第 3 章【学習 11】演習解答

●演習 1

（解答例）

アとイの表現の違い：

縦軸の目盛りが異なる。アは 200 円から始まっており，イは 0 円から始まっている。このため，アのグラフでは 30 歳未満の世帯では漬物を購入していないように見える。

グラフ作成の際に注意すべき点：

縦軸の目盛幅は適切か，ビン幅（階級の幅）は適切か，表現するグラフの種類は適切か。

●演習 2

（解答例）

所得

年齢

独身か 2 人世帯か 3 人以上世帯か

●演習 3

（解答例）

出席番号，Int 型

氏名，String 型

得点，Int 型

得点率，Real 型

●演習 4

（解答例）

TestcsvimportD.ipynb

（読み込んだデータの表示）手順書にしたがってドライブにマウント後，演習 4 の前の本文のコードを実行

```
import pandas as pd

df = pd.read_csv('/content/drive/My Drive/Colab Notebooks/Testdata.csv')#f)#.values.tolist()
 # dfの1行目ヘッダーあり
print(df)
```

（実行結果）

	PlayerNo	Score	Accuracy
0	13	16930	1.00
1	11	15110	0.95
2	37	11300	0.79
3	13	16930	1.00
4	19	10650	0.74
5	1	10910	0.74
6	33	9430	0.74
7	11	15110	0.95
8	26	8950	0.68
9	11	15110	0.95
10	17	8920	0.68
11	21	9950	0.68
12	25	9950	0.68
13	7	9850	0.68
14	3	8600	0.63
15	28	9120	0.63
16	38	8460	0.63
17	2	7760	0.58
18	23	8260	0.58
19	16	7860	0.58
20	30	8430	0.58
21	27	7780	0.58
22	32	7740	0.58
23	18	8120	0.58
24	22	7550	0.58
25	34	7440	0.53
26	39	6340	0.53
27	10	6030	0.47
28	20	6310	0.47
29	8	6620	0.47
30	12	6480	0.47
31	15	5390	0.42
32	35	6170	0.42
33	4	7180	0.42
34	36	4940	0.37
35	24	5960	0.37
36	9	5080	0.37
37	5	6030	0.37
38	31	5010	0.37
39	29	4650	0.32
40	14	4850	0.32
41	6	4570	0.32

（データの削除）

```
#演習4　dataflameから特定の行(0行目)を除く
datadroped = df.drop(0,axis=0)
print(datadroped)

# 追加課題
#重複したデータ（全く同じデータ）がデータベースに格納されていないか確認
print(df[df.duplicated()])
#PlayerNoが重複したデータをさがすこともできる
print(df[df.duplicated(subset='PlayerNo')])
# dfから重複データを削除(一つ目を残す))
df.drop_duplicates(subset='PlayerNo',keep='first',inplace=True)
print(df)
```

（実行結果）

	PlayerNo	Score	Accuracy
1	11	15110	0.95
2	37	11300	0.79
3	13	16930	1.00
4	19	10650	0.74
5	1	10910	0.74
6	33	9430	0.74
7	11	15110	0.95
8	26	8950	0.68
9	11	15110	0.95
10	17	8920	0.68
11	21	9950	0.68

●演習 5

（解答例）

（1）興味を持った分野：

　　疫学・予防医学

（2）つながりを見つけた分野：

　　統計科学，衛生学，獣医学，食生活学，ウイルス学，救急医学など

●演習 6

（解答）

演習本文自体が解答となっています。

＜更に詳しく学びたい人へ参考に＞

・演習 5 の学問分野の関連検索は以下の方が詳細である。

　研究者キーワード相関図（2016 年 3 月 18 日時点のデータより）

　京都大学　学際融合教育研究推進センター（http://www.cpier.kyoto-u.ac.jp/keyword-map/20160318/）

　授業では，研修で扱った Schola Scope と Your Schola 診断を生徒の進路研究と関連づけて取り入れる方法もある。

・ドキュメント型データベースにはクラウド環境で試すことができる Mongo Atlas もある。

・グラフ型データベースについては Neo4j がある。Neo4j 公式サイト（https://neo4j.com）ではドキュメンテーション以外に，グラフ型データベースに関する無料の学習教材（英語，コース修了証も発行される）も提供されている。

　いずれもオープンソース。

・データ保持の形式には本文で紹介した以外もある。一例として，LOD（Linked Open Data）を挙げておく。

　The Linked Open Data Cloud（https://lod-cloud.net）や DBpedia（http://ja.dbpedia.org）でその形式を体験することができる。

第 3 章【学習 12】演習解答

●演習 1

ブラウザを用いて USGS からデータを取得するには，WebAPI のエンドポイントの URL にアクセスすればよい。ここでは例として，2019 年 12 月 1 日から 2 日に発生した地震のデータを GeoJSON 形式で取得するためのエンドポイントを例として載せる。

https://earthquake.usgs.gov/fdsnws/event/1/query?format=geojson&starttime=2019-12-01&endtime=2019-12-02

これにより，次のようなデータが得られる。


```
{"type":"FeatureCollection","metadata":
{"generated":1581855330000,"url":"https://earthquake.usgs.gov/fdsnws/event/1/query?format=geojson&starttime=2019-12-01&endtime=2019-12-02","title":"USGS Earthquakes","status":200,"api":"1.8.1","count":379},"features":
[{"type":"Feature","properties":{"mag":1.1799999999999999,"place":"8km NNW of Redwood Valley,
CA","time":1575244115740,"updated":1575408903432,"tz":-480,"url":"https://earthquake.usgs.gov/earthquakes/eventpage/nc73310796","detail":"https://earthquake.usgs.gov/fdsnws/event/1/query?
eventid=nc73310796&format=geojson","felt":null,"cdi":null,"mmi":null,"alert":null,"status":"reviewed","tsunami":0
,"sig":21,"net":"nc","code":"73310796","ids":",nc73310796,","sources":",nc,","types":",geoserve,nearby-
cities,origin,phase-data,scitech-
link,","nst":8,"dmin":0.1356,"rms":0.059999999999999998,"gap":271,"magType":"md","type":"earthquake","title":"M
1.2 - 8km NNW of Redwood Valley, CA"},"geometry":{"type":"Point","coordinates":
[-123.2415,39.3333333,6.29]},"id":"nc73310796"},
{"type":"Feature","properties":{"mag":1.8,"place":"3km SE of Newport, Rhode
Island","time":1575244011621,"updated":1580876392765,"tz":-300,"url":"https://earthquake.usgs.gov/earthquakes/eventpage/us70006f5x","detail":"https://earthquake.usgs.gov/fdsnws/event/1/query?
eventid=us70006f5x&format=geojson","felt":11,"cdi":3.2999999999999998,"mmi":null,"alert":null,"status":"reviewed"
,"tsunami":0,"sig":53,"net":"us","code":"70006f5x","ids":",us70006f5x,","sources":",us,","types":",dyfi,geoserve,
origin,phase-
data,","nst":null,"dmin":0.48899999999999999,"rms":0.31,"gap":181,"magType":"ml","type":"earthquake","title":"M
1.8 - 3km SE of Newport, Rhode Island"},"geometry":{"type":"Point","coordinates":
[-71.2851,41.465899999999998,5]},"id":"us70006f5x"},
{"type":"Feature","properties":{"mag":4.7000000000000002,"place":"237km SE of Amahai,
Indonesia","time":1575243860836,"updated":1578622057040,"tz":540,"url":"https://earthquake.usgs.gov/earthquakes/eventpage/us70006f38","detail":"https://earthquake.usgs.gov/fdsnws/event/1/query?
eventid=us70006f38&format=geojson","felt":null,"cdi":null,"mmi":null,"alert":null,"status":"reviewed","tsunami":0
,"sig":340,"net":"us","code":"70006f38","ids":",us70006f38,","sources":",us,","types":",geoserve,origin,phase-
data,","nst":null,"dmin":2.2389999999999999,"rms":0.78000000000000003,"gap":41,"magType":"mb","type":"earthquake"
,"title":"M 4.7 - 237km SE of Amahai, Indonesia"},"geometry":{"type":"Point","coordinates":
[130.68600000000001,-4.5366,61.810000000000002]},"id":"us70006f38"},
{"type":"Feature","properties":{"mag":1.3,"place":"39km WSW of Sandy Valley,
Nevada","time":1575243645009,"updated":1575327155494,"tz":-480,"url":"https://earthquake.usgs.gov/earthquakes/eventpage/nn00712109","detail":"https://earthquake.usgs.gov/fdsnws/event/1/query?
eventid=nn00712109&format=geojson","felt":null,"cdi":null,"mmi":null,"alert":null,"status":"reviewed","tsunami":0
,"sig":26,"net":"nn","code":"00712109","ids":",nn00712109,","sources":",nn,","types":",geoserve,origin,phase-
data,","nst":13,"dmin":0.251,"rms":0.15210000000000001,"gap":132.25999999999999,"magType":"ml","type":"earthquake
","title":"M 1.3 - 39km WSW of Sandy Valley, Nevada"},"geometry":{"type":"Point","coordinates":
[-116.0271,35.671900000000001,0]},"id":"nn00712109"},
```


図表 1　USGS から取得した地震データの一部

データを取得し，表示するプログラムは次のようになる。

```
import json
import requests
import datetime

url = 'https://earthquake.usgs.gov/fdsnws/event/1/query?'
param = {
      'format': 'geojson',
      'starttime': '2019-12-01',
      'endtime': '2019-12-02'
}

r = requests.get(url, params=param)
r_json = r.json()
print(json.dumps(r_json, indent=2))
```

