機械学習の数理100問

鈴木 譲

平成 31 年 2 月 22 日

1 線形回帰

以下では、N, p を正の整数とする。

1. $x_1, \cdots, x_N, y_1, \cdots, y_N \in \mathbb{R}$ について、$\sum_{i=1}^{N}(y_i - \beta_0 - \beta_1 x_i)^2$ を最小にする $\beta_0, \beta_1 \in \mathbb{R}$ を $\hat{\beta}_0, \hat{\beta}_1$ とおくとき、以下の等式を示せ。ただし、$\bar{x}$ および $\bar{y}$ は、$\frac{1}{N}\sum_{i=1}^{N} x_i$ および $\frac{1}{N}\sum_{i=1}^{N} y_i$ で定義されるものとする。

   (a) $x_1, \cdots, x_N$ がすべて等しくはないとき、

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

   (b) $\hat{\beta}_0 + \hat{\beta}_1 \bar{x} = \bar{y}$

2. 問題 1 で得られた $\hat{\beta}_0$ および $\hat{\beta}_1$ をそれぞれ切片および傾きにする直線 l を考える。$x_1 - \bar{x}, \cdots, x_N - \bar{x}$ および $y_1 - \bar{y}, \cdots, y_N - \bar{y}$ から得られた直線を l' の切片と傾きを求めよ。また、$\hat{\beta}_1$ が求まってから、l の切片 $\hat{\beta}_0$ を求めるにはどうすればよいか。

3. 問題 2 の直線 l, l' の関係を可視化したい。空欄 (1)、空欄 (2) をうめて、グラフを図示せよ。

```
N=100                                  # サンプル数を N とする
a=rnorm(1); b=rnorm(1);                # 直線の係数をランダムに生成
x=rnorm(N); y=a*x+b+rnorm(N)           # 直線の周りの点をランダムに生成
plot(y~x)                              # 点のプロット
abline(lm(y~x),col="red")              # あてはめ直線
abline(h=0); abline(v=0)               # x 軸と y 軸
x=x-空欄 (1);  y=y-空欄 (2)             # N 組の点の重心を原点にうつす
plot(y~x,col="blue")                   # 点のプロット
abline(lm(y~x),col="blue")             # あてはめ直線
abline(h=0); abline(v=0)               # x 軸と y 軸
```

4. m, n を正の整数として、行列 $A \in \mathbb{R}^{m \times m}$ が、ある行列 $B \in \mathbb{R}^{n \times m}$ を用いて $A = B^T B$ とかけるとき、

   (a) A の 2 次形式が非負であることを示せ。

(b) $\lambda_1, \cdots, \lambda_m$ を A の固有値として、$A = P^T \begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \lambda_m \end{bmatrix} P$ なる直交行列 $P \in \mathbb{R}^{m \times m}$ が存在することを用いて、任意の $x \in \mathbb{R}^m$ および $y = Px = \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix}$ について、$x^T A x = \sum_{i=1}^m \lambda_i y_i^2$ および $\lambda_1, \cdots, \lambda_m \geq 0$ を示せ。ヒント: ある $1 \leq i \leq N$ について、$\lambda_i < 0$ であれば、$y_i = 1$, $y_j = 0$ $(j \neq i)$ となる y から得られる $x = P^{-1}y$ に対して、$0 \leq ||Bx||^2 = \lambda < 0$ となる。

(c) 任意の $z \in \mathbb{R}^m$ について、$Az = 0 \Longleftrightarrow Bz = 0$ を示せ。ヒント: $Az = 0 \Longrightarrow z^T B^T B z = 0 \Longrightarrow ||Bz||^2 = 0$ となる。

(d) A が正則となるための必要十分条件が、B の階数が m であることを示せ。ヒント: A, B の核が等しいので、両者の像の次元 (階数) が等しくなる。

本節の以下では、$X \in \mathbb{R}^{N \times (p+1)}$ を最初の列 (第 0 列) がすべて 1 の行列とする。

5. 以下の各場合について、$X^T X$ が逆行列をもたないことを示せ。

(a) $N < p + 1$ のとき。

(b) $N \geq p + 1$ であって、X のある 2 列が等しい場合。

本節の以下では、さらに $N \geq p + 1$ であって、$X \in \mathbb{R}^{N \times (p+1)}$ の階数が $p + 1$ であることを仮定する。

6. $X \in \mathbb{R}^{N \times (p+1)}$, $y \in \mathbb{R}^N$ から、$L := ||y - X\beta||^2$ が最小となる $\beta \in \mathbb{R}^{p+1}$ を求めたい。ただし、$||\cdot||$ は、$z \in \mathbb{R}^N$ に対して $\sqrt{\sum_{i=1}^N z_i^2}$ であると定義するものとする。

(a) X の第 (i, j) 成分を $x_{i,j}$ かくとき、$L = \sum_{i=1}^N (y_i - \sum_{j=0}^p x_{i,j}\beta_j)^2$ の β_j による偏微分が、$X^T y - X^T X \beta$ の第 j 成分と一致することを示せ。ヒント: $X^T y$ の第 j 成分は $\sum_{i=1}^N x_{i,j} y_i$、$X^T X$ の第 (j, k) 成分は $\sum_{i=1}^N x_{i,j} x_{i,k}$、$X^T X \beta$ の第 j 成分は $\sum_{k=0}^p \sum_{i=1}^N x_{i,j} x_{i,k} \beta_k$ となる。

(b) $\dfrac{\partial L}{\partial \beta} = 0$ なる $\beta \in \mathbb{R}^{p+1}$ を求めよ。以下では、この値を $\hat{\beta}$ とかくものとする。

7. 未知の定数 $\beta \in \mathbb{R}^{p+1}, \sigma^2 > 0$ と、確率変数 $\epsilon \sim N(0, \sigma^2 I)$ から、$y \in \mathbb{R}^N$ が $X\beta + \epsilon$ の実現値として得られ、問題 6 の手順にしたがって確率変数 $\hat{\beta}$ が得られるものとする。ただし、$I \in \mathbb{R}^{N \times N}$ は単位行列であるとする。

(a) $\hat{\beta} = \beta + (X^T X)^{-1} X^T \epsilon$ を示せ。

(b) $\hat{\beta}$ の平均が β に一致する ($\hat{\beta}$ が不偏推定量である) ことを示せ。

(c) $\hat{\beta}$ の共分散行列 $E(\hat{\beta} - \beta)(\hat{\beta} - \beta)^T$ が $\sigma^2 (X^T X)^{-1}$ となることを示せ。

