課題

ウィスコンシン乳がんデータセットを用いて，細胞診データから乳がん診断に有用な細胞の特徴を探索します．Python または R を用いて以下の課題を行なって下さい．

1. データセットに含まれる 30 の特徴全てについて，良性（Benign）と悪性（Malignant）の場合での密度分布を作成して下さい．
2. Wilcoxson 順位和検定（Mann-Whitney U 検定）を用いて良性と悪性の細胞で有意に分布が異なる特徴を検定して下さい．

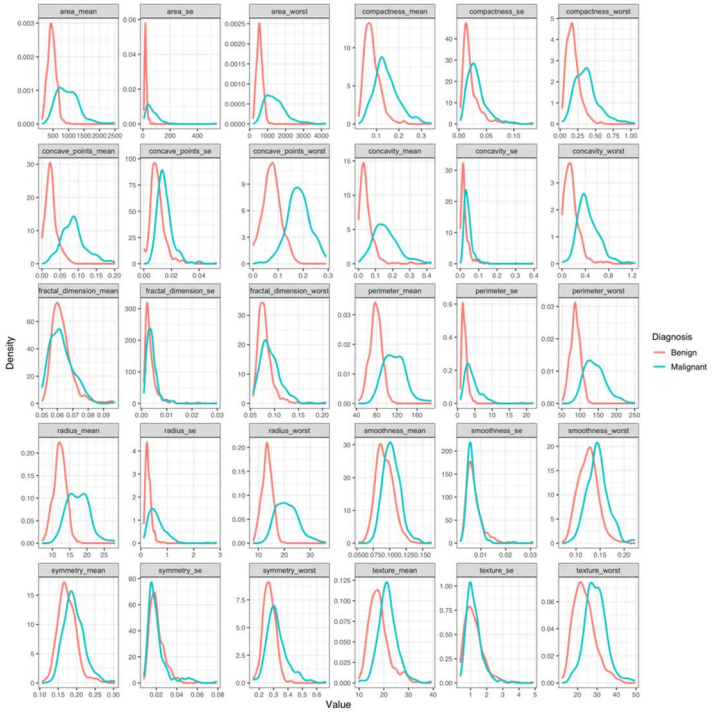
本課題の狙い

ノンパラメトリック検定でよく利用される Wilcoxson 順位和検定を理解する．また，この手法による解析が，さらに高度な機械学習アルゴリズムを用いる解析の前処理として重要なことを理解する．

解答例

R による解析例：

密度分布



Wilcoxson 順位和検定

Feature	Malignant	Benign	Differemce	P value
perimeter_worst	138	86.92	51.08	2.57E-80 ***
radius_worst	20.59	13.35	7.24	1.13E-78 ***
area_worst	1303	547.4	755.6	1.79E-78 ***
concave_points_worst	0.182	0.07431	0.10769	1.85E-77 ***
concave_points_mean	0.08628	0.02344	0.06284	1.00E-76 ***
perimeter_mean	114.2	78.18	36.02	3.54E-71 ***
area_mean	932	458.4	473.6	1.53E-68 ***
concavity_mean	0.15135	0.03709	0.11426	2.15E-68 ***
radius_mean	17.325	12.2	5.125	2.68E-68 ***
area_se	58.455	19.63	38.825	5.74E-65 ***
concavity_worst	0.4049	0.1412	0.2637	1.75E-63 ***
perimeter_se	3.6795	1.851	1.8285	5.08E-51 ***
radius_se	0.5472	0.2575	0.2897	6.19E-49 ***
compactness_mean	0.13235	0.07529	0.05706	8.92E-48 ***
compactness_worst	0.35635	0.1698	0.18655	2.11E-47 ***
concave_points_se	0.014205	0.009061	0.005144	2.36E-31 ***
texture_worst	28.945	22.82	6.125	6.50E-30 ***
concavity_se	0.037125	0.0184	0.018725	3.66E-29 ***
texture_mean	21.46	17.39	4.07	3.42E-28 ***
smoothness_worst	0.14345	0.1254	0.01805	3.63E-24 ***
symmetry_worst	0.3103	0.2687	0.0416	3.14E-21 ***
compactness_se	0.02859	0.01631	0.01228	1.17E-19 ***
smoothness_mean	0.1022	0.09076	0.01144	7.77E-19 ***
symmetry_mean	0.1899	0.1714	0.0185	2.26E-15 ***
fractal_dimension_worst	0.0876	0.07712	0.01048	1.14E-13 ***
fractal_dimension_se	0.0037395	0.002808	0.0009315	1.57E-06 ***
symmetry_se	0.0177	0.01909	-0.00139	0.02781792 *
smoothness_se	0.0062095	0.00653	-0.0003205	0.21353455
fractal_dimension_mean	0.061575	0.06154	3.50E-05	0.53701169
texture_se	1.1025	1.108	-0.0055	0.64350365

Malignant と Benign の値は各特徴の中央値．Difference はその差

留意点

分布の正規性を仮定できないデータが決して珍しくないこと，そしてそのようなデータに対しては分布の形状に仮定を置かないノンパラメトリック検定が有効であることを説明する．

課題

独立行政法人統計センターから総務省・家計調査データが教育用に使いやすい形に編集したデータセットが公開されている（https://www.nstac.go.jp/SSDSE/）。このデータセットを用いて、以下の課題に取り組んでください。.

1. 相関が高いと思われる変量を２つ選んで、その単回帰分析を実施せよ。
   （Python などの開発言語に慣れていない場合は、エクセルでも取り組めます）

2. 単回帰分析によって得られた結果に対して、考察を加えよ。

本課題の狙い

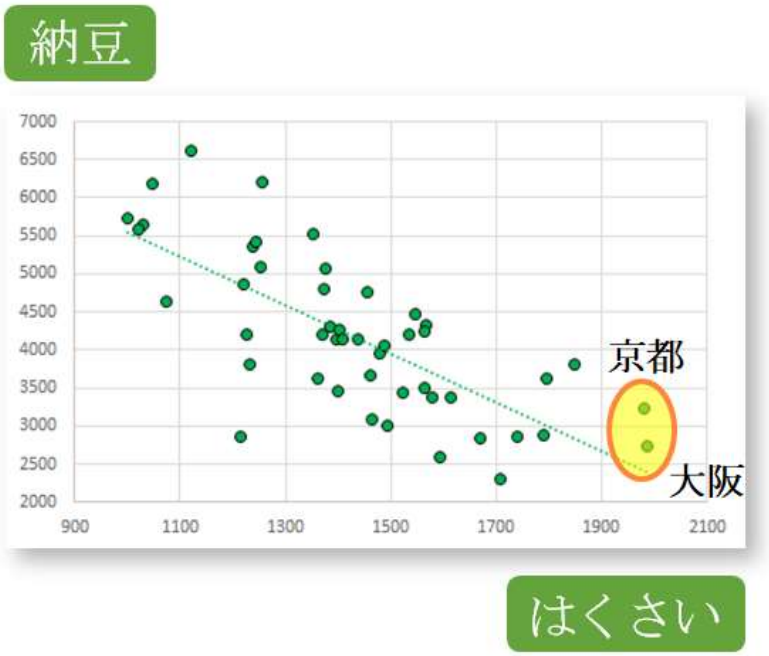
実際のデータに対して、単回帰分析を実施することによって、理論的な理解を深めるとともに、単回帰分析の有効性を経験する。さらには、分析スキルが高められることを期待する。

解答例

１．各都道府県にておける、「はくさい」および「納豆」の１世帯当たりの年間支出金額の平均値をデータセットから取得する。横軸に「はくさい」、縦軸に「納豆」の支出額をとり、各都道府県を２次元空間に布置する。その散布図は以下となる。

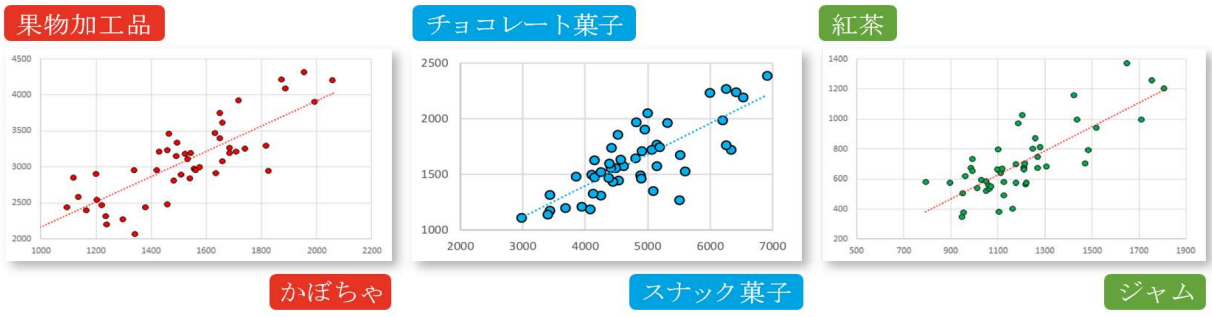
さらに、単回帰分析を行うと「はくさい」と「納豆」の支出額の関係は、散布図上の点線のようになる。つまり、「はくさい」への支出額が大きい都道府県では、「納豆」への支出額が小さい傾向がある。なお、相関係数を計算すると-0.732 であり、負の相関があることを定量的にも確認できる。