このプログラムの実行結果は次のようになる（地震データは1件だけ掲載）。


```
{
  "type": "FeatureCollection",
  "metadata": {
    "generated": 1581860124000,
    "url":       "https://earthquake.usgs.gov/fdsnws/event/1/query?format=geojson&starttime=2019-12-01&endtime=2019-12-02",
    "title": "USGS Earthquakes",
    "status": 200,
    "api": "1.8.1",
    "count": 379
  },
  "features": [
    {
      "type": "Feature",
      "properties": {
        "mag": 1.18,
        "place": "8km NNW of Redwood Valley, CA",
        "time": 1575244115740,
        "updated": 1575408903432,
        "tz": -480,
        "url": "https://earthquake.usgs.gov/earthquakes/eventpage/nc73310796",
```

```
            "detail":
"https://earthquake.usgs.gov/fdsnws/event/1/query?eventid=nc73310796&format=geojson",
            "felt": null,
            "cdi": null,
            "mmi": null,
            "alert": null,
            "status": "reviewed",
            "tsunami": 0,
            "sig": 21,
            "net": "nc",
            "code": "73310796",
            "ids": ",nc73310796,",
            "sources": ",nc,",
            "types": ",geoserve,nearby-cities,origin,phase-data,scitech-link,",
            "nst": 8,
            "dmin": 0.1356,
            "rms": 0.06,
            "gap": 271,
            "magType": "md",
            "type": "earthquake",
            "title": "M 1.2 - 8km NNW of Redwood Valley, CA"
        },
        "geometry": {
            "type": "Point",
            "coordinates": [
                -123.2415,
                39.3333333,
                6.29
            ]
        },
        "id": "nc73310796"
    },
（以下省略）
```


●演習 2

文部科学省の新着情報のページから掲載日ごとに項目名を取得するプログラムは次のようになる。

```
import requests
from bs4 import BeautifulSoup
url = 'https://www.mext.go.jp/b_menu/news/index.html'
r = requests.get( url )
soup = BeautifulSoup( r.content, 'html.parser' )
date = soup.find_all( 'h3', 'information-date' )
links = soup.find_all( 'ul', 'news_list' )
for d, l in zip(date, links):
    titles = l.find_all('a')
    for t in titles:
        print( d.string, t.string, sep=',' )
```

本文では項目名だけが表示されるプログラムであるが，解答では日付も表示するプログラムに変更している。このプログラムを実行することにより，次のように表示される。（実行日を含む1ヶ月分が表示される）

令和 2 年 2 月 14 日，研究環境基盤部会　共同利用・共同研究拠点及び国際共同利用・共同研究拠点に関する作業部会（第 10 期）（第 4 回）　配付資料
令和 2 年 2 月 14 日，宇宙開発利用部会（第 53 回）の開催について
令和 2 年 2 月 14 日，宇宙開発利用部会　国際宇宙ステーション・国際宇宙探査小委員会（第 35 回）の開催について
（以下省略）

●演習 3

e-Stat からダウンロードしたデータに対しては，次のような修正を行う。

- 不要な行や列を削除する
- 項目名を修正する
- 数値の表示形式をカンマが付かない形式に修正する

演習 3 のデータにはなかったが，他にも

- 表記を統一する（高校と高等学校，2020 年と令和 2 年など）
- 複数行にまたがっているデータを 1 行ごとに修正する
- 行や列を結合して値をまとめて表示している場合には，それぞれのデータに対して値を持たせる

などの修正が必要になる。

●演習 4

学習 12 演習 4 の本文に示したプログラムにより，本文中の実行結果が得られる。

●演習 5

学習 12 演習 5 の本文に示したプログラムにより，本文中の実行結果が得られる。

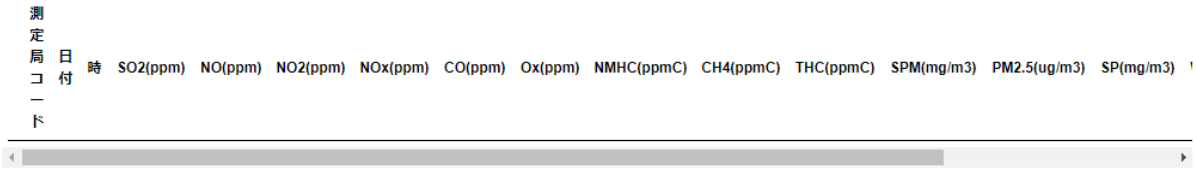
●演習 6

学習 12 演習 6 の本文中ではプログラムを示したが，欠損値の処理結果が掲載されていないためここで示す。

すべての列にデータがあるものを使う場合のプログラムと実行結果

```
df1=df.dropna()
df1
```

Out[2]:



図表 2　欠損値の処理 1

必要な列を選び欠損値がある行を除く場合のプログラムと実行結果

```
df2 = df[['日付','時','NO(ppm)']].dropna()
df2
```

Out[4]:

	日付	時	NO(ppm)
0	2019/12/1	1	0.017
1	2019/12/1	2	0.006
2	2019/12/1	3	0.005
3	2019/12/1	4	0.001
4	2019/12/1	5	0.000
5	2019/12/1	6	0.000
6	2019/12/1	7	0.001
7	2019/12/1	8	0.002

図表 3　欠損値の処理 2

欠損値を 0 として扱う場合のプログラムと実行結果

```
df3 = df.fillna(0)
df3
```

Out[5]:

	測定局コード	日付	時	SO2(ppm)	NO(ppm)	NO2(ppm)	NOx(ppm)	CO(ppm)	Ox(ppm)	NMHC(ppmC)	CH4(ppmC)	THC(ppmC)	SPM(mg/m3)	PM2.5(ug
0	13101010	2019/12/1	1	0.000	0.017	0.041	0.058	0.0	0.001	0.0	0.0	0.0	0.014	
1	13101010	2019/12/1	2	0.000	0.006	0.037	0.043	0.0	0.001	0.0	0.0	0.0	0.013	
2	13101010	2019/12/1	3	0.000	0.005	0.033	0.038	0.0	0.001	0.0	0.0	0.0	0.017	
3	13101010	2019/12/1	4	0.000	0.001	0.021	0.022	0.0	0.007	0.0	0.0	0.0	0.016	
4	13101010	2019/12/1	5	0.000	0.000	0.017	0.017	0.0	0.009	0.0	0.0	0.0	0.015	
5	13101010	2019/12/1	6	0.000	0.000	0.012	0.012	0.0	0.014	0.0	0.0	0.0	0.007	
6	13101010	2019/12/1	7	0.000	0.001	0.012	0.013	0.0	0.016	0.0	0.0	0.0	0.015	

図表 4　欠損値の処理 3

以下については，表示される範囲では欠損値のままになっているので,プログラムのみ示す。

前の値で埋める場合のプログラム

```
df4 = df.fillna(method='ffill')
df4
```

平均値で埋める場合のプログラム

```
df5 = df.fillna(df.mean())
df5
```

第 3 章【学習 13】演習解答

●演習 1

（解答例）

Step.0　特性要因図

「特性要因図」は，問題の解決を目指す手法として QC(Quality Control)7 つ道具の一つに挙げられ，広く世界中で児童生徒から企業の業務課題解決に至るまで活用されている。

問題解決では，「何の（対象）の何を（特性値，特徴量，変数，指標，KPI，KGI）をどうしたいのか？」の問いを明確にすることから始まる。例えば，個人の成績を上げたい場合と，クラス全体の成績を上げたい場合では，当然，取るべきデータ項目（変数）や何で(原因系の変数)何を(結果系の変数)を動かすのか，その方略が違ってくる。

特性要因図の「特性」が結果系の指標 Y，「要因」がその値を変化させる原因系の指標 X となる。特性要因図をグループディスカッションで最初に作成する際には，「特性」も「要因」も具体的な指標（データ）が対応していない「概念」でも構わないが，実際にデータの取得を計画する段階では，その「概念」を具体的な指標に置き換える必要がある。「成績」の場合は，「いつの何の試験の得点」とするなど。作成の手順は，

① Y を決める
② Y に向かって矢線(大骨)をひく
③ 特性に影響する要因（原因）X を洗い出す・・中骨を入れる
④ 更に要因（原因）を深掘りする・・小骨を入れる

- ５－Why

  要因を探る際に，「なぜなぜ問答（５－Why）」といわれる，なぜ Y の値が変わる？ X1 の値が変わるから，なぜ X1 の値が変わる？， X2 の値が変わるから・・・と要因を表層から深層（根本的原因）に迫っていく方法がある。グループディスカッションでの話の切り出し方，話のまとめ方（図の作成）の技法ともなる。

- 5M

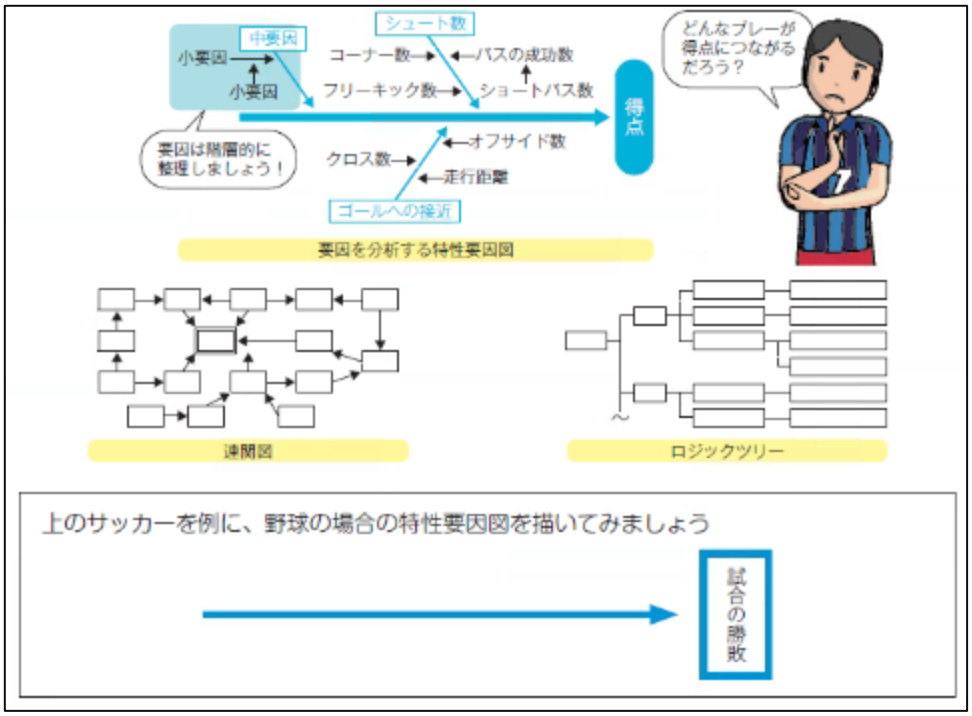
  要因系の変数の洗い出しには，5M と呼ばれる観点から考えることもメーカー系企業ではよく使用されている。5M とは，

