y) dy \\
&= \frac{1}{f_X(x)} f_X(x) = 1
\end{aligned}$$

2.2 Supervised Learning

Proof.

したがって $\mathbb{E}[(Y - g(\boldsymbol{X}))^2] \geq \mathbb{E}[(Y - g^*(\boldsymbol{X}))^2]$ となり，これは g^* が最小の 2 乗誤差損失リスクを与えることを意味する． □

この定理のから，$\boldsymbol{X} = \boldsymbol{x}$ を条件とするとき，ランダムな応答 Y は

$$Y = g^*(\boldsymbol{x}) + \varepsilon(\boldsymbol{x})$$

と書くことができる．ここで $\varepsilon(\boldsymbol{x})$ は $\boldsymbol{x}$ における応答の条件付き平均からのランダムな偏差とみなすことができる．このランダムな偏差は $\mathbb{E}[\varepsilon(\boldsymbol{x})] = 0$ を満たす．さらに，ある正の関数 ν を用いて $\mathrm{var}[\varepsilon(\boldsymbol{x})] = \nu^2(\boldsymbol{x})$ と書くことができる．一般に $\varepsilon(\boldsymbol{x})$ は $\boldsymbol{x}$ に依存するが，$\nu^2(\boldsymbol{x})$ の確率分布は特定できない．

2.2 Supervised Learning

- 最適予測関数 g^* は通常，未知である $(\boldsymbol{X}, Y)$ の同時分布に依存するため，実際には利用できない
- その代わりに利用できるものは，同時確率密度 $f(x, y)$ から独立に得られた有限個の実現値のみ
- これらのサンプルを

$$\mathcal{T} = \{(\boldsymbol{X}_1, Y), \ldots, (\boldsymbol{X}_n, Y)\}$$

とし、これを training set(訓練セット) と呼ぶ.

- traning set $\mathcal{T}$ とその実現値 $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)$ を区別することが重要．後者は τ とし，サイズを強調したいときは τ_n と書く．(τ は実際に観測された具体的なデータで、$\mathcal{T}$ はまだ観測していないランダムなデータ集合のことか？)

2.2 Supervised Learning

- 目標：training set $\mathcal{T}$ にある n 個のサンプルから未知の g を"learn"(学習) すること
- $g_{\mathcal{T}}$ を，$\mathcal{T}$ から構築できる g^* の（何らかの基準による）最良の近似とする．g はランダムな関数であることに注意
- 特定の結果に関しては g_τ とする．
- ここで，教師と学習者という比喻を用いて教師あり学習について考えることができる
  - $g_{\mathcal{T}}$ は未知の関数 $g^* : \boldsymbol{x} \mapsto y$ を訓練データ $\mathcal{T}$ から学習する leaner(学習者)
  - 出力 Y_i と入力 $\boldsymbol{X}_i$ の真の関係を表すようなサンプルを n 個提供する teacher(教師) は，新しい入力 X に対する出力を予測するために学習者 $g_{\mathcal{T}}$ を train(訓練) する．この X に対する正しい出力 Y は教師によって提供されない（未知）

2.2 Supervised Learning

- 例：メールのスパム判定
  - 各メールを特徴ベクトル $\boldsymbol{x}$（例：単語の出現回数など）で表現し、ラベル y（スパム/非スパム）を予測.
  - 学習データには多くのメールとそのラベルが含まれる.
  - 学習済みの g を使い、新しいメールがスパムかどうかを判定できる.
- このような問題では、条件付き確率密度 $f(y|\boldsymbol{x})$ が分かれば、最適な予測関数 $g^*(\boldsymbol{x})$ を理論的に求めることができる.

2.2 Supervised Learning

- Unspervised Learning(教師なし学習) では、応答変数（ラベル）は与えられず、データ $\boldsymbol{x}$ の構造や分布を学習することが目的.
- 例：スーパーの顧客購買行動の分析
  - 100 種類の商品について、各顧客が購入したかどうかを 0/1 のベクトル $\boldsymbol{x} \in \{0,1\}^{100}$ で表現.
  - ラベルは存在せず、顧客の購買パターンやクラスタを自動的に発見する.
  - 例：似た購買傾向を持つ顧客グループの発見、異常な購買パターンの検出など.
- 教師なし学習では、$f(\boldsymbol{x})$（データの分布）を推定し、$g(\boldsymbol{x})$ はその近似となる.
- リスクは $\ell(g) = \mathbb{E}[\mathrm{Loss}(f(\boldsymbol{X}), g(\boldsymbol{X}))]$ で評価される.

2.2 Supervised Learning

教師なし学習の主な手法として

- クラスタリング：データを似たグループに分ける
- 主成分分析（PCA）：高次元データを低次元に圧縮し、特徴を抽出
- カーネル密度推定：データの分布をなめらかに推定

がある．これらの手法は第 4 章で詳しく説明される．

今後の章の内容

- 次の章では主に教師あり学習（回帰・分類）に焦点を当てて解説
- 主な手法：回帰（regression）、分類（classification）
- より高度な手法（再生カーネル Hilbert 空間、決定木、深層学習など）は第 6, 8, 9 章で扱う