

統計的機械学習論

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)

1

クラスの進行

- ガイダンス
 - この講義の達成目標
 - クラスの進行
 - 講義スケジュール
 - 成績評価について
- 1.はじめに
- 1.1.データサイエンスと機械学習
 - 1.1.1.データサイエンティストの役割
 - 1.1.2.データサイエンスの全体像
 - 1.2.アルゴリズムの分類
 - 1.3.使用するデータの紹介
 - 1.4.分析ツールについて
 - 1.5.数学的な記述のおさらい

3

この講義の達成目標

達成目標

- 機械学習アルゴリズムの基礎を理解すること
- データのモデル化とパラメータの最適化の手続きを理解する
- データサイエンスにおける機械学習の役割を理解する

教科書・参考書

- 「機械学習理論入門」中井悦司著 技術評論社
- 「自然科学の統計学」東京大学出版会
- 「kerasによるディープラーニング」F. Chollet (著)マイナビ出版

2

講義スケジュール

- 09/02 : 3限「ガイダンス」
- 09/03 : 3限「線形回帰アルゴリズムの基礎：最小二乗法」
- 09/03 : 4限「最尤推定法：確率を用いた推定理論」
- 09/04 : 4限「分類アルゴリズムの基礎：パーセプトロン」
- 09/05 : 3限「ロジスティック回帰とROC曲線」
- 09/05 : 4限「演習：線形判別による新規データの分類」
- 09/06 : 3限「教師なし学習モデルの基礎：k平均法」
- 09/06 : 4限「k近傍法」
- 09/09 : 3限「EMアルゴリズム」
- 09/10 : 3限「ベイズ推定とベイズの定理」「ベイズ推定の回帰分析への応用」
- 09/10 : 4限「ベイズ推定の演習」
- 09/11 : 3限「畳み込みニューラルネットワークを用いた機械学習について」
- 09/12 : 3限「kerasを用いたディープラーニング1」
- 09/12 : 4限「kerasを用いたディープラーニング2」
- 09/13 : 確認テスト

4

成績評価について

ミニッツペーパー（4割）

+

最後の授業での確認テスト（6割）

5

1.1 データサイエンスの役割

データサイエンスの目的：

データを活用してより質の高い判断を行うこと



過去のデータに含まれる事実を抽出

+

仮説を立て、検証し、未来予測することができる

7

1. データサイエンスと機械学習

データサイエンスにおける機械学習の役割を理解する

どのような仕組み／考え方でデータ分析が行われるか

データのモデル化とパラメータの最適化の手続き

機械学習アルゴリズムによる問題解決の様子を観察する

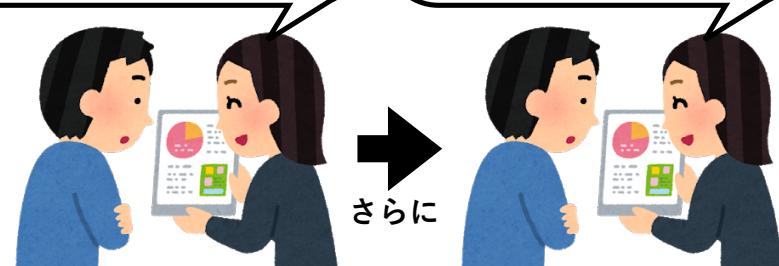
6

1.1.1 データサイエンティストの役割

期待されていること（売り上げ予測の例）：

最高気温が32°Cを超える日はミネラルウォーターの売り上げが30%増加しました

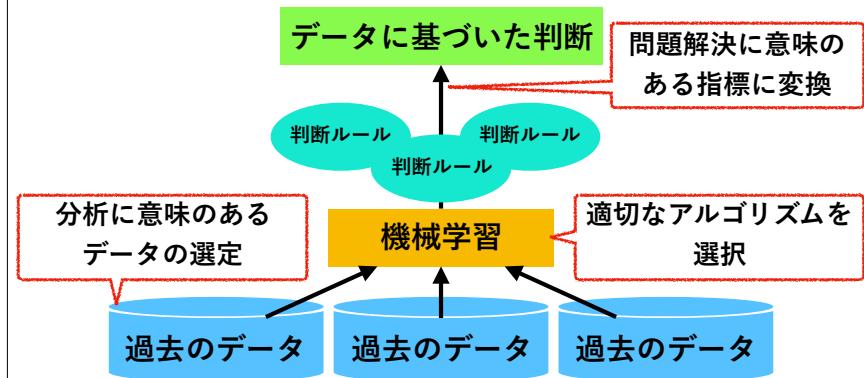
最高気温が32°Cを超える地域の店舗でXX%在庫を増やせばYY%利益が向上します