  Man ： 人に係わる要因　　Material ： 材料に係わる要因
  Machine ： 道具や機材に係わる要因　　Measure ： 測定に係わる要因
  Method ： 方法，やり方に係わる要因

  である。これに，環境 Environment 要因も含めて，5M+E と呼ばれたりもする。これらは決まったカテゴリーというわけではないので，対象に応じて柔軟に考える必要がある。

Step.1　特性要因図の事例

●対象は？ ⇒ 公式試合
●問題を評価する指標は？　⇒　勝敗
●原因となる要因系の指標は？　対戦相手のチーム特性、イニングごとの結果



出典：総務省「生徒のための統計活用基礎編」
https://www.soumu.go.jp/toukei_toukatsu/info/guide/stkankyo.htm

特性要因図は，魚骨図，フィッシュボーンチャートと言われ，情報処理技術者試験にも出題される内容である。問題例は検索サイトで参照することができる。

同様に，興味のある分野や対象が決まったら，特性要因図やその他の同義語とその分野名をキーワードとして，検索サイトで画像検索してみると，先行事例が参照できる。

- 注意事項

  生徒が作成した分析内容例は，日本統計学会統計教育委員会の教育教材ページから参照できる。また，この研究の中で，要因の影響度の強弱を重回帰分析の出力 t値（検定統計量）の絶対値で行っているが，これは，偏回帰係数または標準編回帰係数の絶対値等で行うのが適切である。検定統計量は有意差のみの判定に利用する。

  特性要因図は研究や分析の設計図であり，正解が決まっているものではない。生徒の探究活動において，このような論理図で思考を表現する習慣づけが知識・情報処理過程で重要である。

●演習2

（解答例）

Step.1 「科学の道具箱」のサイトをチェック。

「理科ねっとわーく」の「デジタル教材」の中に，『算数・数学の資料の活用やデータの分析のための科学の道具箱』，通称，「科学の道具箱」がある。いろいろな自習用コンテンツがあるので，生徒にも紹介してほしい。

https://rika-net.com/contents/cp0530/contents/index.html

「データライブラリー」の中には，右記のコンテンツがある。

1段目の「体力測定データ」を選択。他のデータに関しても，どのような文脈で使用できるのかのストーリーも連動しているので，参照してほしい。

例えば，ここで高校の体力測定データをダウンロードしてみる。下が全体のデータ表で，高校 1 年生から 3 年生までの某県の抽出データ 947 名×21 項目(変数)が入手できる。

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
1	校種・学年	性	身長	体重	握力	垂力	上体起こし	長座体前屈	反復横跳び	シャトルラン	50m走	立ち幅跳び	ハンドボール投げ	握力得点	上体起こし得点	長座体前屈得点	反復横跳び得点	シャトルラン得点	50m走得点	立ち幅跳び得点	ハンドボール投げ得点		
2	高1	男	167.6	56.2	89.8	35	33	55	49	112	7	235	31	5	9	8	6	8	8	7	8		
3	高1	男	157.1	50.5	85.8	33	29	48	57	70	7.4	205	29	5	7	6	8	5	6	5	7		
4	高1	男	165.4	61	85.2	34	31	45	54	76	8	237	22	5	8	6	7	6	4	7	5		
5	高1	男	168	60	91.1	40	31	55	52	76	7.5	225	23	6	8	8	6	6	6	6	5		
6	高1	男	165.9	49	89.7	37	32	62	56	87	7.8	240	26	5	8	9	8	6	5	7	6		
7	高1	男	170	61.5	91.2	36	31	48	55	68	8.2	212	28	5	8	6	7	5	4	5	7		
8	高1	男	168.7	57	92.3	40	42	50	62	102	6.9	240	37	6	10	7	9	8	8	7	10		
9	高1	男	173.1	57.5	91.7	47	38	47	63	95	7.1	270	28	8	10	6	10	7	7	10	7		
10	高1	男	168.2	51.5	90.3	33	35	53	60	88	7.2	245	31	5	10	8	9	6	7	8	8		
11	高1	男	167.3	51	88.6	36	32	55	61	82	7.5	240	31	5	8	8	9	6	6	7	8		
12	高1	男	166.1	49.9	89.3	34	35	41	51	51	7.7	238	14	5	10	5	6	4	5	7	2		
13	高1	男	165.6	64.3	87.6	41	30	61	56	106	6.6	245	17	6	8	9	8	8	10	8	3		
14	高1	男	161.4	52.5	86.7	30	30	40	44	82	7.3	225	18	4	8	5	4	6	6	6	3		

941	高3	女	157	53.1	84.5	30	30	61	48	65	8.5	209	19	6	10	9	8	8	8	7	7		
942	高3	女	146	47.5	79.8	26	30	57	45	54	8.9	180	13	6	10	8	7	7	6	7	5		
943	高3	女	167.6	52.3	88.8	30	31	47	50	55	7.9	230	24	8	10	6	9	7	9	10	10		
944	高3	女	157.6	50.2	85.7	26	28	54	47	64	8.3	212	13	6	9	8	7	8	8	10	5		
945	高3	女	153.2	53.8	86.2	28	33	64	45	73	9	171	12	7	10	10	7	8	5	6	5		
946	高3	女	160.1	48.2	87.2	31	29	55	53	50	7.7	210	21	8	10	8	10	6	10	10	9		
947	高3	女	164.1	55.3	88.9	28	33	54	54	73	8.5	210	26	7	10	8	10	8	7	10	10		
948	高3	女	147	47.5	81	22	29	46	49	51	9.3	175	14	4	10	6	8	6	5	6	6		
949	高3	女	151.5	52.7	80.4	26	32	55	48	66	8.2	210	18	6	10	8	8	8	8	10	8		
950																							
951																							
952																							

Step.2　分析の前準備

データの項目(変数)に関して，体力測定の項目の情報をインターネットで調べたりして，測定単位やどのようにして測定されているのかなど，データの背景情報をあらかじめ調べる習慣をつけることが大切である。また，「文部科学省の体力テスト実施要領」などを参考に，データ収集の目的や集計結果の公表形態，活用の仕方なども調べ，同時に，最近の児童・生徒・成人の体力に関する傾向や課題，どのようなことがメディアで問題として捉えられているのかなどを調べる機会を設け，そのもととなる個票（生徒個人個人の匿名化されたデータ，マイクロデータ）データを分析することの関心を高めることも大切である。

出典　放送大学「データサイエンス基礎から応用」第 1 回スライドより

Step.3　分析ための仮説（リサーチクエッション）を想定・先行研究を調べる

今回の演習は，重回帰分析を使って予測モデルを構築することではあるが，なぜ，その予測をすることに意味があるのか，など，常に「探究」の方法論であることの位置付けを明確にしておく必要がある。

体力測定など，テーマと重回帰分析のキーワードをかけて，Google Scholar （論文など学術文献用の検索 https://scholar.google.com/）等で先行研究を調べると，重回帰分析でどのようなことをすると，課題研究になるのかが分かる。また，論文で分析結果の出力をどのように読み解いて記述するのかも分かる。

例えば、Google Scholar で、「体力測定　重回帰分析　高校生」を検索すると、
いくつかの項目の中に、「高校生自転車競技選手の 1kmTT の記録を推定する体力テストの検討」がヒットし、クリックすると、下記の内容が記載されている。

＊研究の背景
自転車競技選手にとって，1kmTT(タイムトライアル)の記録は，短距離から長距離の競技種目を問わずパフォーマンスを評価するための指標として用いられているが，高校生年代では 1kmTT 記録と関連する体力テスト項目の検討はほとんどなされていない。

＊研究の目的
高校生自転車競技選手の 1kmTT の記録を推定する体力テスト項目の検討を行うこと。

＊研究の方法
全国大会入賞レベルを含む青森県内の自転車競技部に所属する男子高校生 28 名を対象に，1kmTT の記録，体力測定の項目から，握力，背筋力，膝伸展及び屈曲筋力，垂直跳び，最大無酸素パワー及び平均パワーを測定し，相関分析，重回帰分析を行う。

＊分析の結果
1）1kmTT の記録と握力，背筋力，膝伸展及び屈曲筋力，垂直跳び，最大無酸素パワー及び平均パワーとの間に有意な負の相関関係が見られた。
2）重回帰分析の結果，1kmTT の記録は体重あたりに発揮できる平均パワー及び握力の 2 項目で推定できる可能性があることが示された。

＊結論
高校生自転車競技選手において，体重あたりに発揮できる平均パワー及び握力を測定することで 1kmTT の記録を推定できることが示唆された。

これらを参考に，何で（説明変数，独立変数），何を（目的変数，被説明変数，従属変数）予測すると現実世界で価値があるのかなど考えて，分析の目的を意識させて，重回帰分析の演習に誘うとよい。