8. $H := X(X^T X)^{-1} X^T \in \mathbb{R}^{N \times N}$、$\hat{y} = X\hat{\beta}$ とおくとき、以下の等式を証明せよ。

(a) $H^2 = H$

(b) $(I - H)^2 = I - H$

(c) $HX = X$

(d) $\hat{y} = Hy$

(e) $y - \hat{y} = (I - H)\epsilon$

(f) $||y - \hat{y}||^2 = \epsilon^T (I - H)\epsilon$

9. 以下を証明せよ。

(a) H の像の次元 (階数) は $p + 1$ である。ヒント: X の階数が $p + 1$ であることを仮定している。

(b) H が固有値 0 の $N-p-1$ 次の固有空間と、固有値 1 の $p+1$ 次の固有空間をもつ。ヒント: H は N 個の列をもつが、これは像の次元と核の次元の和になっている。

(c) $I-H$ が固有値 0 の $p+1$ 次の固有空間と、固有値 1 の $N-p-1$ 次の固有空間をもつ。ヒント: 任意の $x \in \mathbb{R}^{p+1}$ について、$(I-H)x=0 \Longleftrightarrow Hx=x$ および $(I-H)x=x \Longleftrightarrow Hx=0$ が成立する。

10. $P(I-H)P^T$ が対角行列 (最初の $N-p-1$ 個が 1、それ以外の $p+1$ 個が 0) となる直交行列 P を用いて、$v=P\epsilon$ を定義するとき、以下を示せ。

(a) $RSS := \epsilon^T(I-H)\epsilon = \sum_{i=1}^{N-p-1} v_i^2$。ヒント: P は直交行列なので $P^TP=I$ となり、$\epsilon = P^{-1}v = P^Tv$ を代入する。そして、$P^T(I-H)P$ は N 個の固有値を対角成分にもつ対角行列となる。特に、$I-H$ は固有値 1 が $N-p-1$ 個、固有値 0 が $p+1$ 個となる。

(b) $Evv^T=\sigma^2 I$。ヒント: $Evv^T = P(E\epsilon\epsilon^T)P^T$ と変形する。

(c) $RSS/\sigma^2 \sim \chi^2_{N-p-1}$ (自由度 $N-p-1$ の χ^2 分布)。ヒント: (a)(b) より、RSS の統計的性質がわかる

ただし、正規分布にしたがう 2 個以上の確率変数の共分散行列が対角行列になることとそれらが独立であることが同値になることは、証明無しで用いて良い。

11. (a) $E(\hat{\beta}-\beta)(y-\hat{y})^T = 0$ を示せ。ヒント: $(\hat{\beta}-\beta)(y-\hat{y})^T = (X^TX)^{-1}X^T\epsilon\epsilon^T(I-H)$ および $E\epsilon\epsilon^T=\sigma^2 I$ を用いる。

(b) $(X^TX)^{-1}$ の対角成分を $B_0,\cdots,B_p$ とおくとき、$(\hat{\beta}_i-\beta_i)/(\sqrt{B_i}\sigma)$ と RSS/σ^2 が独立であることを示せ。ただし、$i=0,1,\cdots,p$ とする。ヒント: RSS が $y-\hat{y}$ の関数なので、$y-\hat{y}$ と $\hat{\beta}-\beta$ の独立性に帰着される。正規分布にしたがうので、共分散が 0 であることと独立であることが同値になる。

(c) $\hat{\sigma} := \sqrt{\dfrac{RSS}{N-p-1}}$(残差標準誤差, residual standard error, σ の推定値)、および $SE(\hat{\beta}_i) := \hat{\sigma}\sqrt{B_i}$($\hat{\beta}_i$ の標準偏差の推定値) とおくとき

$$\frac{\hat{\beta}_i-\beta_i}{SE(\hat{\beta}_i)} \sim t_{N-p-1}$$

(自由度 $N-p-1$ の t 分布), $i=0,1,\cdots,p$ を示せ。ヒント:

$$\frac{\hat{\beta}_i-\beta_i}{SE(\hat{\beta}_i)} = \frac{\hat{\beta}_i-\beta_i}{\sigma\sqrt{B_i}} \Big/ \sqrt{\frac{RSS}{\sigma^2}/(N-p-1)}$$

を導き、右辺が t 分布にしたがうことを示す。

(d) $p=1$ のとき、第 1 列が $(x_{1,1},\cdots,x_{N,1})=(x_1,\cdots,x_N)$ であるとして、B_0 および B_1 を求めよ。ヒント: 以下を導く。

$$(X^TX)^{-1} = \frac{1}{\sum_{i=1}^N (x_i-\bar{x})^2} \begin{bmatrix} \frac{1}{N}\sum_{i=1}^N x_i^2 & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix}$$

ただし、正規分布にしたがう確率変数 $U_1,\cdots,U_m, V_1,\cdots,V_N$ の大きさ $m\times n$ の共分散行列が 0 であることと、U_i, V_j, $i=1,\cdots,m$, $j=1,\cdots,n$ が独立であることは、証明無しで用いてよい。

12. 帰無仮説 $H_0: \beta_i=0$、対立仮説 $H_1: \beta_i \neq 0$ の検定を行いたい。$p=1$ として、H_0 のもとで、

$$t = \frac{\hat{\beta}_i-0}{SE(\hat{\beta}_i)} \sim t_{N-p-1}$$

となることを用いて、以下の処理を構成した。ただし、関数 `pt(x,m)` は、自由度 m の t 分布の確率密度関数を f_m として、$\int_x^\infty f_m(t)dt$ の値をかえすものとする。

```
x=rnorm(N); y=rnorm(N)
x.bar=mean(x); y.bar=mean(y)
beta.0=sum(y.bar*sum(x^2)-x.bar*sum(x*y))/sum((x-x.bar)^2)
beta.1=sum((x-x.bar)*(y-y.bar))/sum((x-x.bar)^2)
RSS=sum((y-beta.0-beta.1*x)^2); RSE=sqrt(RSS/(N-1-1))
B.0=sum(x^2)/N/sum((x-x.bar)^2); B.1=1/sum((x-x.bar)^2)
se.0=RSE*sqrt(B.0); se.1=RSE*sqrt(B.1)
t.0=beta.0/se.0; t.1=beta.1/se.1
p.0=2*(1-pt(abs(t.0),N-2))  # p 値（その値より外側にある確率）
p.1=2*(1-pt(abs(t.1),N-2))  # p 値（その値より外側にある確率）
beta.0;se.0;t.0;p.0;
beta.1;se.1;t.1;p.1
```