8

1.1.2 データサイエンスの全体像

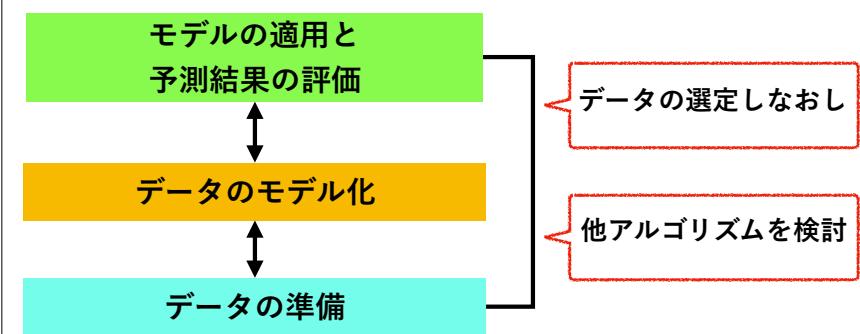
機械学習を使えばさまざまな判断ルールを生み出せるが
それらが全て未来予測に役立つわけではない



9

1.1.2 データサイエンスの全体像

未来予測のためには、仮説の構築、データのモデル化、
モデルの検証を繰り返さなければならない



10

判断の質を向上させるためには…

- 課題を理解し適切なアルゴリズムを選ぶ
→ 機械学習アルゴリズムの正確な理解
- 分析に利用すべきデータを選別できる
→ 必要に応じたデータ収集の提案



この講義ではアルゴリズムの基本部分を紹介



さらに高度な機械学習へ

11

1.2 機械学習アルゴリズムの分類

- 分類
複数のクラスに分類された既存のデータを元に新規データがどのクラスに属するかを予測する（スパムメールの判定）
- 回帰
既存のデータからデータの値を決定する何らかの関数を推測し次に得られるデータの値を予測する（売り上げ目標と広告宣伝費）
- クラスタリング
教師データなしでグループ化する（顧客のグループ化と特徴づける要素の発見）

教師あり学習

教師なし学習

12

1.2.4 その他のアルゴリズム

この講義では取り上げないもの：

- 類似マッチング
新しく得られたデータが既存のどのデータと類似しているか判定
(既存の優良顧客と類似した顧客の発見)
- 共起分析(アソシエーション分析・アフィニティー分析)
既存のデータから同時に発生する事象を発見する
(Aを買った人はBも買っています)
- リンク予測
データ間の潜在的なつながりを予測する
(SNSにおける人間関係のつながり)