Step.4　データサイエンス・サイクルの確認

出典　放送大学「データサイエンス基礎から応用」第 1 回スライドより

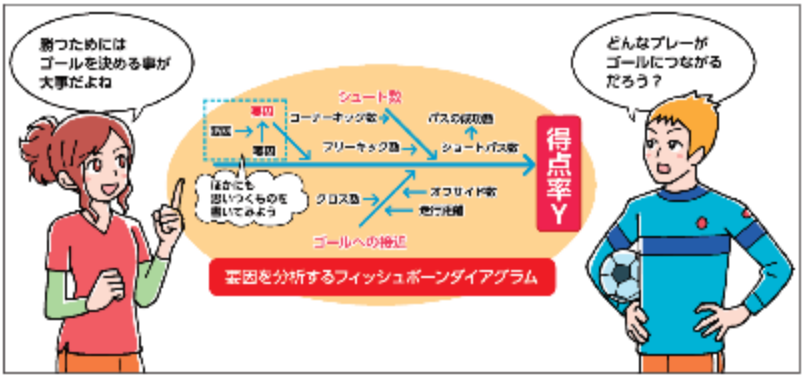
Step.5　データサイエンス・サイクル①　問題の設計

＊分析の目的

高校生男子の 50m 走のタイムを予測するためのその他の体力テスト項目の検討を行うこと。

＊研究の方法

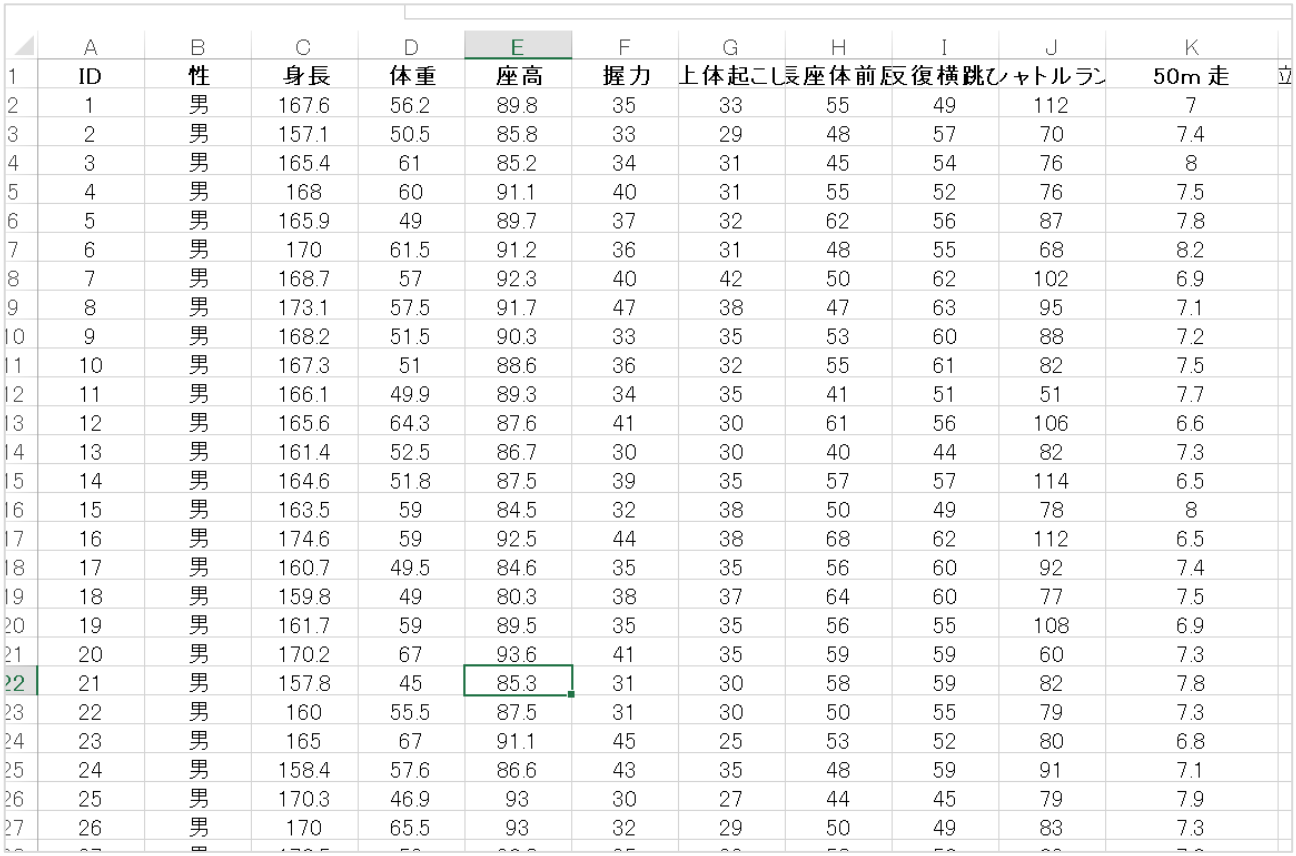
科学の道具箱から取得した男子高校生（学年を 1 年のみにするのか，全学年にするのかなども考えて）？？名を対象に，50m 走の記録と他の体力測定の項目からの間の相関分析，重回帰分析を行う。



特定要因図の作成（「情報Ⅱ」教員研修用教材P126）

Step.6　データサイエンス・サイクル②　データの取得・構造化

例）リスト形式のデータ行列の作成

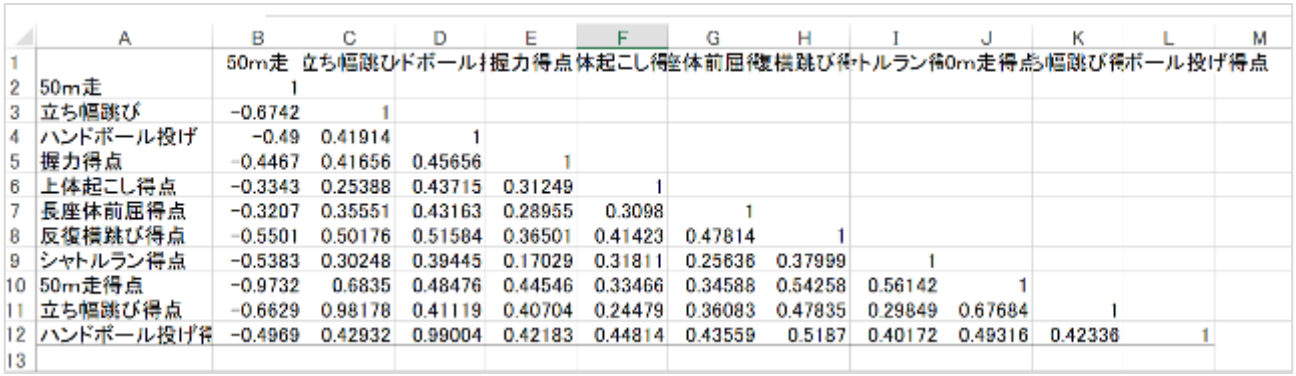


	A	B	C	D	E	F	G	H	I	J	K
1	ID	性	身長	体重	座高	握力	上体起こし	長座体前屈	反復横跳び	シャトルラン	50m 走
2	1	男	167.6	56.2	89.8	35	33	55	49	112	7
3	2	男	157.1	50.5	85.8	33	29	48	57	70	7.4
4	3	男	165.4	61	85.2	34	31	45	54	76	8
5	4	男	168	60	91.1	40	31	55	52	76	7.5
6	5	男	165.9	49	89.7	37	32	62	56	87	7.8
7	6	男	170	61.5	91.2	36	31	48	55	68	8.2
8	7	男	168.7	57	92.3	40	42	50	62	102	6.9
9	8	男	173.1	57.5	91.7	47	38	47	63	95	7.1
10	9	男	168.2	51.5	90.3	33	35	53	60	88	7.2
11	10	男	167.3	51	88.6	36	32	55	61	82	7.5
12	11	男	166.1	49.9	89.3	34	35	41	51	51	7.7
13	12	男	165.6	64.3	87.6	41	30	61	56	106	6.6
14	13	男	161.4	52.5	86.7	30	30	40	44	82	7.3
15	14	男	164.6	51.8	87.5	39	35	57	57	114	6.5
16	15	男	163.5	59	84.5	32	38	50	49	78	8
17	16	男	174.6	59	92.5	44	38	68	62	112	6.5
18	17	男	160.7	49.5	84.6	35	35	56	60	92	7.4
19	18	男	159.8	49	80.3	38	37	64	60	77	7.5
20	19	男	161.7	59	89.5	35	35	56	55	108	6.9
21	20	男	170.2	67	93.6	41	35	59	59	60	7.3
22	21	男	157.8	45	85.3	31	30	58	59	82	7.8
23	22	男	160	55.5	87.5	31	30	50	55	79	7.3
24	23	男	165	67	91.1	45	25	53	52	80	6.8
25	24	男	158.4	57.6	86.6	43	35	48	59	91	7.1
26	25	男	170.3	46.9	93	30	27	44	45	79	7.9
27	26	男	170	65.5	93	32	29	50	49	83	7.3

Step.7　データサイエンス・サイクル③　探索的データ分析：相関行列の作成

重回帰分析の前に，どのような項目と単相関があるのか，相関行列で探索的に確認する。

例）　相関行列は，Excel の「データの分析」メニューの「相関」でも出力できる。「情報Ⅰ」教員研修用教材参照。



	A	B	C	D	E	F	G	H	I	J	K	L	M
1		50m走	立ち幅跳び	ドボール投	握力得点	体起こし得	座体前屈	復横跳び得	トルラン得	0m走得点	幅跳び得	ボール投げ得点	
2	50m走	1											
3	立ち幅跳び	-0.6742	1										
4	ハンドボール投げ	-0.49	0.41914	1									
5	握力得点	-0.4467	0.41656	0.45656	1								
6	上体起こし得点	-0.3343	0.25388	0.43715	0.31249	1							
7	長座体前屈得点	-0.3207	0.35551	0.43163	0.28955	0.3098	1						
8	反復横跳び得点	-0.5501	0.50176	0.51584	0.36501	0.41423	0.47814	1					
9	シャトルラン得点	-0.5383	0.30248	0.39445	0.17029	0.31811	0.25636	0.37999	1				
10	50m走得点	-0.9732	0.6835	0.48476	0.44546	0.33466	0.34588	0.54258	0.56142	1			
11	立ち幅跳び得点	-0.6629	0.98178	0.41119	0.40704	0.24479	0.36083	0.47835	0.29849	0.67684	1		
12	ハンドボール投げ得	-0.4969	0.42932	0.99004	0.42183	0.44814	0.43559	0.5187	0.40172	0.49316	0.42336	1	
13													

どういう体力測定項目と「50m 走」の相関が高いかを相関係数で読み取っておく。ただし，この段階はあくまでも単相関係数なので，これが重回帰モデル式の中でそのまま，有意に効いてくるとは限らない。ここが重回帰分析の難しいところである。インターネット講義　gacco「統計学Ⅲ　多変量解析法」などで発展的に学ぶと理解が深まる。