２．京都や大阪では納豆への支出額が低い。関西では納豆を食べる習慣があまりないということがわかります。一方、はくさいの支出額が高い。その理由の１説として、はくさいを購入して漬物やキムチを作る人・業者が多いと言われている。漬物やキムチなどがご飯のお供として好まれるなら、納豆が食べられる機会が減るのかもしれない。統計データおよび回帰分析がこの１説を支持する。



留意点

できる限り多くの品目間の回帰分析を行って、分析手順の手続きを習得し、分析結果の見方を学習するように努めてください。参考までに以下の分析結果も掲示します。



課題

ピマ族の糖尿病に関するデータベースを用いて，糖尿病の発症要因をロジスティック回帰で探索します． Python または R を用いて以下の課題を行なって下さい．

1. データセットの数値を標準得点に変換した上で糖尿病発症に対するロジスティック回帰モデルを構築し，当てはめた結果の標準偏回帰係数を求めて下さい．
2. Pregnant（妊娠回数）と Age（年齢）は相関する場合が多いことを考慮して，この 2 変数の交互作用を入れたロジスティック回帰モデルでデータを再解析し，その結果の違いについて考察して下さい．

本課題の狙い

機械学習における分類器としてではなく，探索的データ解析の手法としての一般化線形モデル（ロジスティック回帰）について理解する．

解答例

R による解析例：

```
library(tidyverse)

data("PimaIndiansDiabetes2", package = "mlbench")
PimaIndiansDiabetes2 <- na.omit(PimaIndiansDiabetes2)
df <- scale(PimaIndiansDiabetes2[,-9]) %>% as.data.frame()
df$diabetes <- PimaIndiansDiabetes2$diabetes

fit_lr <-glm(formula = diabetes~., family    = binomial(),df)
summary(fit_lr)$coefficients
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.0002571	0.14327356	-6.9814492	2.92E-12***
pregnant	0.26384878	0.17799495	1.48233852	0.13825024
glucose	1.18102733	0.17799599	6.63513444	3.24E-11**

pressure	-0.0177481	0.14787119	-0.1200239	0.90446422
triceps	0.11800889	0.17965953	0.6568474	0.51127904
insulin	-0.0980816	0.15525941	-0.631727	0.52756526
mass	0.49571409	0.19215123	2.57981222	0.00988541**
pedigree	0.39417029	0.14767324	2.66920592	0.00760308**
age	0.34633292	0.18750783	1.84703177	0.06474254

```
fit_lr <-glm(formula = diabetes~.+age:pregnant, family    = binomial(),df)
summary(fit_lr)$coefficients
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.8963196	0.16493141	-5.4344992	5.50E-08***
pregnant	0.37925477	0.20024338	1.89396902	0.05822912
glucose	1.16786973	0.17862509	6.53810567	6.23E-11***
pressure	-0.020838	0.14792711	-0.1408666	0.88797536
triceps	0.11283572	0.18024726	0.62600515	0.53131156
insulin	-0.0968432	0.1570648	-0.616581	0.53751115
mass	0.50473161	0.19283291	2.61745578	0.0088588**
pedigree	0.39026975	0.14799775	2.63699793	0.00836433**
age	0.41995805	0.1977015	2.12420267	0.0336532*
pregnant:age	-0.1762055	0.14010668	-1.2576522	0.20851756

留意点

回帰分析における適切な変数選択には，問題に対する事前知識が重要となる．glucose と insulin といった明らかに関係しそうな変数についても交互作用を考えるよう教示する．

課題

ピマ族の糖尿病の発症要因に関するロジスティック回帰モデルで予測に重要な変数を探索するため，AIC を用いたステップワイズ法でモデル選択を行います． Python または R を用いて以下の課題を行なって下さい．

1. データセットの数値を標準得点に変換した上で 7 個の変数全てを用いてロジスティック回帰モデルを構築し，これに対して AIC によるステップワイズ法で変数選択を行って下さい．
2. さらに，7 個の変数全ての組み合わせの 2 次の交互作用を入れたロジスティック回帰モデルを構築し，これに対して AIC によるステップワイズ法で変数選択を行って下さい．

本課題の狙い

回帰モデルでの変数選択という代表的な事例を通して，情報量基準を用いたモデル選択について理解する．

解答例

R による解析例：

1． 7 変数のロジスティック回帰での変数選択

```
fit_lr <-glm(formula = diabetes~., family   = binomial(),df)
fit_lr_step <- step(fit_lr)
summary(fit_lr_step)$coefficients
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.9964042	0.14261227	-6.9868055	2.81E-12***
pregnant	0.26960875	0.17672673	1.5255686	0.12711734
glucose	1.12511498	0.15362084	7.32397357	2.41E-13***
mass	0.54913186	0.14480726	3.79215698	0.00014934***
pedigree	0.39762663	0.14657068	2.71286609	0.00667041**
age	0.35050234	0.18167195	1.9293146	0.05369182

Insulin と triceps が回帰の変数から外れる

2．7 変数とその全ての 2 次交互作用を考慮したロジスティック回帰での変数選択

```
fit_lr <-glm(formula = diabetes~.^2, family  = binomial(),df)
fit_lr_step <- step(fit_lr)
summary(fit_lr_step)$coefficients
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.9071099	0.17149577	-5.289401	1.23E-07***
glucose	1.53619107	0.22262286	6.90041917	5.18E-12***
pressure	0.01604384	0.16001738	0.10026309	0.92013546
triceps	0.19197612	0.19962514	0.96168305	0.33620884
insulin	-0.1305771	0.19847752	-0.6578938	0.51060636
mass	0.68226876	0.21807103	3.12865385	0.00175609**
pedigree	0.60778746	0.16337386	3.72022476	0.00019905***
age	0.53815448	0.16887375	3.18672669	0.00143893**
glucose:pressure	0.41444527	0.22151876	1.8709263	0.0613553
glucose:triceps	0.50111859	0.23175569	2.16227095	0.0305973*
glucose:mass	-0.6385115	0.25270596	-2.5266974	0.01151407*
glucose:age	-0.4772772	0.19915628	-2.396496	0.01655267*
pressure:insulin	-0.4302531	0.20682261	-2.0803001	0.03749802*
triceps:mass	-0.2991351	0.16659664	-1.795565	0.07256373
triceps:age	-0.3186151	0.16398796	-1.9429175	0.05202613
insulin:pedigree	-0.3184398	0.10230312	-3.1127083	0.00185379**
insulin:age	0.58358714	0.19811227	2.94573957	0.0032218**4

主効果として age が有意になるとともに pregnant は変数から外れる．正の交互作用（相乗効果）があるものとしては glucose・triceps と insulin・age，負の交互作用があるものとしては glucose・mass，glucose・age，　pressure・insulin，insulin・pedigree が見出された．

留意点

変数の数が多い場合の回帰分析において，ステップワイズ法は適切な変数選択に有効な方法であるが，あくまでも AIC という情報量基準に基づいて見出されたものであることは説明する（他の基準では別の変数が選ばれることもある）．

