1　データサイエンスII 第１回　序

1.1　今日の予定

- 講義の形式
- データサイエンスの基本問題
- 今日の例：平面の N 点を通る p 次代数曲線
- 演習１：$N = 5, p = 2$（５点を通る二次曲線）
  - jupyter, python の復習、numpy の練習
- 演習２：$N \geq 6, p = 2$（６点以上を通る近似二次曲線）
- 演習３：$p = 3$（三次曲線）
- レポート：jupyter notebook 提出
- アンケート

1.2　クラウド環境についての注意

- 作業を続けているときは４時間まで使える。
- 作業を中断し、ページが表示されていない時間が１０分を越すと、サーバが終了し、データが消失する
- 作業を中断するときは、作業中の notebook をダウンロードすることが必要.

1.3　次回までにしておくこと

- 対面授業を受けない人は、自分の PC で jupyter notebook が使えるようにする。

1.4　各回の流れ

- 前半に講義：テーマの解説。例への適用
- 後半に演習：資料の notebook で実習。質疑は zoom で。

1.5 各回の資料

- jupyter notebook(manaba コンテンツ）
- 演習用 補助動画資料 (Panopto 動画）
- 授業録画

1.6 提出物

- 各回
  - 実習済 jupyter notebook を Manaba report にアップロード
  - アンケート回答
- オンライン演習を若干
- 最終課題

1.7 コミュニケーション

- 授業中の質疑（Zoom ではチャット、ブレークアウトルーム）
- メール 21ds2@tjst.ac-net.org
- Manaba アンケートでの質疑
- Manaba 掲示板 Q&A スレッド

1.8 データサイエンス II の概要（シラバスへ）

1.9 「データサイエンス」とは

- 問題 1. データ集合 $D \subset X$ を理解し、$x \in X$ が与えられた時、x は D と同類かどうか判定せよ.
  - 「理解せよ」= 内包的記述（特徴づけ）
  - D は有限集合だが、最近は「ビッグデータ」で数万から数億。
  - X は属性の組 $X_1 \times X_2 \times \cdots \times X_k$
    - 画像や時系列の場合は、k も巨大（百万）

1.10　集合の与え方

- 外延的：$X = \{ a, b, c, d, e \}$
  - 外延性公理：集合は要素で決まる.

$$A = B \Longleftrightarrow \forall x : x \in A \Longleftrightarrow x \in B$$

- 内包的：$X = \{ x \mid P(x) \}$.

1.11　問題 1 を数学の問題にするのに必要なこと

- 部分集合族を指定しないと問題にならない。
- 問題 2: 部分集合族 $\mathscr{A} = \{ A_i \subset X \mid i \in I \}$ を指定した時、
  - 与えられたデータ集合 $D \subset X$ を含む部分集合 A_i を求めよ。
  - 可能性は３通り.
    - 複数存在
    - 唯一存在
    - 存在しない

1.12　特殊化: 問題 2A

- $D \subset X$ の特性写像 $\chi_D : X \to \{ 0, 1 \}$ を求めよ. すなわち
  $\chi_D(x) = 1 \iff x$ が D の要素と似ている

1.13　特殊化: 問題 2B

- 問題 2A の一般化: 部分集合族 $\{ D_v \mid v \in V \}$ が与えられた時、
  次のような分類写像 $\chi : X \to V$ を求めよ：
  $\chi(x) = v \iff x$ が D_v の要素に似ている

1.14　特殊化: 問題 2C 回帰（関数の推定）

- 問題 2B の一般化 $D = \{ (x_i, y_i) \in X \times V \mid i \in [1..N] \}$ が与えられた時、
  次のような写像 $F : X \to V$ を探す:
  $\forall i : F(x_i) \approx y_i$

1.15 問題 3

- $D = \coprod_i D_i$ と分割し、D_i の中の要素は似ているようにする。

1.16 可視化

- 1 次元：ヒストグラム、密度関数
- ２次元：散布図、密度関数 (heat map)
- ３次元：散布図

1.17 外延から内包を求める数学の問題の例

- 問題 A：平面上の与えられた N 点を通る p 次代数曲線を求めよ.
- 解決可能性: p 次以下の単項数 K_p は $K_p = 1 + 2 + 3 + \cdots + p + 1 = \frac{(p+1)(p+2)}{2}$

[]@cccccc@ p 1 2 3 4 5
K_p – 1 2 5 9 14 20

- $N < K_p$ の時は、解が存在する
- $N \geq K_p$ の時は、一般には存在しない。

1.18 データサイエンスは $N \geq K_p$ の時が主.

- 問題. 最良近似曲線を求めよ.

1.19 $N = 5$, $p = 2$ の時

- $I = \mathbb{R}^6 \setminus \mathbf{0}$
- $A_{\mathbf{k}} = \{ (x, y) \in \mathbb{R}^2 \mid f_{\mathbf{k}}(x, y) = 0 \}$, ただし $f_{\mathbf{k}}(x, y) := k_0x^2 + k_1xy + k_2y^2 + k_3x + k_4y + k_5$
- 問題：平面上の点集合 $D = \{ (x_i, y_i) \mid 0 \leq i < N \}$ を通る二次曲線を求めよ.
  - $N < 5$ なら無数にある
  - $N = 5$ なら通常は唯一
  - $N > 5$ なら通常は存在しない

1.20 $N \leq 5$ の場合の解法

- D は $(N, 2)$ 行列
- $f_{\mathbf{k}}(x, y) = [x^2, xy, y^2, x, y, 1]\mathbf{k}$
- $f_{\mathbf{k}}(x_i, y_i) = 0$ は　$D_2\mathbf{k} = 0$, ただし $D_2 = [x_i^2, x_i y_i, y_i^2, x_i, y_i, 1]_{0 \leq i < N}$
- D_2 は $(N, 6)$ 行列で、$N < 6$ の時 0 でない解がある。
- ${}^tD_2 D_2\mathbf{k} = 0$ と書き換えると、6 次正方行列 $B := {}^tD_2 D_2$ のゼロ固有ベクトル $\mathbf{k}$ が求めるもの。

1.21 $|D| > 5$ の場合の解法

- 最もよく近似する二次曲線を求めよ
  - 解は「近似の度合」の尺度が重要
  - ここでは　$\varepsilon(\mathbf{k}) := \sum_i f_{\mathbf{k}}(x_i, y_i)^2$ が最小となるような $\mathbf{k} \in S^5$ を求める.
- $\varepsilon(\mathbf{k}) = (D_2\mathbf{k}, D_2\mathbf{k}) = (B\mathbf{k}, \mathbf{k})$ は $\mathbf{k}$ の二次形式。

1.22 復習：二次形式の最小値

- 実対称行列は直交行列で対角化でき、固有値は実数。
  固有値を $\lambda_1 \leq \cdots \leq \lambda_6$ とすると
  - $\|\mathbf{k}\| = 1$ の時 $\lambda_1 \leq (B\mathbf{k}, \mathbf{k}) \leq \lambda_6$
  - $\varepsilon(\mathbf{k}) = \lambda_1 \iff \mathbf{k}$ が λ_1 固有空間に属する

1.23 以上を jupyter で計算

- クラウド：データサイエンス II

1.24 演習１ $N = 5$ の場合

1.25 演習２ $N > 5$ の場合

1.26 演習３三次曲線