Step.8　データサイエンス・サイクル④　探索的データ分析：統計モデルによる分析：重回帰モデル

「4　重回帰分析のコンピュータでの実行と出力」（130 ページ）に沿って，実行

Step.9　データサイエンス・サイクル⑤　分析結果の解釈・説明・実装・可視化

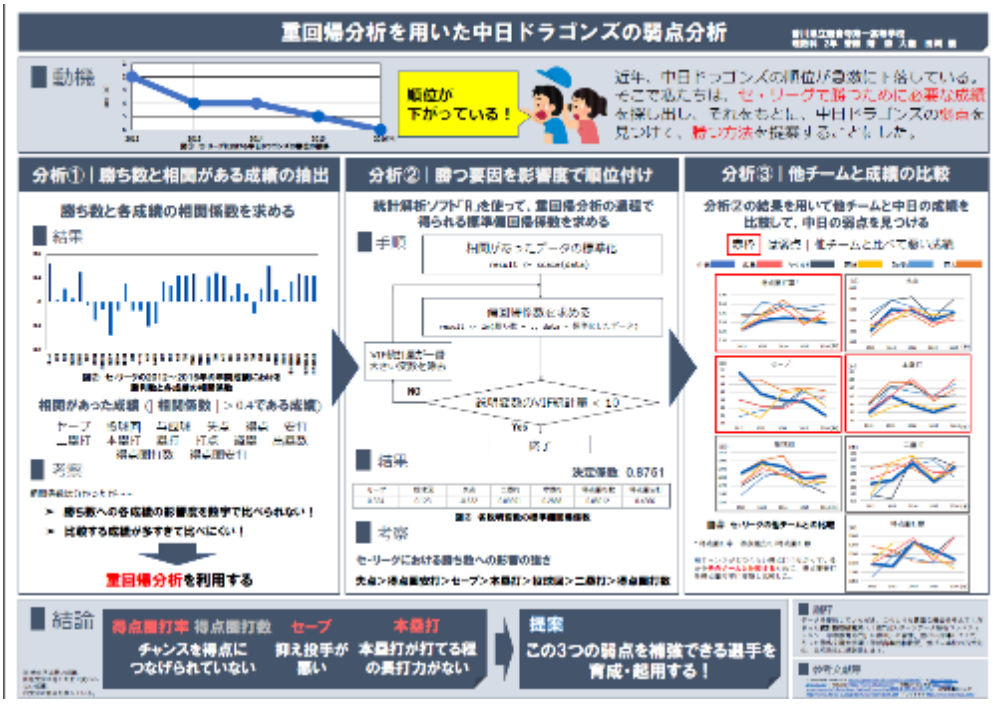
得られたモデル式や寄与率，自由度調整済み寄与率を分かりやすく表示し，説明できるようにする。

例）

50m 走(秒)の予測値＝−0.012（秒／cm）×立ち幅跳び(cm)
−0.014（秒／m）×ハンドボール投げ(m)
−0.040（秒／kg）×握力得点(kg)
−0.025（秒／回）×上体起こし(回)
＋10.819（秒）

また，予測モデル式が確定したら，「実装」として，説明変数の値を具体的に入れると予測値を出力する Excel シートやアプリを作成し，モデルをデータから作成することのサービス価値を生徒に実感させる。

参考：生徒探究事例



出典：日本統計協会（https://estat.sci.kagoshima-u.ac.jp/cse/sports07.htm）

第 3 章【学習 14】演習解答

●演習 1

（解答例）

Step.0　R での主成分分析出力

■出力

Rotation (n x k) = (8 x 8)：主成分負荷量

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
握力	0.3252018	0.2682176	-0.53297421	0.39993408	-0.3653663	-0.31441687	0.34209544	-0.17004275
上体起こし	0.3141190	0.4351668	0.42225757	0.40834395	0.4032249	-0.33321281	-0.29431157	0.08168542
長座体前屈	0.3077864	0.3745785	0.01503113	-0.75987597	-0.2411453	-0.28776668	-0.10238851	0.18941208
反復横跳び	0.3933948	0.1203619	0.05183489	-0.20404673	0.4967487	0.35638527	0.61198108	-0.19529718
シャトルラン	0.3132617	-0.4444223	0.59760197	0.01703693	-0.3900527	-0.21759749	0.17541898	-0.34157859
X50m走	-0.4057185	0.4620511	0.11729178	-0.10636452	-0.0709927	0.04215936	-0.08597965	-0.76329592
立ち幅跳び	0.3681042	-0.3669386	-0.40018514	-0.13933339	0.3055848	-0.10049579	-0.50594605	-0.43684157
ハンドボール投げ	0.3844997	0.1955680	0.06075045	0.15245958	-0.3852838	0.72184877	-0.34234695	0.01636705

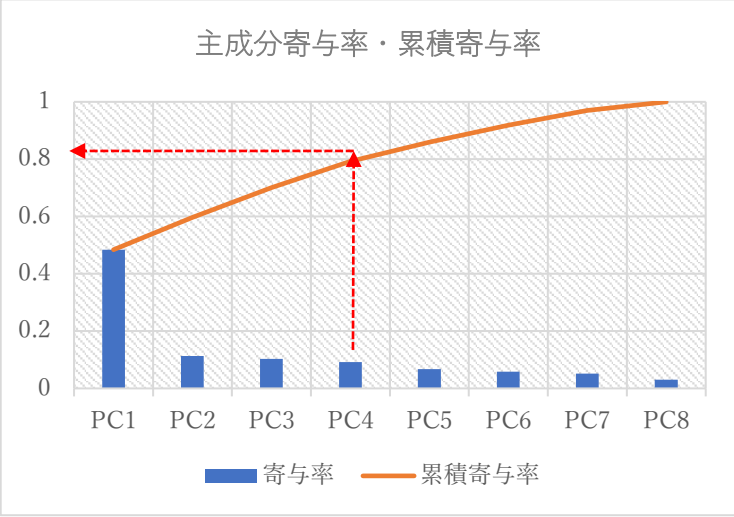
Importance of components:

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	
Standard deviation	1.968	0.9525	0.9096	0.85538	0.73271	0.68576	0.6400	0.49495	主成分の標準偏差
Proportion of Variance	0.484	0.1134	0.1034	0.09146	0.06711	0.05878	0.0512	0.03062	寄与率
Cumulative Proportion	0.484	0.5974	0.7008	0.79229	0.85940	0.91818	0.9694	1.00000	累積寄与率

Step.1　寄与率・累積寄与率

R の出力から，寄与率：Proportion of Variance，累積寄与率：Cumulative Proportion を読み取る(右図)。

PC1(Principle Component)主成分 1 の寄与率は 48.4%，主成分2の寄与率は 11.3%，・・・主成分 4 までの累積寄与率は，79.2%で，もともと 8 変数の体力測定の情報の約 80%が 4 つの主成分に集約されることが分かる。



Step.2　主成分の解釈：主成分負荷量

Rotation 行列の値（主成分負荷量）を下の表のように整理すると解釈が分かりやすくなる。

「50m 走」は値が小さいほど記録がよいと判断される指標なので，－の負荷量を＋と読み替えている。

主成分を解釈する場合は，負荷量の絶対値が相対的に大きなものを拾ってみることになる。

下の表では，マーカーで色を付けた部分になる。

主成分1（総合的な体力得点軸）				主成分2（上半身の筋力（＋）⇔下半身の俊敏性（－））			
＋の負荷量		－の負荷量		＋の負荷量		－の負荷量	
反復横跳び	0.393	X50m走	-0.406	X50m走	0.462	シャトルラン	-0.444
ハンドボール投	0.384			上体起こし	0.435	立ち幅跳び	-0.367
立ち幅跳び	0.368			長座体前屈	0.375		
握力	0.325			握力	0.268		
上体起こし	0.314			ハンドボール投	0.196		
シャトルラン	0.313			反復横跳び	0.120		
長座体前屈	0.308						

主成分３				主成分４			
＋の負荷量		－の負荷量		＋の負荷量		－の負荷量	
シャトルラン	0.598	握力	-0.444	上体起こし	0.408	長座体前屈	-0.760
上体起こし	0.422	立ち幅跳び	-0.367	握力	0.400	反復横跳び	-0.204
X50m走	0.117			ハンドボール投	0.152	X50m走	-0.106
ハンドボール投	0.061			シャトルラン	0.017	立ち幅跳び	-0.139
反復横跳び	0.052						
長座体前屈	0.015						

主成分１は，値が＋で大きいほど総合的に全ての記録がよいことを意味し，逆に，－になるほど全ての記録が悪いことを示す評価軸となっている。主成分２は，値が＋で大きいほど，「上体起こし」と「握力」の記録が「50m 走」，「シャトルラン」，「立ち幅跳び」に比べてよいことを意味し，逆に，－になるほど「50m 走」，「シャトルラン」，「立ち幅跳び」の記録が「上体起こし」と「握力」に比べてよいことを示す評価軸となっている。上半身の筋力や俊敏性などが関連していると考えられる。

負荷量の＋と－の対比が，対応する体力測定指標のバランスのずれに相当する。そのような軸が主成分として構成されるということは，双方向にバランスがくずれた記録を示す生徒が存在するということを意味する。

主成分分析は，主成分（生徒の体力の評価軸）の解釈で終わるのではなく，その評価軸上で生徒のポジショニング（散布図）を行い，よりその特徴を調べていくことが大切である。

「科学の道具箱」からの体力測定のデータをダウンロード方法については当資料の P12を，主成分分析を実行してみましょう。

Step.3　先行研究を調べる

体力測定など，テーマと主成分分析のキーワードをかけて，Google Scholar （論文など学術文献用の検索）等で先行研究を調べてみると，主成分の解釈や主成分得点の 2 次的分析，主成分分析を行う目的などが分かる。