課題

ウィスコンシン乳がんデータセットを用いて，細胞診データの特徴量の関係を探索します．Python または R を用いて以下の課題を行なって下さい．

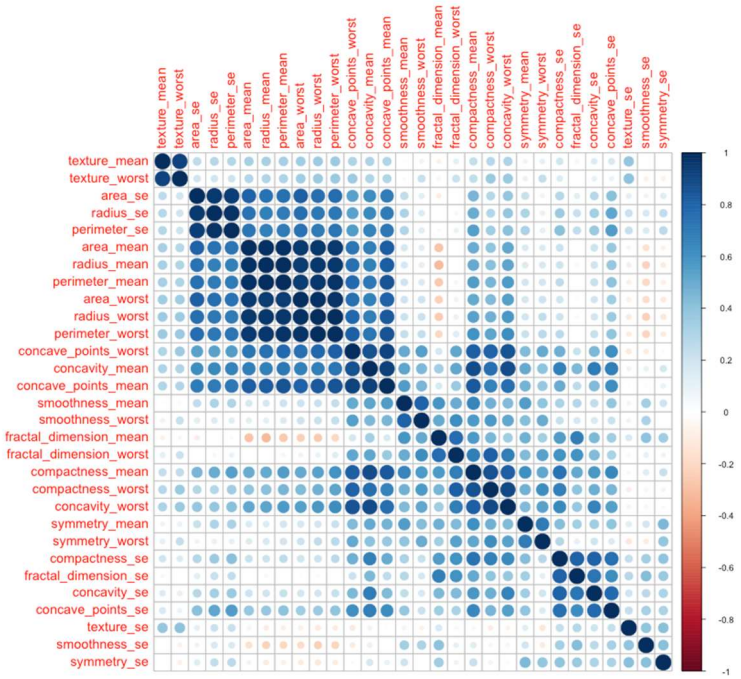
1. データセットに含まれる 30 の特徴に関する相関行列を求め，相関のない特徴の数について考察して下さい．
2. 全ての特徴について主成分分析を行い，主成分得点の分布を確認するとともに重要な成分の数について考察して下さい．

本課題の狙い

主成分分析における次元削減の原理が，相関のない合成変数を作成することであることを理解する．また，主成分分析による有効な次元数の見積りについて理解する．

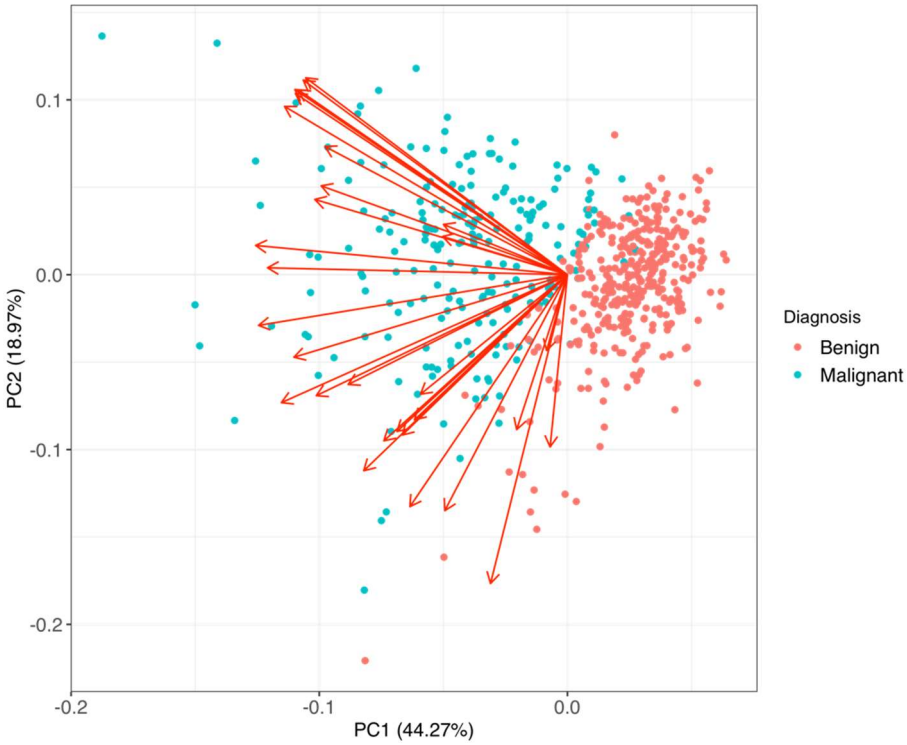
解答例

R による解析例：
相関行列



ブロック対角行列となっている各ブロックで 1 つの変数のみを選択すると，およそ 11 から 12 個程度の変数の相関がないと考えられる

主成分得点とバイプロット



主成分の累積寄与率を計算すると 10 個の成分で 95%を超える

留意点

主成分分析の利用の仕方として，可視化だけでなく，特徴の次元を削減することで予測性能がより良い回帰モデルの構築にもつながることを説明する．

■課題

1. 1973 年のアメリカの 50 州における犯罪検挙率（10 万人あたりの検挙数）のデータセット（USArrests）について主成分分析を行って下さい．その際に，各変数（Murder：殺人検挙率，Assault：暴行検挙率，UrbanPop：都市人口，Rape：レイプ検挙率）を正規化しない場合とする場合を比較して結果の違いについて考察して下さい．

2. 手書き数字のデータセット（MNIST）の中の「8」の数字のデータについて主成分分析を行い，主成分得点の分布を描画するとともに，代表的な固有ベクトルについて考察して下さい．

■狙い

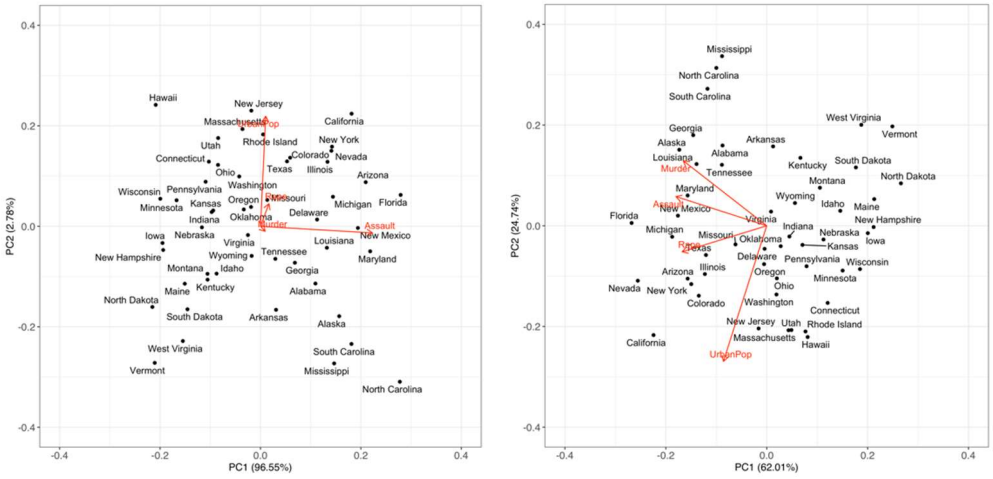
主成分分析における変数の正規化の効果を理解する．また，主成分得点だけではなく固有ベクトルに表現される情報の有用性を理解する．

■解答例

1

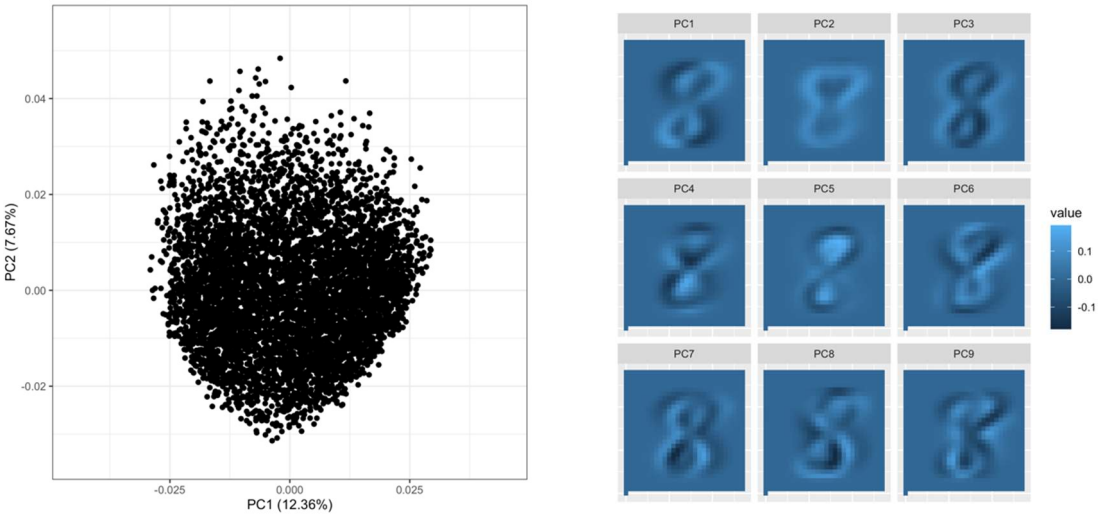
```
data("USArrests")
pca_data <- USArrests
pca.out. unscaled <- prcomp(pca_data, scale=F) #正規化なし
pca.out.scaled <- prcomp(pca_data, scale=T) #正規化あり
#  各変数の分散
apply(USArrests,2,var)
```

Murder	Assault	UrbanPop	Rape
18.97047	6945.16571	209.51878	87.72916



左の正規化してないものでは，Murder と Rape の寄与がほとんど反映されていない．主成分分析では分散の大きい変数の影響が大きくなるので，各変数の寄与を同等に評価するためには正規化を行っておく必要がある，

2.