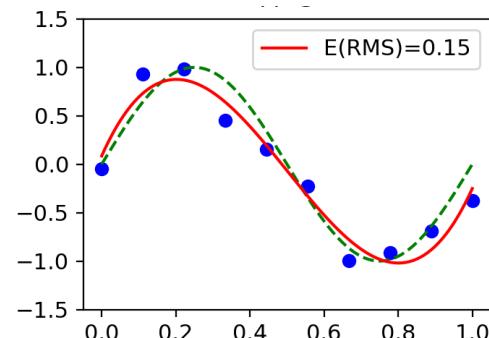
13

1.2.2 回帰分析

既存データの背後に何らかの関数が隠れていると考える：

最小二乗法 最尤推定法 ベイズ推定

具体例：売り上げ目標に応じた広告費の設定



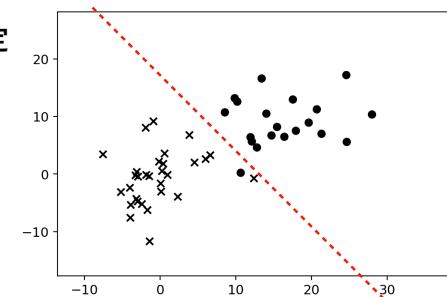
15

1.2.1 分類

過去のデータをもとに新規データがどのクラスに属するかを予測する：

パーセプトロン ロジスティック回帰

具体例：スパムメールの判定



14

1.2.3 クラスタリング

問題の答えがわかっていない（教師なし）の状態から何らかの解析で自然に形成されているグループを見つけ出す。

類似性の高いデータごとにグループを分けることで特徴づける要素を発見できることもあるため、特徴ベクトルを見つけることにも使われることが多い：

k平均法 k近傍法 EMアルゴリズム

具体例：手書き文字の分類、画像の減色処理

16

主な数学記号と基本公式

- 和の記号

▶ 記号 Σ は和を表す。

$$\sum_{n=1}^N x_n = x_1 + x_2 + \dots + x_N$$

- 積の記号

▶ 記号 Π は積を表す。

$$\prod_{n=1}^N x_n = x_1 \times x_2 \times \dots \times x_N$$

17

主な数学記号と基本公式

- 指数関数

▶ 記号 \exp は自然対数の底 $e \approx 2.718$ を用いた指数関数を表す。

$$\exp x = e^x$$

▶ 指数関数の積は引数の和に変換される

$$\prod_{n=1}^N e^{x_n} = e^{x_1} \times e^{x_2} \times \dots \times e^{x_N} = \exp \left\{ \sum_{n=1}^N x_n \right\}$$

▶ 指数関数 e^x は、微分しても関数が変化しない

$$\frac{d}{dx} e^x = e^x$$

18

主な数学記号と基本公式

- 対数関数

▶ 記号 \ln は自然対数の底 $e \approx 2.718$ を用いた対数関数を表す。

$$\ln x = \log_e x$$

$$\ln e = 1$$

▶ 対数関数は次の法則を満たす。

$$\ln \frac{ab}{c} = \ln a + \ln b - \ln c$$

$$\ln a^b = b \ln a$$

▶ このことから、指数関数を対数関数に代入すると式が簡単になる。

$$\ln \left(\exp \sum_{n=1}^N x_n \right) = \sum_{n=1}^N x_n \times \ln e = \sum_{n=1}^N x_n$$

▶ 対数関数の微分系

$$\frac{d}{dx} \ln x = \frac{1}{x}$$

主な数学記号と基本公式

- 偏微分

複数の変数を持つ関数について、特定の変数で微分することを偏微分と呼ぶ

▶ y を固定して x で微分する

$$\frac{\partial f(x, y)}{\partial x}$$

▶ x を固定して y で微分する

$$\frac{\partial f(x, y)}{\partial y}$$

▶ 偏微分の合成関数微分の公式

$$\frac{\partial f(g(x, y))}{\partial x} = f'(g(x, y)) \times \frac{\partial g(x, y)}{\partial x}$$

$$f'(x) = \frac{\partial f(x)}{\partial x}$$

19

20

主な数学記号と基本公式

- ベクトルの内積と直積

▶ 数式中の太字の変数はベクトルを表し基本的には縦ベクトルと考える。

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}$$

▶ 表記の都合上、横ベクトルで記載する時は、転置記号Tを用いて縦ベクトルであることを示す。

$$\mathbf{x} = (x_1, x_2, x_3)^T$$

21

主な数学記号と基本公式

- ベクトルの内積と直積

▶ ベクトルとして扱うと特定の成分について偏微分することが可能になる。

$$\frac{\partial f(\mathbf{w}^T \mathbf{x})}{\partial w_i} = f'(\mathbf{w}^T \mathbf{x}) \times \frac{\partial \mathbf{w}^T \mathbf{x}}{\partial w_i} = f'(\mathbf{w}^T \mathbf{x}) x_i$$