lm 関数で、数値が正しいことを確認せよ。

```
lm(y~x)
summary(lm(y~x))
```

13. 下記は、前問の $\hat{\beta}_1$ を $r = 1000$ 回繰返し推定して、$\hat{\beta}_1/SE(\beta_1)$ のヒストグラムをとったものである。ただし、毎回発生させたデータから `beta.1/se.1` が計算され、(大きさ r の) ベクトルとして `T` に蓄積される。

```
N=100; r=1000
T=NULL
for(i in 1:r){
x=rnorm(N); y=rnorm(N); x.bar=mean(x); y.bar=mean(y)
fit=lm(y~x);beta=fit$coefficients
RSS=sum((y-fit$fitted.values)^2); RSE=sqrt(RSS/(N-1-1))
B.1=1/sum((x-x.bar)^2); se.1=RSE*sqrt(B.1)
T=c(T,beta[2]/se.1)
}
hist(T,breaks=sqrt(r),probability=TRUE, xlab="t の値",ylab="確率密度",
main="t の値のヒストグラムと理論値 (赤)")
curve(dt(x, N-2),-3,3,type="l", col="red",add=TRUE)
```

`y=rnorm(N)` を `y=0.1*x+rnorm(N)` におきかえて実行し、2 個のグラフの差異を説明せよ。

14. $W \in \mathbb{R}^{N\times N}$ の各成分を $1/N$ とし、$\bar{y} := \dfrac{1}{N}\sum_{i=1}^{N} y_i = Wy$ とかくとき、

    (a) $HW = W$ および $(I-H)(H-W) = 0$ を示せ。ヒント: W の各列は、H の固有値 1 の固有ベクトルなので、$HW = W$。

    (b) $ESS := ||\hat{y} - \bar{y}||^2 = ||(H-W)y||^2$ および $TSS := ||y - \bar{y}||^2 = ||(I-W)y||^2$ を示せ。

    (c) $RSS = ||(I-H)\epsilon||^2 = ||(I-H)y||^2$ と ESS は独立であることを示せ。ヒント: $(I-H)\epsilon$ と $(H-W)y$ の共分散行列は、$(I-H)\epsilon$ と $(H-W)\epsilon$ のそれと同じになる。共分散行列 $E(I-H)\epsilon\epsilon^T(H-W)$ を評価する。その際に (a) を用いる。

    (d) $||(I-W)y||^2 = ||(I-H)y||^2 + ||(H-W)y||^2$、すなわち $TSS = RSS + ESS$ を示せ。ヒント: $(I-W)y = (I-H)y + (H-W)y$ と変形する。

15. $X \in \mathbb{R}^{N\times(p+1)}$, $y \in \mathbb{R}^N$ に対して、

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$$

を決定係数 (coefficient of determination) という。$p = 1$ のとき、$x = [x_1, \cdots, x_N]^T$ として、

(a) $\hat{y} - \bar{y} = \hat{\beta}_1(x - \bar{x})$ を示せ。ヒント: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ および問題 1(b) を用いる。

(b) $R^2 = \dfrac{\hat{\beta}_1^2||x - \bar{x}||^2}{||y - \bar{y}||^2}$ を示せ。

(c) $p = 1$ のとき、R^2 の値が、相関係数の 2 乗に一致することを示せ。ヒント: $||x-\bar{x}||^2 = \sum_{i=1}^N (x_i - \bar{x})^2$ および問題 1(a) を用いる。

16. 下記の関数は、決定係数を求めている。ただし、入力の $X \in \mathbb{R}^{N\times p}$ には、すべて 1 の列が含まれていないものとする。

```
R2=function(x,y){
y.hat=lm(y~x)$fitted.values; y.bar=mean(y)
RSS=sum((y-y.hat)^2); TSS=sum((y-y.bar)^2)
return(1-RSS/TSS)
}
N=100; m=2; x=matrix(rnorm(m*N),ncol=m); y=rnorm(N); R2(x,y)
```

`N=100; m=1` として、`x=matrix(rnorm(m*N),ncol=m); y=rnorm(N); R2(x,y); cor(x,y)^2` を実行せよ。

17. 決定係数は、説明変数によって目的変数がどれだけ説明できているかをあらわす (最大値 1)。説明変数どうし冗長なものがないかどうかを測る尺度として、VIF(variance inflation factor) が用いられる。

$$VIF := \frac{1}{1 - R^2_{X_j|X_{-j}}}$$

ただし、$R^2_{X_j|X_{-j}}$ で、目的変数として $X \in \mathbb{R}^{N\times p}$ の第 j 列説明変数としてそれ以外の $p-1$ 個の列を用いる ($y \in \mathbb{R}^N$ は用いない)。VIF の値が大きいほど (最小値 1)、他の説明変数によって説明されている (共線形性 (colinearity) が強い) ことを意味する。MASS という R パッケージをインストールして、Boston というデータについて、VIF を計算せよ (下記を実行するだけでよい)。

```
library(MASS); X=as.matrix(Boston); p=ncol(X)
T=NULL; for(j in 1:p)T=c(T, 1/(1-R2(X[,-j],X[,j]))); T
```

18. 係数の推定値 $\hat{\beta}$ を用いて、N サンプルとは別の新しい点 $x_* \in \mathbb{R}^{p+1}$(列ベクトル、最初の成分が 1 で、それ以外には p 変数の値がはいる) における予測値 $x_*^T\hat{\beta}$ を計算できる。

(a) $x_*^T\hat{\beta}$ の分散が $\sigma^2 x_*^T(X^TX)^{-1}x_*$ となることを示せ。ヒント: $V(\hat{\beta}) = \sigma^2(X^TX)^{-1}$ を用いよ。

(b) $SE(x_*^T\hat{\beta}) := \hat{\sigma}\sqrt{x_*^T(X^TX)^{-1}x_*}$ とおくとき、

$$\frac{x_*^T\hat{\beta} - x_*^T\beta}{SE(x_*^T\hat{\beta})} \sim t_{N-p-1}$$

となることを示せ。ただし、$\hat{\sigma} = \sqrt{RSS/(N-p-1)}$ とした。

(c) 実際に生じる y の値は、$y_* := x_*^T\beta + \epsilon$ とできる。したがって、$y - x_*\hat{\beta}$ の分散は、σ^2 だけ大きくなる。

$$\frac{x_*^T\hat{\beta} - y}{\hat{\sigma}\sqrt{1 + x_*^T(X^TX)^{-1}x_*}} \sim t_{N-p-1}$$