右の固有ベクトルの中で第 1 固有ベクトル(PC1)は文字の傾き, 第 2 固有ベクトル(PC2)は中心部の交差の有無，第 3 固有ベクトル（PC3 ）は文字の大きさと特徴を表現していると考えられる．

■留意点

主成分分析の利用の仕方として，次元削減だけでなく，固有ベクトルによる特徴抽出の重要性を説明する．

課題

1. ピマ族の糖尿病に関するデータベースから全ての被験者のデータを標準得点に変換してヒートマップで表示し，糖尿病の発症要因と被験者の両方で階層的クラスタリングを行って下さい．

2. Ekman (1954) による 14 種類の波長光の類似性評定のデータから，多次元尺度法(MDS)を用いてそれぞれの波長光を互いの類似性を反映するように 2 次元平面に配置して下さい．

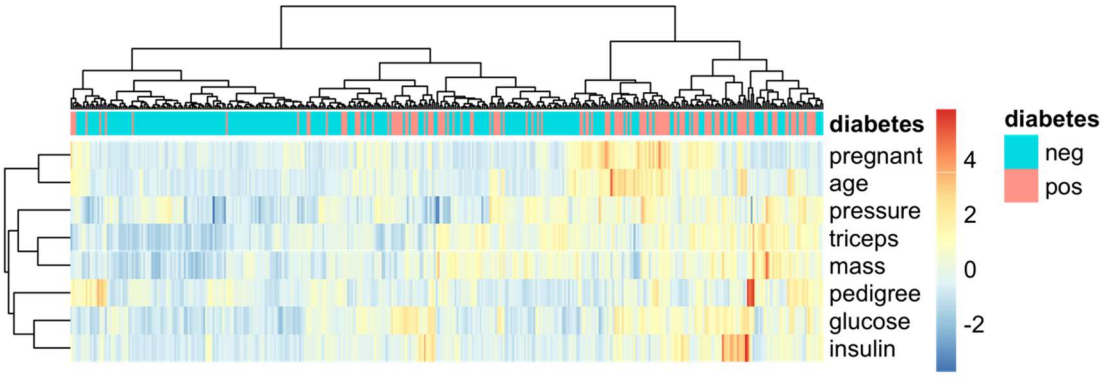
本課題の狙い

データ間の距離に基づいたクラスタリングの代表的手法である階層的クラスタリング，およびデータの類似性・親近性に基づいて低次元空間への射影を行う多次元尺度法について理解する．

解答例

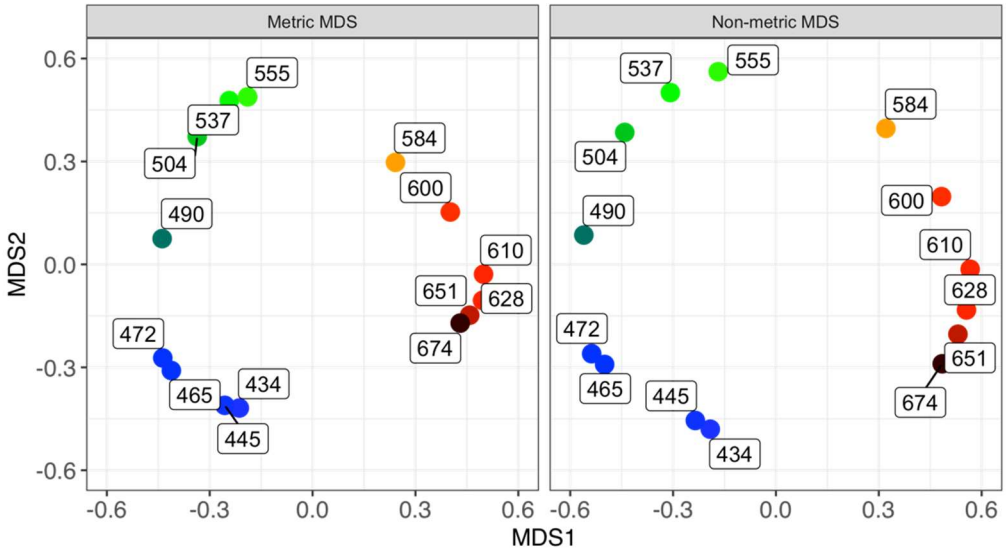
R による解析例：

1. ヒートマップとデンドログラム（樹形図）



被験者間では左側の糖尿病未発症者がほとんどのクラスターと右側の糖尿病発症者が多いクラスターに分かれる．前者では発症要因の検査値が低い，後者では高い傾向がある．発症要因に関しては，pregnant と age の組み，glucose と insulin の組みが同じクラスターになっていることがわかる．

2. 多次元尺度法による色空間



計量 MDS でも非計量 MDS でも色相環に対応した円環状の配置が得られるが，後者の方がより円に近い配置になっている．

留意点

発展事項として階層的クラスタリングからのクラスター数推定の方法について説明する．また，円周上の弧長で計量が定義されるような構造を持ったデータに関しては，計量 MDS や主成分分析などのユークリッド距離に基づいた計量を用いた分析では元の構造が復元できない場合があること，その場合には非計量 MDS が有効な手法になることを説明する．

課題

1. アヤメデータの Sepal Length の中央値（5.8）の経験分布をブートストラップ法で求めてそのヒストグラムを描画して下さい．さらに，その平均値の 95%信頼区間を正規分布近似（Normal），basic 法，パーセンタイル法および BCa 法で求めて下さい．

2. 31 本の桜の倒木の外周長（Girth），樹高（Height）および容積のデータベース（trees）を用いて，外周長から樹高を予測する単回帰モデルを作成し，その予測の信頼区間を誤差のリサンプリングによるブートストラップで求めて下さい．

本課題の狙い

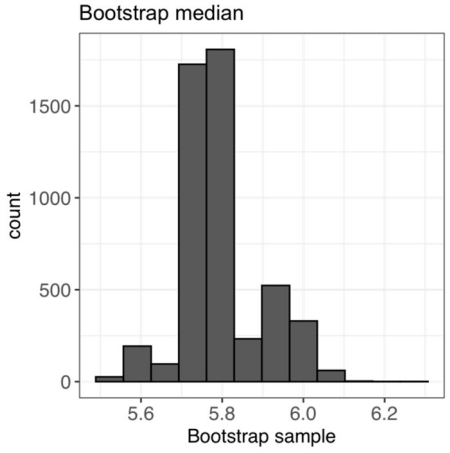
母集団の分布が明らかではない統計量に関してもブートスラップにより経験分布が推定できることを理解する．

解答例

R による解析例：

1.

```
library(boot)
library(simpleboot)
data(iris)
bootNum <- 5000 # ブートスラップサンプルの数
bs.median <- one.boot(iris$Sepal.Length,median,bootNum)
boot.ci(bs.median)
```



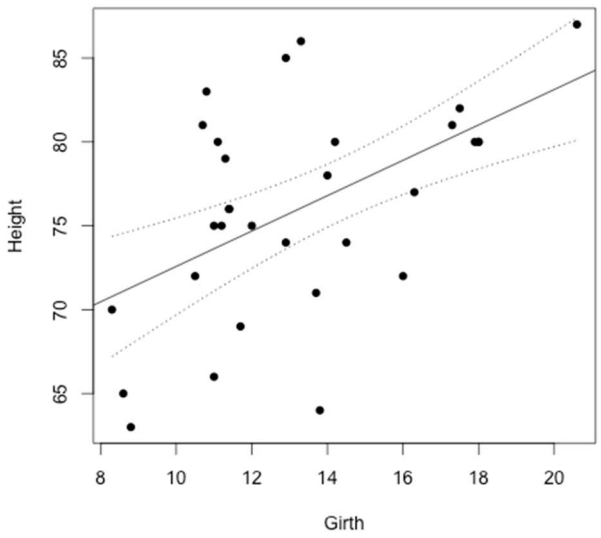

```
Intervals :
Level      Normal              Basic
95%   ( 5.612,  6.013 )   ( 5.600,  6.000 )

Level      Percentile            BCa
95%   ( 5.6,  6.0 )          ( 5.6,  6.0 )
```

2.

```
data("trees")


bootNum <- 1000 #  ブートスラップサンプルの数
fit_lm <- lm(Height~Girth,data=trees)
fit_bs <- lm.boot(fit_lm,R=bootNum)
plot(fit_bs,xlab="Girth", ylab="Height",pch=16)
```