分析に関しては，「情報Ⅱ」教員研修用教材の手順に沿って行ってください。

第 3 章【学習 15】演習解答

●演習 1

演習本文自体が解答となっています。

●演習 2

演習本文自体が解答となっています。

第 3 章【学習 16】演習解答

●演習 1

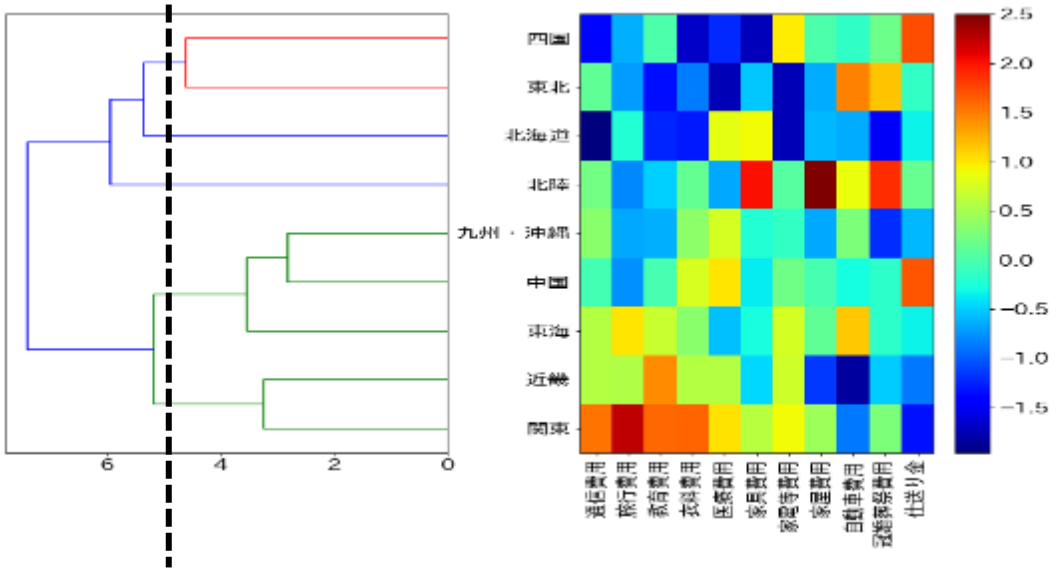
学習 16 の演習 1 に掲載されている，URL からデータをダウンロードして，下記図表 1 を参考に不要な列を削除し、クラスタリングを行う。プログラム及び実行結果は学習 16 の本文中に示したとおりである。

	A	B	C	D	E	F	G	H	I	J	K	L
1	品目区分 (平成 29 年改定)	通信費用	旅行費用	教育費用	衣料費用	医療費用	家具費用	家電等費用	家屋費用	自動車費用	冠婚葬祭費用	仕送り金
2	北海道	13065	5547	6633	1559	1877	1166	3361	7768	19632	4030	2529
3	東北	14672	4985	6414	1726	1367	970	3358	7674	26949	7447	2664
4	関東	15770	8264	12569	2703	1913	1123	4724	9865	18813	6294	1925
5	北陸	14751	4882	8216	2116	1579	1319	4291	13954	24862	8369	2837
6	東海	15026	6903	10658	2199	1602	1005	4642	9197	25831	5730	2533
7	近畿	15019	6394	12205	2288	1825	979	4626	6683	15507	5293	2200
8	中国	14556	4919	9178	2365	1912	990	4363	8916	20875	5735	3781
9	四国	13510	5082	9257	1408	1470	805	4757	8991	21335	6180	3800
10	九州・沖縄	14850	5047	7910	2207	1858	1012	4190	7658	22813	4386	2371

図表 1　演習 1 で使用するデータ

結果として得られたデンドログラムの解釈について一例を挙げる。図表 2 のデンドログラム中に破線で示したところでクラスタリングをした場合には，クラスタ数は 5 になる。四国と東北のクラスタは，旅行費用，衣料費用，医療費用，家具費用が少ない点で類似性があると考えられる。北海道のクラスタは，全体的に費用が少ないが医療費用と家具費用が多めになっている点に特徴がある。北陸のクラスタは，家具費用，家屋費用，冠婚葬祭費用が多い点が特徴的である。九州・沖縄，中国，東海のクラスタは，全体的に平均に近い費用となっている点が共通している。近畿と関東のクラスタでは，自動車費用と仕送り金が少なく，それ以外の費用は多めで，特に教育費用が多い点に特徴があると考えられる。

ここでは解釈の一例を示したが，問題の発見や解決のために何らかの視点に基づいて，クラスタ数を決めたりクラスタの特徴を解釈したりすることが必要である。



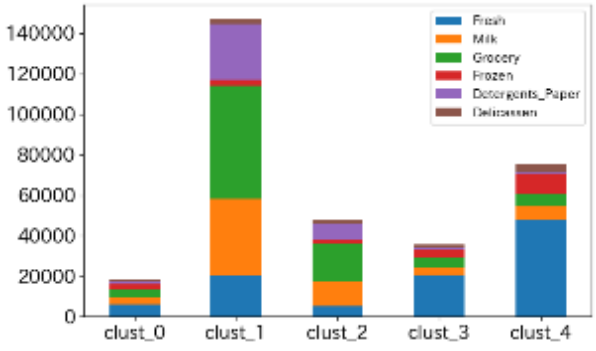
図表 2　演習 1 の結果として得られるデンドログラム

●演習 2

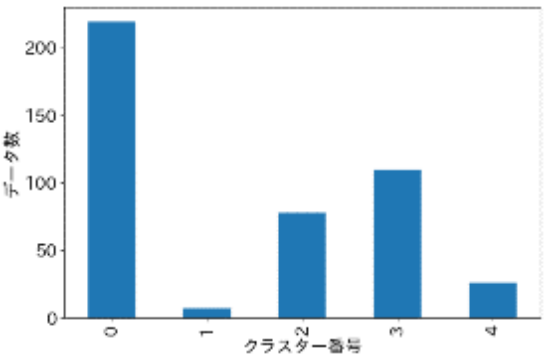
演習本文が回答となっています。

●演習 3

クラスタごとの注文額の平均値とデータ数を再掲し，クラスタごとに考えられる特徴の一例を挙げる。



図表 3　クラスタごとの注文額の平均値



図表 4　クラスタごとのデータ数

クラスタ番号 0（clust_0）の顧客は全体的に注文額が少ないが，200 件を超える多くの顧客がこのクラスタに該当する。クラスタ番号 1（clust_1）の顧客は Fresh，Milk，Grocery，Detergents_Paper の注文額が他のクラスタに比べて圧倒的に高額となっている。しかし，このクラスタの件数は少ない。クラスタ番号 2（clust_2）の顧客は，Grocery，Milk，Detergents_Paper の注文額が多めになっている。クラスタ番号 3（clust_3）の顧客は，クラスタ番号 0 の顧客に比べて Fresh の注文額が多く，他のものはクラスタ番号 0 とほぼ同じ注文額となっている。クラスタ番号 4（clust_4）の顧客は，Fresh の注文額が他のクラスタに比べてかなり多い特徴がある。ここで，新規の顧客が獲得可能な場合は，クラスタ番号 1 の顧客の特徴を分析し，それに合ったマーケティングを行う。新規の顧客が望めない場合は，クラスタ番号 0 の顧客の特徴を分析し，注文量の増加を図ることなどが考えられる。

ここまでについては，クラスタ数を 5 にしてクラスタリングを行った。クラスタ数 5 は，プログラム中の次の行の n_clusters=5 で決定している。この値を変えることで，クラスタ数を変えて分析することができる。

```
KMEANS = KMEANS( INIT='RANDOM', N_CLUSTERS=5, RANDOM_STATE=0 )
```

●演習 4

パンと一緒に買われているものには，コーヒー，弁当，お茶，紅茶がある。これらのものについて，支持度，確信度，リフト値を求めると次の表になる。

X	supp(パンと X)	conf(パン→X)	supp(X)	lift(パン→X)
コーヒー	3/8＝0.375	3/5＝0.6	4/8＝0.5	6/5＝1.2
弁当	3/8＝0.375	4/5＝0.8	7/8＝0.875	32/35≒0.91

お茶	1/8＝0.125	1/5＝0.2	2/8＝0.25	4/5＝0.8
紅茶	1/8＝0.125	1/5＝0.2	2/8＝0.25	4/5＝0.8

図表 5　パンを買った人に勧めるための評価指標

支持度（supp）によりパンと一緒に買われているものを見つけることができる。コーヒーと弁当の値が高く，これらがパンと一緒に買われていることが分かる。

次に，確信度（conf）により，パンを買った人の中で買われているものとして多いものが分かる。これにより，弁当が 5 件中 4 件，コーヒーが 5 件中 3 件の順になっていることが分かる。

リフト値を求めるために，それぞれの商品についての支持度を調べる。これにより，パンの購入の有無に関わらず弁当が売れていることが分かる。

この値を用いてリフト値を求めると，コーヒーでは 1 を超えており最も高い値となっている。これにより，パンを買う人は，お客さん全体と比較してコーヒーを買う傾向が高いことが分かる。

これらの値を用いると，パンを買った人には，リフト値が高いコーヒーを提案するとともに，支持度が高い弁当についても買い忘れることがないよう提案したい。

第 3 章【学習 17】演習解答

●演習 1

(解答例)

選んだ AI 技術:

Human or AI (同じ曲の演奏を聞いて，人間による演奏か AI による演奏かを当てる Web サイト)

更なる活用方法:

・特殊な空間(水中や人が入れないような狭い場所)やパフォーマンス(ダンスと生演奏のコラボレーション)に AI の演奏を利用する。

・演奏ではなくこの技術を更に進化させて人間のスピーチの代替とする。例えば命を狙われている有名人の演説で本人が立たず，画像と AI スピーチを組み合わせる。

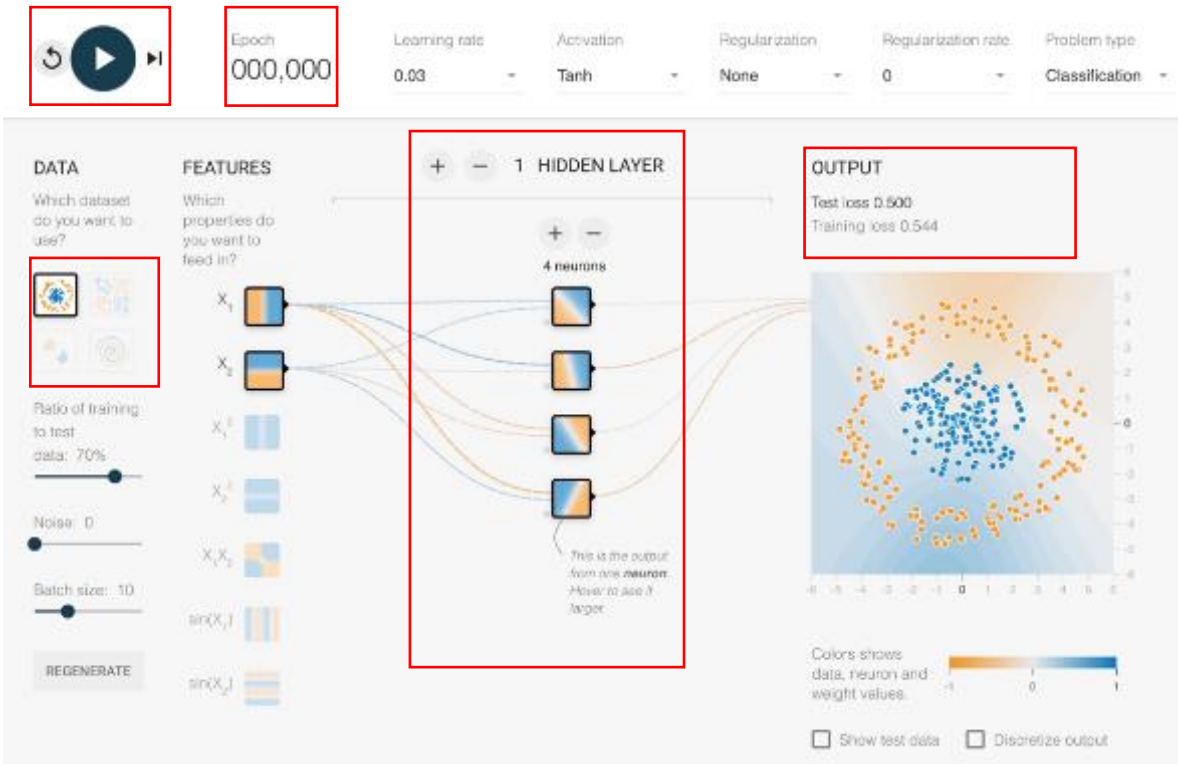
注意すべき点:

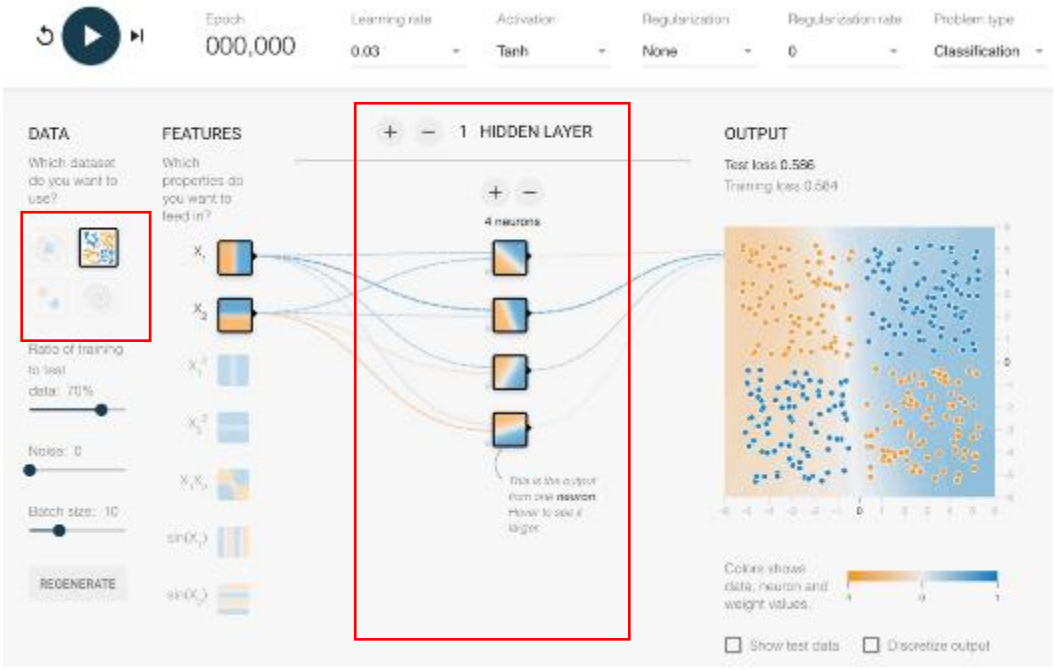
・人間のパフォーマンスか否かをはっきりさせる方がいいのかどうか。権利の問題(誰の作品とするか，オリジナリティをどこまで認めるか)。

・本人よりいいスピーチをする可能性。影武者としての AI を許容するかどうか。

●演習 2

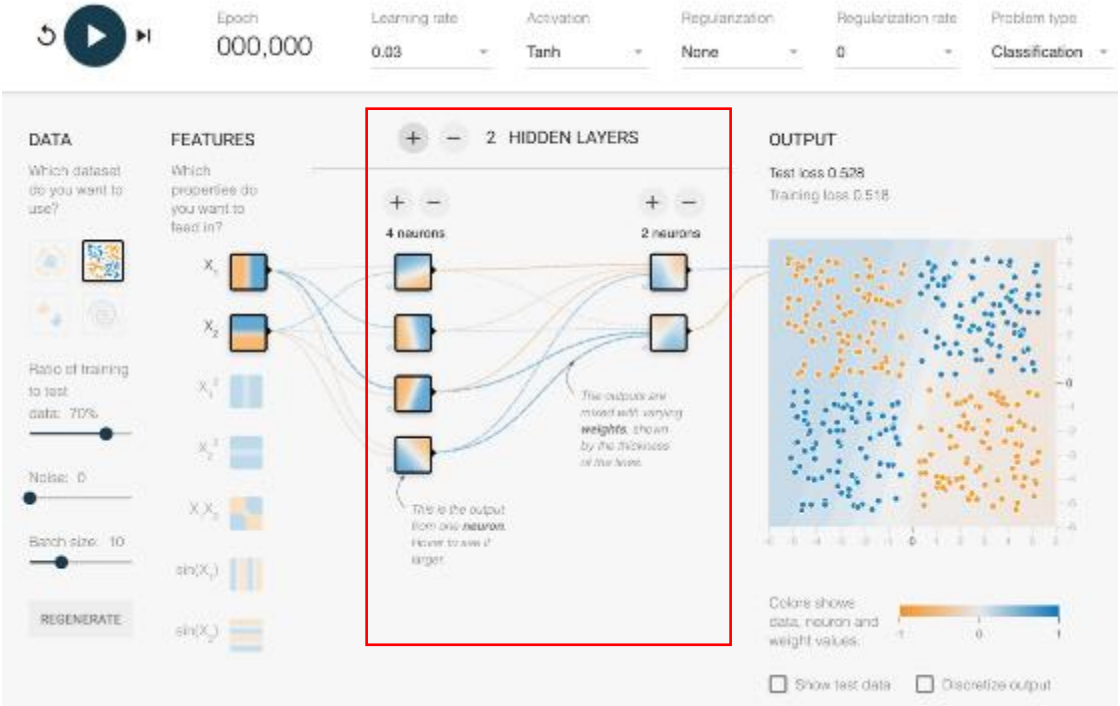
TensorFlow.playground を使用する。(http://playground.tensorflow.org/)