を示せ。

19. 問題 18 より、自由度 $N-p-1$ の t 分布の確率密度関数を f として、$\alpha/2 = \int_t^\infty f(u)du$ なる t を $t_{N-p-1}(\alpha/2)$ とかくと、$y_* = x_*^T\beta + \epsilon$ の信頼区間として、

$$x_*^T\hat{\beta} \pm t_{N-p-1}(\alpha/2)\hat{\sigma}\sqrt{x_*^T(X^TX)^{-1}x_*}$$

(信頼区間, confident interval)、もしくは

$$x_*^T\hat{\beta} \pm t_{N-p-1}(\alpha/2)\hat{\sigma}\sqrt{1+x_*^T(X^TX)^{-1}x_*}$$

(予測区間, prediction interval) とできる。$p=1$ として、x_* を動かしていきながら、前者の区間のペアを赤で、後者の区間のペアを青でプロットしたい。信頼区間に関しては、下記を実行して、上限下限のグラフが得られる。予測区間に関しても、関数 $g(x)$ を定義して、信頼区間のグラフの上に、点線として重ね合わせて出力せよ。

```
#データを生成
N=100; p=1; X=matrix(rnorm(N*p),ncol=p); X=cbind(rep(1,N),X); beta=rnorm(p+1);
   epsilon=rnorm(N); y=X%*%beta*epsilon
#関数 f(x) を定義。 U は t(X)%*%X の逆行列
U=solve(t(X)%*%X); beta.hat=U%*%t(X)%*%y;
RSS=sum((y-X%*%beta.hat)^2); RSE=sqrt(RSS/(N-p-1)); alpha=0.05
f=function(x){
x=cbind(1,x); range=qt(df=N-p-1,1-alpha/2)*RSE*sqrt(x%*%U%*%t(x));
return(list(lower=x%*%beta.hat-range,upper=x%*%beta.hat+range))
}
#グラフで信頼区間を標示
x.seq=seq(-10,10,0.1)
lower.seq=NULL; for(x in x.seq)lower.seq=c(lower.seq, f(x)$lower)
upper.seq=NULL; for(x in x.seq)upper.seq=c(upper.seq, f(x)$upper)
x.lim=c(min(x.seq),max(x.seq)); y.lim=c(min(lower.seq),max(upper.seq))
plot(x.seq,lower.seq,col="blue",xlim=x.lim, ylim=y.lim, xlab="x",ylab="y", type="l")
par(new=TRUE);
plot(x.seq,upper.seq,col="red", xlim=x.lim, ylim=y.lim, xlab="",ylab="", type="l",
     axes=FALSE)
abline(beta.hat[1],beta.hat[2])
```

2 分類

20. $x \in \mathbb{R}^p$ に対して、$Y=1$ となる確率が $\dfrac{e^{\beta_0+\beta^Tx}}{1+e^{\beta_0+\beta^Tx}}$、$Y=-1$ となる確率が $\dfrac{1}{1+e^{\beta_0+\beta^Tx}}$ となるような $\beta_0 \in \mathbb{R}$, $\beta \in \mathbb{R}^p$ が存在する (ロジスティック分布) ことを仮定する。$Y=y \in \{-1,1\}$ となる確率が $\dfrac{1}{1+e^{-y(\beta_0+\beta^Tx)}}$ とかけることを示せ。

21. $\beta > 0$ として、関数 $f(x) = \dfrac{1}{1+e^{-(\beta_0+\beta x)}}$, $x \in \mathbb{R}$ が単調増加、$x < -\beta_0/\beta$ で下に凸、$x > -\beta_0/\beta$ で上に凸であることを示せ。また、β の値を大きくすると、どのような関数になるか。下記の処理を実行して確認せよ。

```
f=function(x)exp(beta.0+beta*x)/(1+exp(beta.0+beta*x))
beta.0=0; beta.seq=c(0,0.2,0.5,1,2,10); m=length(beta.seq)
beta=beta.seq[1]
```

```
plot(f,xlim=c(-10,10),ylim=c(0,1),xlab="x",ylab="y", col=1, main="ロジスティック曲線")
for(i in 2:m){
beta=beta.seq[i]
par(new=TRUE); plot(f,xlim=c(-10,10),ylim=c(0,1),xlab="", ylab="", axes=FALSE,,col=i)
}
legend("topleft",legend=beta.seq,col=1:length(beta.seq),lwd=2,cex=.8)
par(new=FALSE)
```

22. 観測値 $(x_1, y_1), \cdots, (x_N, y_N) \in \mathbb{R}^p \times \{-1, 1\}$ から、尤度 $\prod_{i=1}^{N} \frac{1}{1+e^{-y_i(\beta_0+\beta^T x_i)}}$ を最大にする、もしくは、その対数のマイナスをとった値

$$l(\beta_0, \beta) = \sum_{i=1}^{N} \log(1 + v_i) \ , \ v_i = e^{-y_i(\beta_0+\beta^T x_i)}$$

を最小にすることによって、$\beta_0 \in \mathbb{R}, \beta \in \mathbb{R}^p$ の推定値を得ることを考える (最尤推定)。微分 $\nabla l(\beta_0, \beta)$、2 回微分 $\nabla^2 l(\beta_0, \beta)$ を求め、$l(\beta_0, \beta)$ が凸であることを示せ。ヒント: $\nabla l(\beta_0, \beta)$ は $\frac{\partial l}{\partial \beta_j}$ を第 j 成分にもつ列ベクトル、$\nabla^2 l(\beta_0, \beta)$ は $\frac{\partial^2 l}{\partial \beta_j \partial \beta_k}$ を第 (j, k) 成分にもつ大きさ $(p+1) \times (p+1)$ の行列。非負定値であることを示せば十分である。$\nabla^2 l(\beta_0, \beta) = X^T W X$ の形に導き、W が対角行列であれば、$W = U^T U$ と分解でき (U の対角成分は、W の該当する成分の平方根)、$\nabla^2 l(\beta_0, \beta) = (UX)^T UX$ とできる。

23. 一般に、$f : \mathbb{R}^m \to \mathbb{R}$ について、$\nabla f(v) = 0, z \in \mathbb{R}^m$ という形式の方程式を解くときに、最初に v の初期値を与えてから、以下の式で v の値を更新することを繰り返して、収束を待つ方法がよく用いられる (Newton 法):

$$v_{new} \leftarrow v_{old} - \{\nabla^2 f(v_{old})\}^{-1} \nabla f(v_{old})$$

ここで、$\nabla f(v) \in \mathbb{R}^m$ は、第 i 成分が $\frac{\partial f}{\partial v_i}$ であるようなベクトル、$\nabla^2 f(v) \in \mathbb{R}^{m \times m}$ は、(i, j) 成分が $\frac{\partial^2 f}{\partial v_i \partial v_j}$ であるような正方行列であるとした。以下では、問題 23 で定義された $\nabla l(\beta_0, \beta) = 0$, $(\beta_0, \beta) \in \mathbb{R} \times \mathbb{R}^p$ を解くことを考える。ただし、記法の簡略化のため、$(\beta_0, \beta) \in \mathbb{R} \times \mathbb{R}^p$ を $\beta \in \mathbb{R}^{p+1}$ とあらわしている。