▶ ベクトルの大きさは次の記号で表す。

$$\|\mathbf{x}\| = \sqrt{\mathbf{x}^T \mathbf{x}} = \sqrt{x_1^2 + x_2^2 + x_3^2}$$

23

主な数学記号と基本公式

- ベクトルの内積と直積

▶ 「横ベクトル」「縦ベクトル」の積は内積を表す。

$$\mathbf{w}^T \mathbf{x} = (w_1, w_2, w_3) \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \sum_{i=1}^3 w_i x_i$$

▶ 「縦ベクトル」「横ベクトル」の積は直積を表す。

$$\mathbf{w} \mathbf{x}^T = \begin{pmatrix} w_1 \\ w_2 \\ w_3 \end{pmatrix} (x_1, x_2, x_3) = \begin{pmatrix} w_1 x_1 & w_1 x_2 & w_1 x_3 \\ w_2 x_1 & w_2 x_2 & w_2 x_3 \\ w_3 x_1 & w_3 x_2 & w_3 x_3 \end{pmatrix}$$

22

主な数学記号と基本公式

- 確率変数の期待値と分散

確率的に様々な値をとる変数Xを確率変数と呼び、X=xである確率(Probability)をP(x)として表す。確率変数の期待値Eと分散Vは以下のようない定義する。

$$E[X] = \sum_x x P(x)$$

$$V[X] = E[(X - E(X))^2]$$

▶ 平均と分散に関しては以下の公式が成立する。

$$E[aX + b] = aE[X] + b$$

$$V[aX] = a^2 V[X]$$

$$V[X] = E[X^2] - (E[X])^2$$

$$E[X - \bar{x}] = E[X] - \bar{x} = 0$$

24

主な数学記号と基本公式

- 同時確率

- 2つの確率変数XとYが独立であるとき、 $X=x$ かつ $Y=y$ となる確率 $P(x, y)$ はそれぞれの確率の積で表される。

$$P(x, y) = P_X(x) \times P_Y(y)$$

- 2つの確率変数XとYが独立であるとき、以下の式が成立する。

$$V[X] = E[\{X - E(X)\}^2]$$

$$\begin{aligned} E[(X - \bar{x})(Y - \bar{y})] &= \sum_{x,y} (x - \bar{x})(y - \bar{y})P(x, y) \\ &= \sum_x (x - \bar{x})P_X(x) \sum_y (y - \bar{y})P_Y(y) \\ &= E[X - \bar{x}]E[Y - \bar{y}] = 0 \end{aligned}$$

$$\bar{x} = E[X], \bar{y} = E[Y]$$

25

実行環境の整備

Python :

- NumPy
- SciPy
- matplotlib
- pandas
- PIL(→PILLOW)
- keras

27

主な数学記号と基本公式

- 手書きのベクトル表記にも慣れておこう

A	B	C	D	E	F	G	a	b	c	d	e	f	g
H	I	J	K	L	M	N	h	i	j	k	l	m	n
O	P	Q	R	S	T	U	o	p	q	r	s	t	u
V	W	X	Y	Z			v	w	x	y	z		

26

計算サーバを利用しよう

アドレス : 131.206.57.50

OS : Ubuntu 18.04.1 LTS

hostname : garlic

28

まとめ

- データサイエンスにおける機械学習の役割を理解した
- 講義で紹介するアルゴリズムの概要と分類を把握した
- 講義で使われるデータの特徴を理解した
- 頻出する数学記号と基本公式を復習した

29

ミニッツペーパー

- データサイエンスの目的を簡潔に説明してください
過去のデータを分析することで、モデル（仮説）の構築し検証することで質の高い判断（未来予測）ができる
- 講義で紹介するアルゴリズムについて、大まかに3つに大別してその名称をあげなさい
分類、回帰、クラスタリング
- 上記のアルゴリズムについて、教師あり・なしの観点から2つに大別し示しなさい
教師あり：分類と回帰、教師なし：クラスタリング
- 授業中に紹介した数学記号と基本公式で知らなかったものがあれば教えてください。
（評価には関係のないアンケートです）
- 機械学習アルゴリズムのうち、すでに知っているものがあれば教えてください。
（評価には関係のないアンケートです）

30