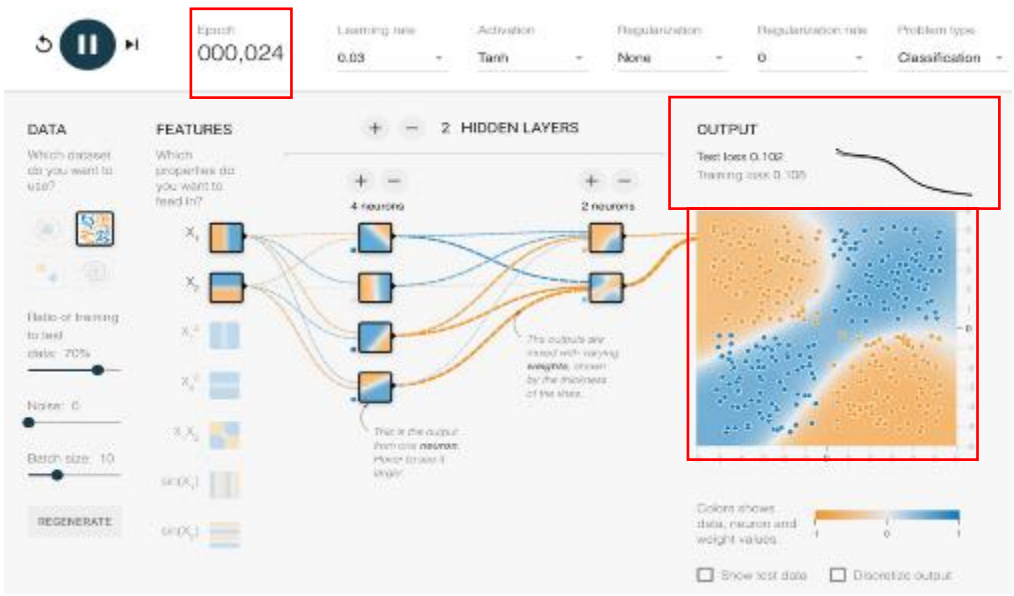


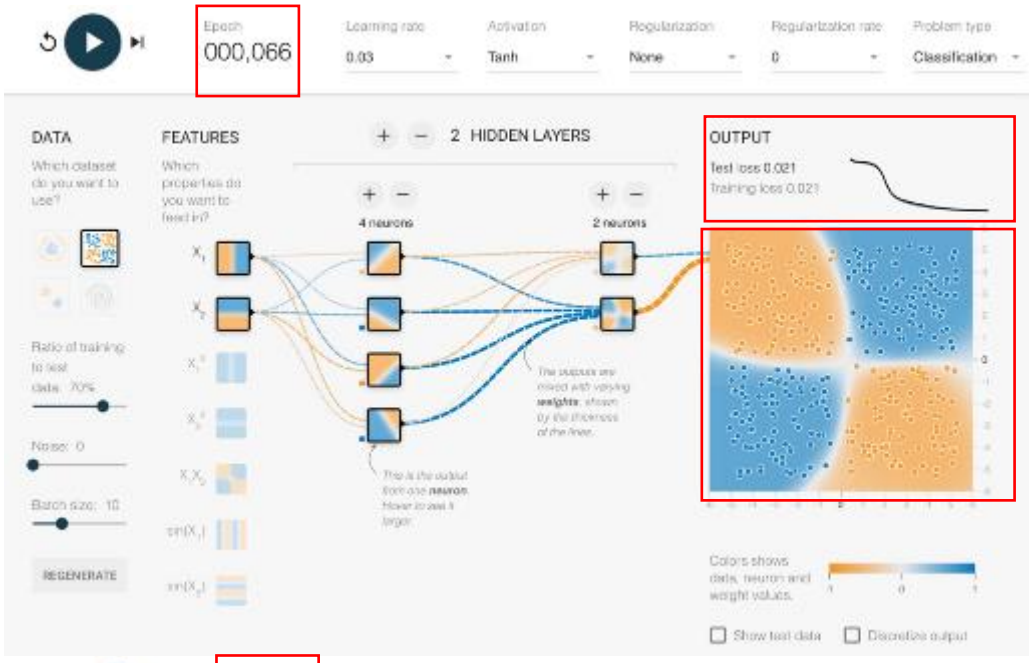
（考察例）

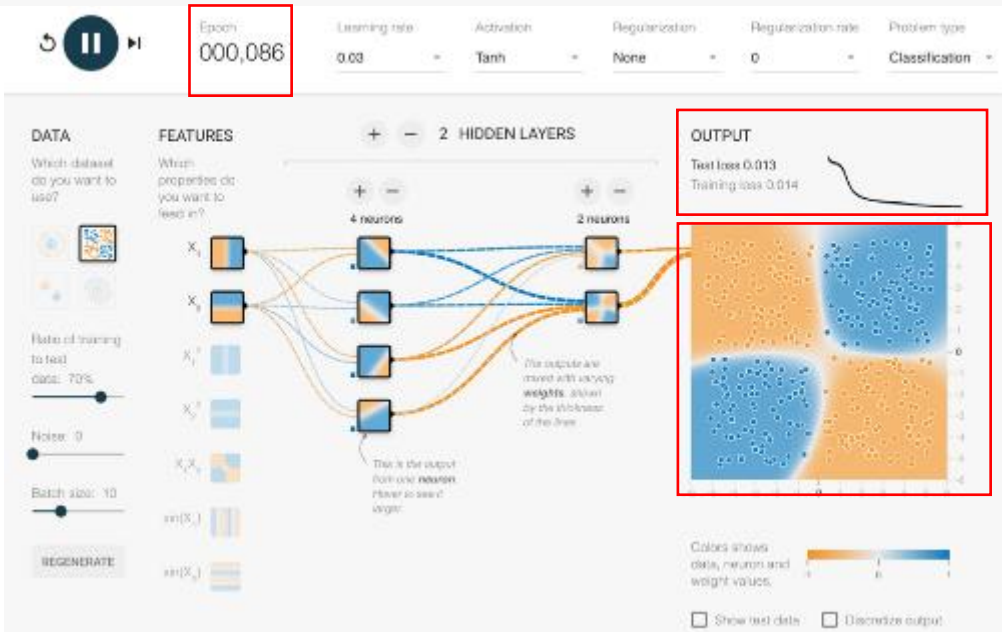
Hidden Layer(隠れ層)を 2 層とすると下図のようになる。

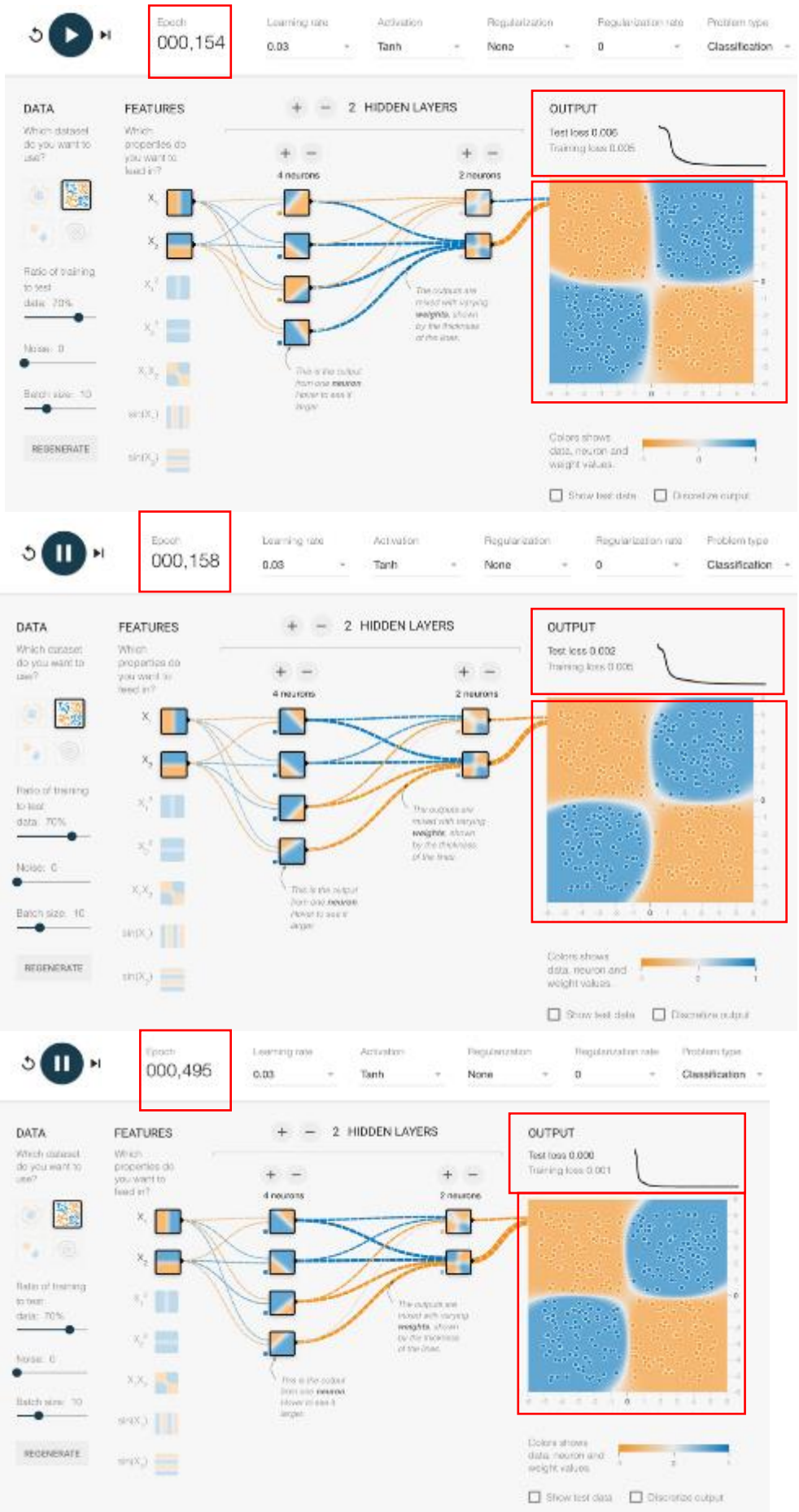


ここから，学習を開始すると，だんだんオレンジの点と青い点を分けている背景色が濃くなり境界も移動し，はっきりとしてくる。このとき，画面上部左上の Epoch の数が増えていくにつれて分類の精度が上がっていくこと，右上の OUTPUT のグラフが滑らかな形に変化していくこと，Training loss の値が限りなく 0 に近づいていくことが確認できる。


Epoch
000,024
Learning rate
0.03
Activation
Tanh
Regularization
None
Regularization rate
0
Problem type
Classification
DATA
FEATURES
2 HIDDEN LAYERS
4 neurons
2 neurons
OUTPUT
Test loss 0.102
Training loss 0.105
Ratio of training to test data: 70%
Noise: 0
Batch size: 10
REGENERATE
Colors shows data, neuron and weight values.
Show test data
Discretize output


Epoch
000,066
Learning rate
0.03
Activation
Tanh
Regularization
None
Regularization rate
0
Problem type
Classification
DATA
FEATURES
2 HIDDEN LAYERS
4 neurons
2 neurons
OUTPUT
Test loss 0.021
Training loss 0.021
Ratio of training to test data: 70%
Noise: 0
Batch size: 10
REGENERATE
Colors shows data, neuron and weight values.
Show test data
Discretize output


Epoch
000,086
Learning rate
0.03
Activation
Tanh
Regularization
None
Regularization rate
0
Problem type
Classification
DATA
FEATURES
2 HIDDEN LAYERS
4 neurons
2 neurons
OUTPUT
Test loss 0.013
Training loss 0.014
Ratio of training to test data: 70%
Noise: 0
Batch size: 10
REGENERATE
Colors shows data, neuron and weight values.
Show test data
Discretize output



学習回数（Epoch）がある程度の回数を超えると変化が見られなくなる。（飽和）

●演習 3

（解答）

計算式： 0.3×1＋0.8×2＋1.2×1＋0.7×1＋0.2

結果：4

●演習 4

（解答）

演習本文自体が解答となっています。

●演習 5

（解答）

演習本文自体が解答となっています。

＊参考＊

実際の授業では，AI の導入に下記のようなブロックプログラミングで AI を利用する体験を加えてもよい。

「Scratch で使える 拡張 AI ブロック- Scratch で AI を使ってみよう -」（https://www.techpark.jp/aiblock）

ただし，上記はあくまで“AI のトレーニングとその利用の体験”である。本文で扱ってきたような仕組みの理解のための教材ではない点に注意し，「情報 II」ではぜひ仕組みの理解とニューラルネットワークの構築まで演習していただきたい。

第 3 章【学習 18】演習解答

●演習 1

意識的にコンピュータのためのデータを人間が作っている例としては，Wikipedia の記事の作成や映画や本のレビュー書き込みなど様々なものが考えられる。

無意識に行っている例としては，IoT のデータのクラウドアップロードやスマートスピーカー等の音声アシスタントの学習，Google 日本語入力の学習などもあげられるだろう。

これらの活動によって作られているものの例として，以下の例を挙げることも面白い。

・シープマーケット http://www.thesheepmarket.com

・絵文字ディック http://www.emojidick.com

●演習 2

これに関しては，演習の解説全体が解答となっています。

●演習 3

これに関しては，演習の解説全体が解答となっています。

●演習4

これに関しては，演習の解説全体が解答となっています。