(a) 更新規則が

$$\beta_{new} \leftarrow (X^T W X)^{-1} X^T W z \tag{1}$$

となることを示せ。ただし、$\nabla l(\beta_{old}) = -X^T u$ なる $u \in \mathbb{R}^{p+1}$ と $\nabla^2 l(\beta_{old}) = X^T W X$ なる $W \in \mathbb{R}^{(p+1)\times(p+1)}$ を用いて、$z \in \mathbb{R}$ は $z := X\beta_{old} + W^{-1}u$ と定義されるものとする。ヒント: 更新規則は、$\beta_{new} \leftarrow \beta_{old} + (X^T W X)^{-1} X^T u$ ともかける。

(b) 更新規則 (1) は、以下のようにもかけることを示せ。

$$\beta_{new} \leftarrow \arg\min_{\beta} (z - X\beta)^T W (z - X\beta) \ .$$

ヒント: $||y - X\beta||^2 \to$ 最小 $\Longrightarrow X^T(y - X\beta) = 0$ と同様に、$(z - X\beta)^T W (z - X\beta) \to$ 最小 $\Longrightarrow X^T W (z - X\beta) = 0$ とできる。

24. 問題 31 を実現する処理を以下のように構成した。空欄 (1)(2)(3) をうめて、ただしく処理が求まることを確認せよ。

```
##  データ生成  ##
N=1000; p=2; X=matrix(rnorm(N*p),ncol=p); X=cbind(rep(1,N),X)
```

```
beta=rnorm(p+1); y=array(N); s=as.vector(X%*%beta); prob=1/(1+exp(s));
for(i in 1:N)if(runif(1)>prob[i])y[i]=1 else y[i]=-1
beta
##  最尤推定  ##
beta=Inf; gamma=rnorm(p+1)
while(sum((beta-gamma)^2)>0.001){
     beta=gamma
     s=as.vector(X%*%beta)
     v=exp(-s*y)
     u= ## 空欄 (1)
     w= ## 空欄 (2)
     z= ## 空欄 (3)
     W=diag(w)
     gamma=as.vector(solve(t(X)%*%W%*%X)%*%t(X)%*%W%*%z)
     print(gamma)
}
```

25. 観測値が $y_i(\beta_0 + \beta^T x_i) \geq 0$, $(x_i, y_i) \in \mathbb{R}^p \times \mathbb{R}$, $i = 1, \cdots, N$ を満足していると、ロジスティック回帰の最尤推定のパラメータを見出すことはできない。なぜか。

26. $p = 1$ として、関数 glm を用いて、$N/2$ 個ののサンプルでロジスティック回帰の係数を推定して (推定値 $\hat{\beta}_0, \hat{\beta}_1$)、$N/2$ 個の推定で用いなかったデータの x だけから、それらの y の値を予測してみた (この問題では、$Y = \pm 1$ ではなく、$Y = 0, 1$ の値をとるものとする)。最後に表示されるテーブルは何をあらわしているか。

```
N=1000
x=rnorm(N); beta=rnorm(2); y=sign(beta[1]+x*beta[2]+rnorm(N))
y=(y+1)/2  ## y を{0,1}に変更
fit=glm(y[1:(N/2)]~x[1:(N/2)],family="binomial")
     ## 1 番目から N/2 番目までのデータから係数を推定
beta.hat=fit$coefficients   ## 推定された係数を変数 beta.hat に格納
z=as.numeric(beta.hat[1]+x[(N/2+1):N]*beta.hat[2]>0)
     ## N/2+1 番目から N 番目までの y を予測
table(y[(N/2+1):N],z)
```

27. 線形判別で、π_k をクラス $Y = k$ の事前確率、$f_k(x)$ でクラス $Y = k$ のもとでの入力の p 変数の値 $x \in \mathbb{R}^p$ (平均 $\mu_k \in \mathbb{R}^p$、共分散行列 $\Sigma_k \in \mathbb{R}^{p \times p}$ の正規分布) の確率密度関数として、

$$\frac{\pi_k f_k(x)}{\sum_{j=1}^K \pi_j f_j(x)} = \frac{\pi_l f_l(x)}{\sum_{j=1}^K \pi_j f_j(x)}$$

となる $x \in \mathbb{R}^p$ の集合 $S_{k,l}$ を考える。

(a) $\pi_k = \pi_l$ とき、$S_{k,l}$ が 2 次曲面

$$-(x - \mu_k)^T \Sigma^{-1}(x - \mu_k) + (x - \mu_l)^T \Sigma^{-1}(x - \mu_l) = \log \frac{\det \Sigma_k}{\det \Sigma_l}$$

上の $x \in \mathbb{R}^p$の集合 になることを示せ。

(b) $\Sigma_k = \Sigma_l$ のとき (Σ とかくものとする)、$S_{k,l}$ が平面 $a^T x + b = 0$ ($a \in \mathbb{R}^P$, $b \in \mathbb{R}$) 上の $x \in \mathbb{R}^p$ になることを示し、a, b を $\mu_k, \mu_l, \Sigma, \pi_k, \pi_l$ を用いてあらわせ。

(c) $\pi_k = \pi_l$ かつ $\Sigma_k = \Sigma_l$ のとき、(b) の平面が $x = (\mu_k + \mu_l)/2$ になることを示せ。

28. 下記の空欄 1、空欄 2 をうめて、$y = \pm 1$ の判別を行う際の境界線をプロット中に加えよ。ただし、$y = \pm 1$ の事前確率は等しいものとする。

```
## データの生成
N=1000; x1=matrix(rnorm(N)+1,ncol=2); x2=matrix(rnorm(N)-1,ncol=2); x=rbind(x1,x2)
## 点のプロット
plot(x[,1],x[,2],type="n", xlab="x1",ylab="x2")
points(x[1:N/2,1],x[1:N/2,2],col="red"); points(x[(N/2+1):N,1],x[(N/2+1):N,2],col="blue")
## 境界線の計算とプロット
mu1=c(mean(x[1:N/2,1]),mean(x[1:N/2,2])); mu2=c(mean(x[(N/2+1):N,1]),mean(x[(N/2+1):N,2]))
inv=solve(cov(x))
A=inv%*%#空欄 (1)#
B=-0.5*t(mu1)%*%inv%*%mu1+#空欄 (2)#
abline(a=-as.numeric(B)/A[2],b=-A[1]/A[2])
```