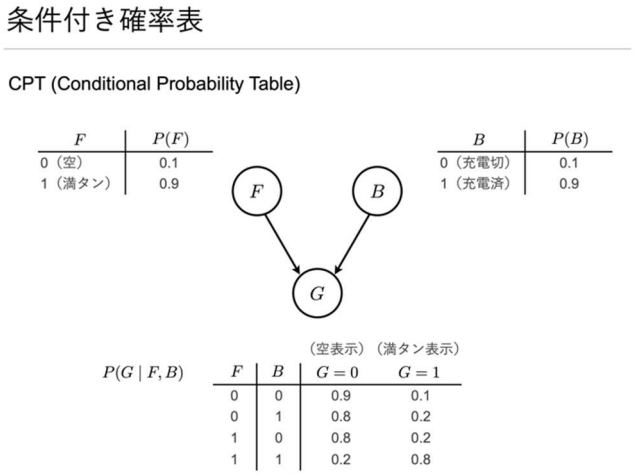
点が元データ，実線が単回帰モデルの当てはめ，点線が 95%信頼区間

留意点

乱数を用いた計算であるため，そのシードを固定しない限り同じ結果にはならない．安定した結果を得るには，できる限りブートストラップサンプルの数を増やすことが有効であることを説明する．

課題

1. 車の電気式燃料計の状態（G）が燃料タンクの状態（F）とバッテリーの状態（B）に依存した確率変数であり，以下の図に示す条件付き確率表が与えられている場合，燃料計の表示が空であったとき（G=0），その原因として燃料タンクが空（F=0）である事後確率を求めて下さい．



2. ピマ族の糖尿病の発症要因のデータセットにおいて，発症の有無（neg と pos）のカテゴリー変数を数値（0 と 1）に変換した上で，糖尿病の発症を予測するガウシアンネットワークの構造推定を Python または R で行って下さい．

本課題の狙い

ベイジアンネットワークの基礎となる条件付き確率からの事後確率の計算，および意思決定システムとしての応用について理解する．

解答例

1.

条件付き確率表より，G=0 の事前確率と F=0 である場合の G=0 の条件付き確率が以下のように求まる．

$$P(G=0) = \sum_{F\in\{0,1\}} \sum_{B\in\{0,1\}} P(G=0 \mid F,B)P(F)P(B) = 0.315$$

$$P(G=0 \mid F=0) = \sum_{B \in \{0,1\}} P(G=0 \mid F=0, B)P(B) = 0.81$$

これよりベイズの定理を用いて F=0 である事後確率が以下のように求まる.

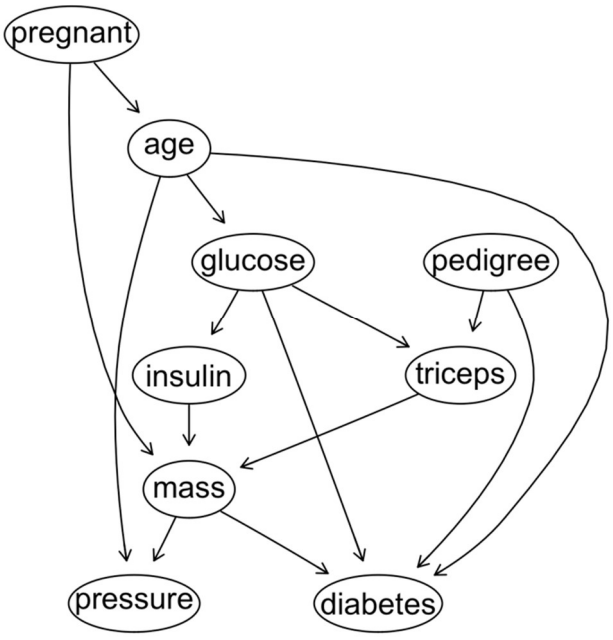
$$P(F=0 \mid G=0) = \frac{P(G=0 \mid F=0)P(F=0)}{P(G=0)} \simeq 0.257$$

2.

R による解析例：

```
library(bnlearn)
library(Rgraphviz)

df <- PimaIndiansDiabetes2
df$diabetes <- as.numeric(PimaIndiansDiabetes2$diabetes)-1 # 発症の変数を数値型に変換
#diabetes は他の変数の原因にはならない，pedigree（遺伝要因）は他の変数の結果にならないという拘束条件の設定
bl <- rbind(tiers2blacklist(list(colnames(df)[-9],"diabetes")),
            tiers2blacklist(list("pedigree",colnames(df)[-c(7,9)])))
dag <-hc(df,blacklist = bl) # 上記の拘束条件のもとでのベイジアンネットワークの構造推定
graphviz.plot(dag, shape = "ellipse") #結果の描画
```



留意点

ベイジアンネットワークは有向グラフィカルモデルなので，結果が原因に戻るようなループがあるネットワーク構造の推定には不向きであることを説明する.

課題

EM アルゴリズムを用いて混合正規分布のパラメータを学習する。以下の問いに答えよ。
適宜、必要な変数やデータを定義して答えること。

1. 混合正規分布の式を答えなさい。

2. 混合正規分布において、隠れ状態は何に相当するか答えなさい。

3. EM アルゴリズムの E-step を答えなさい。

4. EM アルゴリズムの M-step を答えなさい。

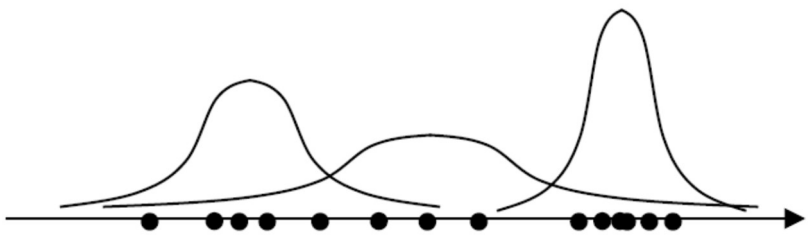
本課題の狙い

EM アルゴリズムを具体的な統計数理モデルに適用することよって、EM アルゴリズムの手順の理解を深めることを期待する。

解答例

1．混合正規分布は以下となる。
$\Theta = (\mu_y, \sigma_y, w_y)$ の各要素は、正規分布の平均、分散、重みパラメータである。

$$p(x|\boldsymbol{\theta}) = \sum_{y}^{K} \frac{w_y}{\sqrt{2\pi {\sigma_y}^2}} exp\left\{ -\frac{(x-\mu_y)^2}{2{\sigma_y}^2} \right\}$$



２．データがどの正規分布から生成されるかは観測できない。したがって、正規分布を指し示す番号 y が隠れ状態となる。

３．隠れ状態の分布を推定する E-step は以下となる。

$$p\left(y\middle|x,\boldsymbol{\theta}^{[t]}\right)=\frac{p\left(y,x\middle|\boldsymbol{\theta}^{[t]}\right)}{\sum_{y}^{K}p\left(y,x\middle|\boldsymbol{\theta}^{[t]}\right)}$$

$$=\frac{\frac{w_y}{\sqrt{2\pi{\sigma_y}^2}}exp\left\{-\frac{\left(x-\mu_y\right)^2}{2{\sigma_y}^2}\right\}}{\sum_{y=1}^{K}\frac{w_y}{\sqrt{2\pi{\sigma_y}^2}}exp\left\{-\frac{\left(x-\mu_y\right)^2}{2{\sigma_y}^2}\right\}}$$

４．混合正規分布のパラメータを最適化する M-step は以下となる。

$$\mu_y=\frac{\sum_{i=1}^{N}p(y\,|x_i,,\boldsymbol{\theta}^{[t]})\,x_i}{\sum_{i=1}^{N}p(y\,|x_i,,\boldsymbol{\theta}^{[t]})}$$

$${\sigma_y}^2=\frac{\sum_{i=1}^{N}p(y\,|x_i,,\boldsymbol{\theta}^{[t]})\left(x_i-\mu_y\right)^2}{\sum_{i=1}^{N}p(y\,|x_i,,\boldsymbol{\theta}^{[t]})}$$

$$w_y=\frac{\sum_{i=1}^{N}p(y\,|x_i,,\boldsymbol{\theta}^{[t]})}{\sum_{y=1}^{K}\sum_{i=1}^{N}p(y\,|x_i,,\boldsymbol{\theta}^{[t]})}$$

留意点

M-step の平均・分散の最適値は、最適性の一次の条件から導出、
重みパラメータの最適値は、ラグランジュ未定乗数法を用いること。

課題

データを予測する数理モデルは複数あります．隠れマルコフモデルやカルマンフィルタでもデータを予測することができます。以下の課題を行なってください．

1. 隠れマルコフモデル、カルマンフィルタで予測するための手順を簡単に説明してください。.
2. 隠れマルコフモデルとカルマンフィルタの数理モデルとして大きな違いを説明してください．