29. Fisher のあやめのデータ (https://archive.ics.uci.edu/ml/datasets/iris) について、4 個の説明変数から、3 種類のあやめのいずれであるかを線形判別してみた。

```
library(MASS)
df=read.table("iris.txt",sep=","); N=nrow(df)
train=sample(N,N/2)  #1-150 のなかで、ランダムに半分を選んで、train とする
lda.fit=lda(V5~V1+V2+V3+V4, data=df[train,])
lda.pred=predict(lda.fit, df[-train,1:4])
table(lda.pred$class,df[-train,5])
```

同様の処理を 2 次判別で行なえ (2 次判別行う関数 qda も、MASS ライブラリに含まれている)。

30. 同様の処理を K-近傍法 (k-nearest neighbour) で行い、k を変えてみて、テストデータでの誤り率が最小の k を見い出せ。ヒント: R パッケージは class、関数は knn を用い、以下の形式で用いる。

```
knn(訓練データの説明変数、テストデータの説明変数、訓練データの目的変数、k=k の値)
```

31. ある病気にかかっている人のある測定値 x についての分布が $f_1(x)$、正常な人の分布が $f_0(x)$、θ を正の範囲で動かして、

$$\frac{f_1(x)}{f_0(x)} \geq \theta$$

であれば、症状にかかっている、そうでなければかかっていない、という判定を行うことにした。病気でない人が病気であると診断される条件付確率を横軸、病気である人が病気であると診断される条件付確率を縦軸であるグラフ (ROC) 曲線をえがき、その AOC(ROC の下の面積) を求めたい。下記では、病気の人、正常な人の分布を $N(1,1)$, $N(-1,1)$ としている。空欄をうめて、出力を pdf で提出せよ。

```
N.0=10000;N.1=1000; mu.1=1; mu.0=-1; var.1=1; var.0=1
x=rnorm(N.0,mu.0,var.0); y=rnorm(N.1,mu.1,var.1)      # x は病気でない人、y は病気の人
plot(1:N,1:N,xlim=c(0,1),ylim=c(0,1),
xlab="False Positive", ylab="True Positive", main="ROC 曲線", type="n")
theta.seq=exp(seq(-10,100,0.1))
U=NULL; V=NULL
for(theta in theta.seq){
u=sum(pnorm(x,mu.1,var.1)/pnorm(x,mu.0,var.0)>theta)/N.0 #病気でない人を病気とみなす
```

```
v= ## 空欄 ##                                                    病気の人を病気とみなす
U=c(U,u); V=c(V,v)
}
lines(U,V)
M=length(theta.seq)-1; AOC=0; for(i in 1:M)AOC=AOC+abs(U[i+1]-U[i])*V[i]
text(0.5,0.5,paste("AOC=",AOC))
```

32. 以下は、あわびの大きさから、大人か否かを予測する処理である。データ・セット abalon の最初の列は"F"(female), "M"(male), "I"(infant) のいずれか、2 番目の列は貝殻の大きさ (mm) である。また、R パッケージ ROCR は、ロジスティック回帰で各あわびが大人か否かを予測し、ROC 曲線を描くためのものである。実際に、UCI Machine Learning Depository (https://archive.ics.uci.edu/ml/datasets/Abalone) からデータをダウンロードして、ROC 曲線を描け。

```
library(ROCR)
rocdata = read.table("abalon.txt",sep=",")
rocdata[,1]=as.integer(rocdata[,1]!="I")    # 最初の列が"I"(子供) か否か
pred = prediction(rocdata[,2],rocdata[,1])  # 各あわびの大きさから大人か子供かを予測
perf = performance(pred, "tpr", "fpr") # 偽陽性率が横軸、真陽性率が縦軸の性能評価
plot(perf)
```

3 リサンプリング

33. $m \geq 1$、 $A \in \mathbb{R}^{m\times m}$ を正則、$b \in \mathbb{R}^m$ が $b^TA^{-1}b \neq 1$ を満足するとして、

$$(A - bb^T)^{-1} = A^{-1} + \frac{1}{1 - b^TA^{-1}b}A^{-1}bb^TA^{-1} \tag{2}$$

を示せ (Sherman-Morroson-Woodbury)。ヒント: $b^TA^{-1}b$ がスカラーになることに注意して、以下を導く。

$$\begin{aligned} & (A - bb^T)(A^{-1} + \frac{1}{1 - b^TA^{-1}b}A^{-1}bb^TA^{-1}) \\ = \;& I - bb^TA^{-1} + \frac{bb^T}{1 - b^TA^{-1}b}A^{-1} - \frac{1}{1 - b^TA^{-1}b}bb^TA^{-1}bb^TA^{-1} = \cdots = I \end{aligned}$$

34. $X \in \mathbb{R}^{N\times(p+1)}$、$y \in \mathbb{R}^N$、$x_i = [x_{i,0}, \cdots, x_{i,p}] \in \mathbb{R}^{1\times(p+1)}$、$X[i] := [x_1, \cdots, x_{i-1}, x_{i+1}, \cdots, x_N]^T \in \mathbb{R}^{(N-1)\times(p+1)}$、$y[i] = [y_1, \cdots, y_{i-1}, y_{i+1}, \cdots, y_N]^T \in \mathbb{R}^{N-1}$ として、

(a) $h_i := x_i(X^TX)^{-1}x_i^T$ を $H = X(X^TX)^{-1}X^T$ の第 i 成分として、

$$(X[i]^TX[i])^{-1} = (X^TX)^{-1} + \frac{(X^TX)^{-1}x_i^Tx_i(X^TX)^{-1}}{1 - h_i}$$

を示せ。ヒント: $X[i]^TX[i] = X'X - x_i^Tx_i$ を用い、$m = p + 1$, $A = X^TX$, $b = x_i^T$ を (2) に適用する。

(b) $e_i = y_i - \hat{y}_i$ として、以下を示せ。

$$\hat{\beta}[i] \;\; := \;\; (X[i]^TX[i])^{-1}X[i]^Ty[i] = \hat{\beta} - \frac{(X^TX)^{-1}x_i^Te_i}{1 - h_i} \;,$$