本課題の狙い

データサイエンスや人工知能では、様々な数理モデルが提案されている。同じ機能を実現することを目的としているが、各数理モデルの仮定・構造・計算手順によって、その性能や特徴は大きく異なる。各モデルの利点・欠点を把握したうえで、利用する数理モデルを取捨選択する必要があることを理解する。

解答例

１．隠れマルコフモデル、カルマンフィルタの予測手続きは以下となる

隠れマルコフモデル：

Step 1 観測データから状態（の分布）を推定する。

Step 2 状態遷移にしたがって次の時刻の状態（の分布）を計算する。

Step 3　計算された状態が持つ出力分布にしたがって、データを生成する。

カルマンフィルタ

Step 1 現在の状態を推定する。
（これは、通常の更新ルールにしたがって計算するだけである）

Step 2 推定された状態を状態方程式に代入すると、次の時刻の状態が計算できる。
（なお、さらにその状態を状態方程式に代入すると、2 時刻先の状態が
計算できることになる。）

Step 3 先述にて求められた状態を観測方程式に代入すると、
その状態にて観測されるであろうデータを計算することができる。

２．カルマンフィルタでは、状態の移り変わりが状態方程式として記述されており、
その方程式はあらかじめ人の手によって設計されている（状態方程式のパラメータは
あらかじめ決められた数値にて固定されている）。

一方、隠れマルコフモデルでは、状態の移り変わりが遷移確率として記述されており、
その確率パラメータは訓練データによって求められることになる。

カルマンフィルタには、訓練データから学習する機能がなく、その性能は、あらかじめ
決められた状態方程式に強く拘束されることになるのに対して、隠れマルコフモデル
では、状態の移り変わりを訓練データから学習することになり、モデルフリーの柔軟
性を有している。

留意点

２の解答だけを見ると隠れマルコフモデルのほうが優位であるように見えるかもしれないが、カルマンフィルタにも優れた利点が複数ある。各モデルの長所・短所を理解するように日ごろから心掛けるようにすること。

課題

サポートベクターマシンによって論理回路を作成する。下の問いに答えよ.

１．以下の真理値表で与えられる論理積

x_1	x_2	y
1	1	1
0	1	0
1	0	0
0	0	0

について、横軸にx_1、縦軸にx_2 として、y=0 のデータを●、y=1 のデータを○として、論理積の 4 つのデータを布置したグラフを描け。

２．y=0 と y ＝ 1 のデータを分離し、かつマージンが最大となる識別面を 1 のグラフに書き込め。

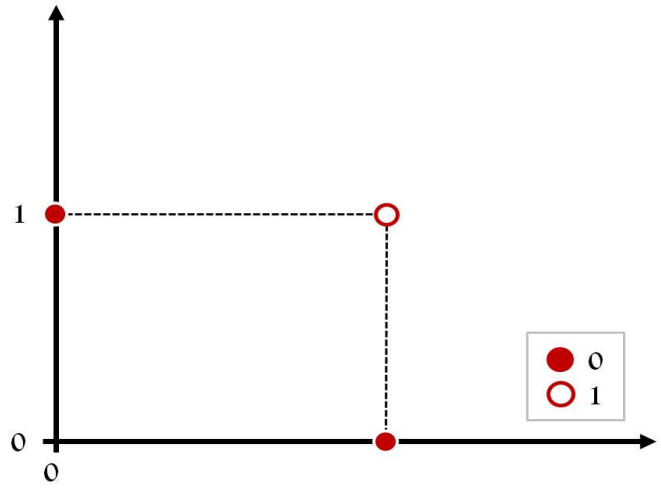
３．２で求めた識別面の式を求めよ。

本課題の狙い

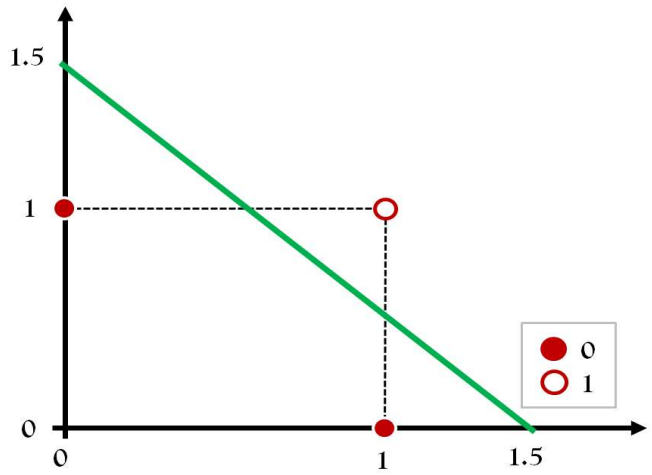
線形分離可能なデータを分離する識別面を求める方法として、サポートベクターマシンが広く利用されている。論理積のデータを具体例として、サポートベクターマシンの方針に従って、識別面を求めることにより理解が深められることを期待する。

解答例

1．論理積の 4 つのデータのグラフは以下となる。



2．マージンが最大となる識別面は以下となる。



3．識別面の式は、

$$-2\ x_1 - 2\ x_2 + 3 = 0$$

となる。

留意点

論理積は簡単な例であるが、実際にどのようなデータの分類にサポートベクターマシンが利用されているか調べて実用性を把握しておくこと。

課題

変分問題および変分ベイズについて、以下の問いに答えよ.

1. 汎関数

$$I[y] = \int_{x_0}^{x_1} F(x, y(x), y'(x))dx$$

を考える。上式は、独立変数x、関数 $y(x)$、およびその導関数 $y'(x)$を変数に持つ関数 $F(x, y, y')$ の積分である。この汎関数を最大もしくは最小にする関数 $y(x)$ を求める問題を変分問題と呼ぶ。

境界条件

$$y(x_0) = y_0, \quad y(x_1) = y_1$$

が与えられる場合に、汎関数の停留関数が満たすべき式を書け。

2. 観測データから統計モデルを学習するときに、1 で求めた関係式をどのように活用すればよいか答えよ.

本課題の狙い

汎関数は関数の関数であり、その汎関数を最大もしくは最小にする関数を求める問題が変分問題である。変分ベイズもその基礎に変分問題がある。詳細な式変形を理解するよりも、まずは変分問題と統計モデルがどのように関係付いて変分ベイズが生まれてくるのかという概ねのながれを理解することを期待する。

解答例

１．汎関数の停留関数は以下の関係式を満足することになる。

$$\frac{\partial F}{\partial y} - \frac{d}{dx}\left(\frac{\partial F}{\partial y'}\right) = 0$$

この関係式オイラー・ラグランジュ方程式と呼ぶ。この式の導出は応用コースの「変分ベイズ」の回で勉強してください。

２．観測データの集合 $X = \{x_1, x_2, \cdots, x_N\}$ の周辺尤度 $P(X)$ の対数尤度下限値は

$$L\big(q(z,\theta)\big) = \int q(z,\theta) \ln \frac{P(X,z,\theta)}{q(z,\theta)}\, dz\, d\theta$$

と表される。隠れ変数 z とモデルパラメータ θ の両方を表す変数を t とおくと

$$L\big(q(t)\big) = \int q(t) \ln \frac{P(X,t)}{q(t)}\, dt$$

となる。

$$I[y] = \int_{x_0}^{x_1} F(x,\, y(x), y'(x))dx$$

と比較して、$F(x,\, y(x), y'(x))$を$q(t) \ln \frac{P(X,t)}{q(t)}$とみなし、これをオイラー・ラグランジュの方程式に代入すると、変数の分布 $q(t)$ が求まる。

留意点

変分ベイズの導出はやや複雑なので、基礎コースで細かい式までは説明しません。「応用コース」で詳細な数学は勉強してください。

課題

データサイエンス基礎コース 13 の内容を踏まえて下記の項目に解答せよ．（適宜，参考書等を活用すること）

1. IRIS データに対して第一主成分と第二主成分以外の組み合わせでプロットせよ．

2. MNIST に対して，MLP やランダムフォレスト，ロジスティック回帰における正解率を比較せよ．

3. Scikit-Learn の LinearRegression と TensorFlow の RNN で予測した結果を比較せよ．

本課題の狙い

ここでは，教師あり，教師なし学習のプログラミングとモデル選択について Python を通して考慮している．設問 2 については，代表的な画像処理の対象として用いられる MNIST に対して，MLP やランダムフォレスト，ロジスティック回帰を用いて正解率を比較している．近年，ランダムフォレストの正解率が非常に良いということがわかるであろう．また，一般的に時系列解析には RNN を用いることが多いが，LinearRegression も良い結果を出すことがわかる．

留意点

各モデルの数学的・統計的なアルゴリズムについては深く議論していないが，残りの講義で各自が適宜，学習してほしい．

解答例

1. IRIS データに対して第一主成分と第二主成分以外の組み合わせでプロットせよ.

   下記の a と b に主成分の数を入力する.
   plt.scatter(feature[:,a],feature[:,b],alpha=0.8,c=list(df.iloc[:,0]))

2. MNIST に対して，MLP,svm,ランダムフォレスト，ロジスティック回帰における正解率を比較せよ.

   MLP：97%
   ランダムフォレスト：96%
   ロジスティック回帰：92%

3. Scikit-Learn の LinearRegression と TensorFlow の RNN で予測した結果を比較せよ.

   Scikit-Learn の LinearRegression と TensorFlow の RNN では共に 99.7%程度の正解率であり，両手法とも有用であった.