ヒント: $X[i]^Ty[i] = X^Ty - x_i^Ty_i$ より、

$$\begin{aligned} & \{(XX^T)^{-1} + \frac{(X^TX)^{-1}x_ix_i^T(X^TX)^{-1}}{1 - h_i}\}(X^Ty - x_i^Ty_i) \\ = \;& \hat{\beta} - (X^TX)^{-1}x_i^Ty_i + \frac{(X^TX)^{-1}x_i^Tx_i(X^TX)^{-1}X^Ty}{1 - h_i} - \frac{(X^TX)^{-1}x_i^Tx_i(X^TX)^{-1}}{1 - h_i}x_i^Ty_i \\ = \;& \hat{\beta} - \frac{(X^TX)^{-1}x_i^T}{1 - h_i}[(1 - h_i)y_i - x_i\hat{\beta} + h_iy_i] \end{aligned}$$

35. $y_i - x_i\hat{\beta}[i] = \dfrac{e_i}{1-h_i}$ を示すことによって、LOOCV (leave one out cross-validation) の N グループの 2 乗誤差の和が $\sum_{i=1}^{N}(\dfrac{e_i}{1-h_i})^2$ となることを示せ。また、$0 \leq h_i \leq 1$ を示せ。ヒント: $y_i - x_i\hat{\beta}[i] = y_i - x_i\{\hat{\beta} - \dfrac{(X^TX)^{-1}x_i^Te_i}{1-h_i}\}$。また、$H$ の固有値が 0,1 のみなので、$I-H$ の固有値もそれらからなる。したがって、ともに非負定値である。第 i 成分のみ 1、それ以外 0 の大きさ N の行ベクトル、列ベクトルを、それぞれそれらの前と後ろからかける。

36. 以下の手順は、$k = 10$ について、クロスバリデーションで、k 近傍法による分類の評価を行っている。空欄をうめ、$k = 5$ の場合と性能を比較せよ。

```
library(class)
df=read.table("iris.txt",sep=","); top.seq=1+seq(0,135,15); S=0
for(top in top.seq){
     index= ## 空欄 (1)
     knn.ans=knn(df[-index,1:4],df[index,1:4],df[-index,5],k=10)
     ans= ## 空欄 (2)
     S=S+sum(knn.ans!=ans)/15
}
S=S/10
```

37. 下記のデータセットの説明変数 V3,V4,V5 の中の最良の組合わせを選びたい
`https://web.stanford.edu/~hastie/StatLearnSparsity/data.html`

(a) `V1~V3+V4` のような関係を 7 通り (説明変数を少なくとも 1 つ用いる) のすべてについてかえて、最適な組合わせを選べ。

```
df=read.table("crime.txt")
fit=glm(V1~V3+V4,data=df)  ## (a) ではここをかえる
cv.fit=cv.glm(df,fit,K=10)  ## (b) ではここをかえる
cv.fit$delta[1]
```

(b) 3 行目をかえてみて、$K = 1$ と $K = 10$ とで、どちらが正しい答えになるか。また、どちらが早いか。ヒント: `cv.fit$delta[1]` の値を比較せよ。

38. 変数 X, Y に関する N 個のデータから、以下の量の標準偏差を推定したい。

$$\frac{v_y - v_x}{v_x + v_y - 2v_{xy}}\ ,\ \left\{\begin{array}{lcl} v_x & := & \frac{1}{N-1}[\sum_{i=1}^{N} X_i^2 - N\{\sum_{i=1}^{N} X_i\}^2] \\ v_y & := & \frac{1}{N-1}[\sum_{i=1}^{N} Y_i^2 - N\{\sum_{i=1}^{N} Y_i\}^2] \\ v_{xy} & := & \frac{1}{N-1}[\sum_{i=1}^{N} X_iY_i - N\{\sum_{i=1}^{N} X_i\}\{\sum_{i=1}^{N} Y_i\}] \end{array}\right.$$

そのために、重複を許して、N 行をランダムに選んでその値を計算し、それを r 回繰り返して、その値の標準偏差を推定した (Bootstrap). 空欄 (1)(2) をうめて処理を完成させ、 標準偏差を推定していることを確認せよ。

```
func.1=function(data,index){
     X=data$X[index]; Y=data$Y[index]
     return((var(Y)-var(X))/(var(X)+var(Y)-2*cov(X,Y)))
}
bt=function(df,func,r){
     m=nrow(df); org=## 空欄 (1)
     u=array(dim=r)
```

```
    for(j in 1:r){
        index=sample(## 空欄 (2)
        u[j]=func(df,index)
    }
    return(list(original=org, bias=mean(u)-org, stderr=sd(u)))
}
library(ISLR); bt(Portfolio,func.1,1000)  ## 実行例
```

39. 線形回帰の係数を推定する場合、雑音が正規分布にしたがうことを仮定すると、標準偏差の理論値が計算できる。その値と Bootstrap で求めた値とを比較したい。空欄をうめて、実行せよ。ただし、今回は、boot パッケージをインストールして、boot 関数を実行せよ。

```
library(boot); df=read.table("crime.txt")
boot.fn=function(df,index)coef(lm(V1~V3+V4,data=df,subset=index))
boot(## 空欄
summary(lm(V1~V3+V4,data=df,subset=index))
```

4 情報量基準

この節では、以下のようにおくものとする。

$$X = \begin{bmatrix} x_1 \\ \vdots \\ x_N \end{bmatrix} \in \mathbb{R}^{N\times(p+1)}, y = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} \in \mathbb{R}^N, z = \begin{bmatrix} z_1 \\ \vdots \\ z_N \end{bmatrix} \in \mathbb{R}^N, \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} \in \mathbb{R}^p$$

すなわち $x_1, \cdots, x_N$ は行ベクトルとし、またそれらの最初の成分が 1 であるものとする。また、

$$f(y|x,\beta) := \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{(y-x\beta)^2}{2\sigma^2}\}$$

とおく。さらに、確率密度関数 $\prod_{i=1}^N f(y_i|x_i,\beta)$ による平均の操作を $E_Y[\cdot]$ とおくものとする。

41. $X \in \mathbb{R}^{N\times(p+1)}$, $y \in \mathbb{R}^N$ から、下記の量を最大にする $\beta \in \mathbb{R}^{p+1}$、$\sigma^2 > 0$ を求めたい。以下のそれぞれを示せ。

    (a) $\sigma^2 > 0$ を既知として、$l := \sum_{i=1}^N \log f(y_i|x_i,\beta)$ を最大にする $\beta \in \mathbb{R}^{p+1}$ は、問題 1 の最小 2 乗解と一致する。ヒント:
    $$l = -\frac{N}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}||y - X\beta||^2$$