課題

ある画像データに対して、スパースモデリングを適用することを検討する。
以下の問いに答えよ。

1. 白黒の 8×8 のパッチ画像をベクトル $\boldsymbol{x}$ にて表現する場合、そのベクトルの次元数はいくらになるか答えよ。

2. パッチ画像を辞書ベクトル $\boldsymbol{d}$ と係数ベクトル $\boldsymbol{c}$ の積和として表現する場合、辞書ベクトルの次元数はいくらになるか答え。

$$\boldsymbol{x} = [\boldsymbol{d_1}, \boldsymbol{d_2}, \cdots, \boldsymbol{d_n}\ ]\ \boldsymbol{c}$$

3. 係数ベクトルの非ゼロの要素数が少数になるようにパッチ画像をスパース表現として近似する。スパース表現ができることの利点を説明せよ。

本課題の狙い

データが複雑・高次元化すると、メモリ・計算量といった様々な観点からデータを低次元化することが求められる。スパースモデリングも低次元化の一種であり、多様な分野で活用されている。その一例として、高次元の画像データにスパースモデリングを利用することを想定して、その有用性を具体的に理解することを期待する。

解答例

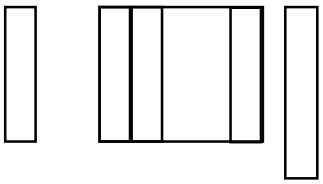
1. 64 次元

各画素値（例えば 0～255）をベクトルの各次元に対応させることによって、
パッチ画像を 8×8＝64 次元のベクトルとして表現できる。
画像中の(i,j) (i=1,2,…8, j=1,2,..8)の位置にある画素は、ベクトルの 8×(i-1) + j

次元目に対応することになる。i=j=1 の画素はベクトルの第 1 次元、i=j=8 の画素はベクトルの第 64 次元に対応することになる。

2．ベクトル・行列の式の形から、辞書ベクトルの次元数はパッチ画像ベクトルの次元数と同じになる。

$$\boldsymbol{x} = [\boldsymbol{d_1}, \boldsymbol{d_2}, \cdots, \boldsymbol{d_n}] \, \boldsymbol{c}$$



3．64 次元の画像データを係数ベクトルの少数の非ゼロ要素数で近似できることは、その数の係数のみで画像データが表現できることにある。すなわち、少数の数値データで表せるということなので、情報を圧縮していることに相当する。

また、ある画像をスパースモデリングを通じて係数ベクトルに近似・圧縮し、その係数ベクトルと辞書ベクトルの積和によって、画像を復元することもできる。
ノイズが乗っている画像に対して、上述のような変換を行うことによって、ノイズを除去できる可能性がある。

留意点

上記の問題は、スパースモデリングの活用例の一部に過ぎない。他の領域・問題に利用されている事例を調査して、理解を深めるように努めてください。

課題

以下のような２層ニューラルネットワークを考える。

$$\mathrm{z} = \sum_i w_i \ x_i$$

$$y = \frac{1}{1+\exp(-z)}$$

ここで、第１層の出力をx_i、第１層と第２層のシナプス結合の強さ（重みパラメータ）を w_i 第２層の入力値の積算をz、第２層の出力をｙとしている。出力層のノードは１個である。以下の問いに答えよ。

1. 訓練データとして、入力値が$(x_1, x_2, \cdots, x_n)$、出力値が t が与えられたとき、
訓練データとニューラルネットワークの誤差を求めよ。

２．訓練データの入出力関係を表現するニューラルネットワークを学習するための
目的関数を求めよ。

３．目的関数を最適化する重みパラメータの導出手順を述べよ。

本課題の狙い

ニューラルネットワークの学習法の１つである誤差逆伝搬法について、実際に式を書き、必要な計算を手で書き下すことによって、学習法の理解を深める。

解答例

１．入力値が$(x_1, x_2, \cdots, x_n)$であるとき、ニューラルネットワークの出力値は

$$z = \sum_i w_i \ x_i$$

$$y = \frac{1}{1+\exp(-z)}$$

となる。訓練データとの出力値との誤差 r は、上記 y を用いると

$$r = y - t$$

である。

２．上述の誤差は、正の時もあれば、負の時もある。誤差の大きさを評価するとして、
目的関数の候補の１つ（これを φ と置く）として、以下が考えられる。

$$\varphi = \frac{1}{2} r^2$$

３．勾配降下法を用いて最適な重みパラメータを数値的に解くことを考える。
勾配降下法における重みパラメータの更新則は、

$$w_i := w_i - \eta \frac{\partial \varphi}{\partial w_i}$$

$$\frac{\partial \varphi}{\partial w_i} = \frac{\partial \varphi}{\partial r} \frac{\partial r}{\partial y} \frac{\partial y}{\partial z} \frac{\partial z}{w_i}$$

のチェーンルールによって表現することができる。ただし、

$$\frac{\partial \varphi}{\partial r} = r$$

$$\frac{\partial r}{\partial y} = 1$$

$$\frac{\partial y}{\partial z} = \frac{\exp(-z)}{(1 + \exp(-z))^2}$$

$$\frac{\partial z}{w_i} = x_i$$

である。

留意点

この問題では訓練データを 1 個与えた時の重みパラメータの更新則を書き下した。実際は訓練データは複数個あるので、訓練データごとに上記計算を行い、それらを足し合わせるなどの手続きが必要になる。ただし、基本的な演算は上式であるので、この式変形をしっかり理解することを求める。

課題

畳み込みニューラルネットワークは画像認識などで優れた性能を実現している。画像処理における畳み込みニューラルネットワークに焦点を当て、以下の問いに答えよ。

1. 畳み込みニューラルネットワークでは、畳み込む層、プーリング層、パディング処理が広く利用されている。これらの機能・目的を簡潔に説明せよ。

２．２つの画像

0	1	0	0
0	1	0	0
0	1	0	0
0	1	0	0

0	0	0	0
1	1	1	1
0	0	0	0
0	0	0	0

に対して、以下のカーネルを施した時の結果を求めなさい

1	2	1
0	0	0
-1	-2	-1

３．上述のカーネルの機能を説明しなさい。

本課題の狙い

畳み込みニューラルネットワークの広く利用されるテクニックやその意味について、画像を題材にして理解することを目指す。

解答例

1．畳み込む層：エッジなどの特徴検出

プーリング層：画像の縮小、データ量の低減

パディング処理：画像の端でのデータ処理をするための画素の補足

2．

0	0
0	0

0	0
4	4

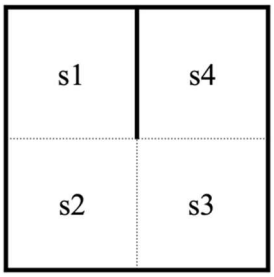
3. 縦方向の微分に相当し、水平方向のエッジを検出することができる。

留意点

深層ニューラルネットワークを単にブラックボックスと見るのでなく、その内部の処理が物理的にどういう意味があるのか把握するように努めること。

課題

図に示す 2×2 格子世界でのエージェントの最適行動を強化学習で学習します．この世界では，エージェントの状態は{s1,s2,s3,s4}の中からランダムに選ばれるとし，各状態において{up,down,left,right}の方向にある状態に移る行動を選択します．但し，格子世界の壁を乗り越えることはできませんし，s1 と s4 の間に設定された壁のためこの 2 つの状態間での直接の遷移はできません．また，最終ゴール s4 の状態にたどり着くと 10 の報酬が得られますが，一つ一つの行動については-1 の報酬とします（最短距離を学習したいため）．



1. 以上の設定のもとで，エージェント現在の状態（state）と行動（action）から，次の状態（next_state）と報酬（reward）を出力するプログラムを記述して下さい

2. Q 学習のアルゴリズムを用いて 1000 回のランダム行動による強化学習を行なった場合と，この学習結果に基づいて ε -greedy 法でさらに 1000 回の行動選択による強化学習を行なった場合で累積報酬を比較して下さい．また最終的な行動価値関数（Q 関数）を求めて下さい

本課題の狙い

強化学習の基本的な学習アルゴリズムである Q 学習を通して，強化学習の原理を理解する．

解答例

1.

```
function (state, action)
{    next_state <- state
     if (state == state("s1") && action == "down")
          next_state <- state("s2")
     if (state == state("s2") && action == "up")
```

```
        next_state <- state("s1")
    if (state == state("s2") && action == "right")
        next_state <- state("s3")
    if (state == state("s3") && action == "left")
        next_state <- state("s2")
    if (state == state("s3") && action == "up")
        next_state <- state("s4")
    if (next_state == state("s4") && state != state("s4")) {
        reward <- 10
    }
    else {
        reward <- -1
    }
    out <- list(NextState = next_state, Reward = reward)
    return(out) }
```

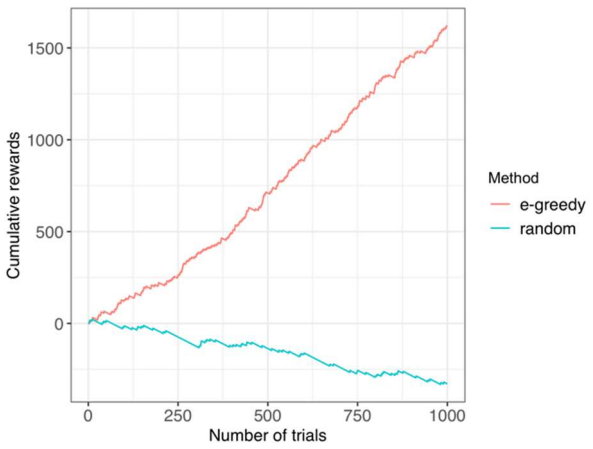
2.

R による解析例：

```
library(ReinforcementLearning)
# 状態と行動
states <- c("s1", "s2", "s3", "s4")
actions <- c("up", "down", "left", "right")
# 2x2 世界での行動と報酬の関数
env <- gridworldEnvironment
# ランダム行動
data_rnd <- sampleExperience(N = 1000, env = env, states = states, actions = actions) model_rnd<-
ReinforcementLearning(data_rnd, s = "State", a = "Action", r = "Reward",
s_new = "NextState")
# ε-greedy 法による行動選択
data_egd <- sampleExperience(N = 1000, env = env, states = states, actions = actions, actionSelection = "epsilon-
greedy",model=model_rnd)
model_ egd <- ReinforcementLearning(data_ egd, s = "State", a = "Action", r = "Reward",
s_new = "NextState",model= model_rnd)
```

ε-greedy法によるQ関数

	right	up	down	left
s1	-0.6488348	-0.6432422	0.7621807	-0.6329177
s2	3.5240753	-0.6507643	0.7609901	0.7498171
s3	3.5421453	9.0476033	3.5372558	0.7593446
s4	-1.9049642	-1.9150310	-1.9522237	-1.9108756



留意点

強化学習では，データが生成される環境を適切にモデリングすることがまず必要となる．この点においてはRよりPythonの方でライブラリが充実しているので，扱う問題のスケールが大きい場合はPython環境を最初から選んだ方がよい．

課題

1. 混合ユニグラムモデルとトピックモデル（潜在ディリクレ配分法：LDA）における単語データの生成過程について，プレート表現を用いた確率的グラフィカルモデルとして図示して下さい．また，両者の違いについて説明して下さい．

2. 1992 年度の AP 通信の配信記事のデータベース（AssociatedPress）から，潜在ディリクレ配分法を用いて文書データを 4 つのトピック分布でクラスタリングし，各トピック分布がどのような記事内容を表現しているのか考察して下さい．

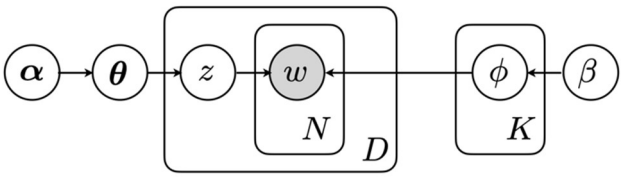
本課題の狙い

文書データ処理で広く使われている潜在ディリクレ配分法によるトピックモデルについて理解する．また，テキスト処理のプログラミングに習熟する．

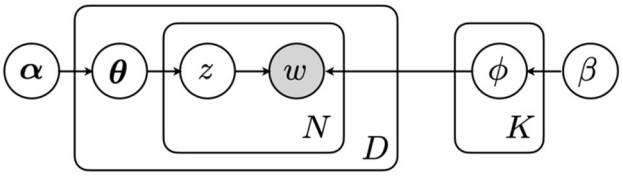
解答例

1.

混合ユニグラムモデル



トピックモデル（潜在ディリクレ配分法）



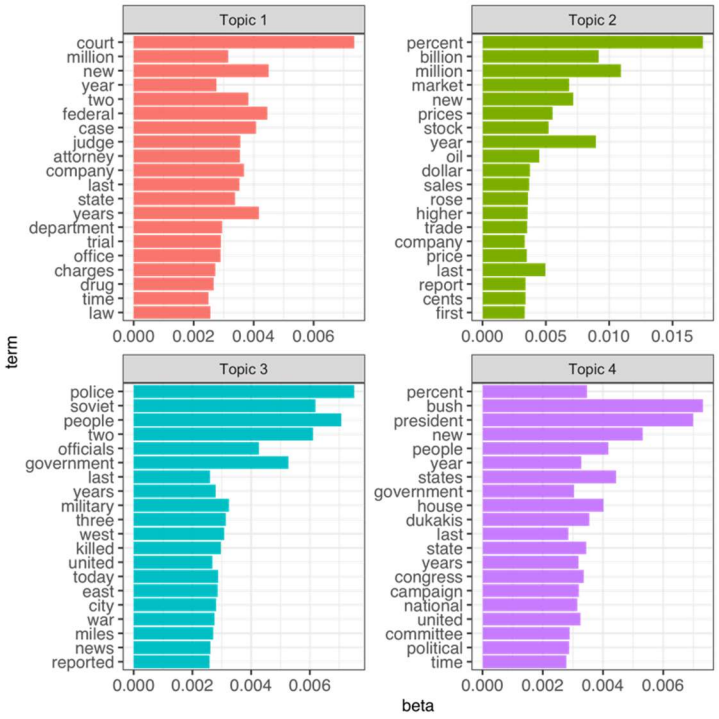
ここではパラメータ α をもつ K 個のトピック分布 θ それぞれについてパラメータ β をもつ単語分布 ϕ があるとしている．混合ユニグラムモデルではトピック分布 θ が D 個の文章集合に対して 1 つだけあり，文章ごとに 1 つのトピック z が割り当てられて N 個の単語 w が単語分布 ϕ から生成される．一方，トピックモデルでは，トピック分布 θ が D 個の文章そ

れぞれにあり，その分布に基づいて単語にトピック z が割り当てられ，対応する単語分布 ϕ から N 個の単語 w が生成される．

2.

R による解析例：

```
library(topicmodels)
data("AssociatedPress")
text_dtm <- AssociatedPress
## DocumentTermMatrix (DTM) 形式からテキストのリストに変換
dtm2list <- apply(text_dtm, 1, function(x) {paste(rep(names(x), x), collapse=" ")})
## tm パッケージの Corpus 形式に変換
text_corpus <- VCorpus(VectorSource(dtm2list))
## 前処理：単語を全て小文字化，数値の除去，ストップワードの除去
text_corpus_clean <- tm_map(text_corpus,content_transformer(tolower))
text_corpus_clean <- tm_map(text_corpus_clean, removeNumbers)
text_corpus_clean <- tm_map(text_corpus_clean,removeWords, stopwords())
text_dtm <- DocumentTermMatrix(text_corpus_clean) # Corpus から DTM へ変換
## LDA
text_lda <- LDA(text_dtm, k = 4, method = "VEM", control = NULL)
```



Topic1 は司法・犯罪，Topic2 は金融・経済，Topic3 はこの年に起こったソビエト連邦崩壊に起因した軍事関連の事件，Topic 4 はこの年の大統領選（ちなみデュカキス氏は一つ前の大統領選でのブッシュ氏の対立候補）のトピックとなっている．

留意点

テキストデータは一般に構造化されていないので，文章から解析に使うことのできる単語を抽出するための前処理が極めて重要であることを教示する．