    (b) $\beta \in \mathbb{R}^{p+1}$、$\sigma^2 > 0$ ともに未知であるとして、σ^2 の最尤推定量は、以下で与えられる。
    $$\hat{\sigma}^2 = \frac{1}{N}||y - X\hat{\beta}||^2 .$$
    ヒント: l を σ^2 で偏微分すると以下のようになる。
    $$\frac{\partial l}{\partial \sigma^2} = -\frac{N}{2\sigma^2} + \frac{||y-X\beta||^2}{2(\sigma^2)^2} = 0 .$$

    (c) $\sigma^2 > 0$ の値は既知であるとして、$\nabla l \in \mathbb{R}^{p+1}$ の共分散行列が、$\nabla^2 l$ に等しいことを示せ (フィッシャー情報量行列)。

42. $l := \sum_{i=1}^N \log f(y_i|x_i,\beta)$、$\tilde{\beta} \in \mathbb{R}^{p+1}$ を β の任意の不偏推定量として、以下を示せ。

(a) $E\nabla l = 0$。ヒント: $l := \log f^N(y|x,\beta)$ と略記すると、$\nabla l = \dfrac{\nabla f^N(y|x,\beta)}{f^N(y|x,\beta)}$ とできる。

(b) $E[\tilde{\beta}_j \dfrac{\partial l}{\partial \beta_j}] = 1$。ヒント: $\int \tilde{\beta}_j f^N(y|x,\beta)dy = \beta_j$ の両辺を β_j で偏微分する。

(c) $E[(\tilde{\beta}-\beta)^T \nabla l] = p+1$。ヒント: $E[\beta^T \nabla l] = \beta^T E\nabla l = 0$ が成立する。

(d) 確率変数 $U, V \in \mathbb{R}^m$ $(m \geq 1)$ について、$\{E[U^T V]\}^2 \leq E[||U||^2]E[||V||^2]$。ヒント: $E[||tU+V||^2] = t^2 E[||U||^2] + 2tE[U^T V] + E[||V||^2] \geq 0$ であるので、その判別式が非負になる。

(e) $\{E(\tilde{\beta}-\beta)^T \nabla l\}^2 \leq E||X(X^T X)^{-1}\nabla l||^2 E||X(\tilde{\beta}-\beta)||^2$

(f) 行列 $A \in \mathbb{R}^{m\times n}, B \in \mathbb{R}^{n\times m}$ について $(m,n \geq 1)$、積 AB, BA のトレースが等しい。

(g) $E||X(X^T X)^{-1}\nabla l||^2 = \sigma^2(p+1)$。ヒント: 問題 41(c) と問題 42(e) を用いる。

(h) $E||X(\tilde{\beta}-\beta)||^2 \geq \sigma^2(p+1)$

43. $\mathbb{R}$ 上の確率密度関数 f, g について、その Kullback-Leibler 情報量が非負、すなわち

$$D(f||g) := \int_{-\infty}^{\infty} f(x) \log \frac{f(x)}{g(x)} dx \geq 0$$

が成立することを示せ。ヒント: 一般に、$\log x \leq x-1$, $x > 0$ が成立する。

44. $x \in \mathbb{R}^{p+1}$(行ベクトル), $z \in \mathbb{R}$, $\beta \in \mathbb{R}^{p+1}$, $\sigma^2 \neq 0$ について、

$$f(z|x,\beta) := \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{(z-x\beta)^2}{2\sigma^2}\}$$

と定義する。

(a) 以下を示せ。

$$\log f(z|x,\gamma) = -\frac{1}{2}\log 2\pi\sigma^2 - \frac{1}{2\sigma^2}(z-x\beta)^2 + \frac{1}{\sigma^2}(\gamma-\beta)^T x^T (z-x\beta) - \frac{1}{2}(\gamma-\beta)^T x^T x(\gamma-\beta)$$

ヒント: 以下を用いよ。

$$\nabla \log f(z|x,\gamma)|_{\gamma=\beta} = \frac{x^T(z-x\beta)}{\sigma^2}\ ,\ \nabla^2 \log f(z|x,\gamma)|_{\gamma=\beta} = -\frac{x^T x}{\sigma^2}\ ,\nabla^3 \log f(z|x,\gamma)|_{\gamma=\beta} = 0$$

(b) $(x_1,z_1),\cdots,(x_1,z_N) \in \mathbb{R}^{(p+1)} \times \mathbb{R}$ について、以下を示せ。

$$-\sum_{i=1}^N \log f(z_i|x_i,\gamma) = \frac{N}{2}\log 2\pi\sigma^2 + \frac{1}{2\sigma^2}||z-X\beta||^2 - \frac{1}{\sigma^2}(\gamma-\beta)^T X^T(z-X\beta) + \frac{1}{2}(\gamma-\beta)X^T X(\gamma-\beta)$$

ヒント: 以下を用いる: $||z-X\beta||^2 = \sum_{i=1}^N (z_i - x_i\beta)^2$、$X^T X = \sum_{i=1}^N x_i^T x_i$、$X^T(z-X\beta) = \sum_{i=1}^N x_i^T(z_i - x_i\beta)$

(c) $z_1,\cdots,z_N$ が $f(z_i|x_1,\beta),\cdots,f(z_N|x_N,\beta)$ にしたがって発生したと仮定するとき、(b) の平均が以下であたえられることを示せ。

$$-E_Z \sum_{i=1}^N \log f(Z_i|x_i,\gamma) = \frac{N}{2}\log(2\pi\sigma^2 e) + \frac{1}{2\sigma^2}||X(\gamma-\beta)||^2 \tag{3}$$

45. $y_1,\cdots,y_N$ が $f(y_i|x_1,\beta),\cdots,f(y_N|x_N,\beta)$ にしたがって発生したとき、それらから β を推定し、平均的に Kullback-Leibler 情報量

$$E_Z \sum_{i=1}^N \log \frac{f(Z_i|x_i,\beta)}{f(Z_i|x_i,\gamma)}$$

を最小にする不偏推定量を γ として用いたい。

(a) 最小 2 乗法を用いる (推定量を $\hat{\beta} \in \mathbb{R}^{p+1}$ とかく) と、$E||X(\hat{\beta}-\beta)||^2 = \sigma^2(p+1)$ になることを示せ。

(b) γ として、最小 2 乗法の推定量 $\hat{\beta}$ を用いると、不偏推定量の中で、(3) の右辺が平均的に最小になり、

$$-E_Y E_Z \sum_{i=1}^{N} \log f(Z_i|x_i, \hat{\beta}) = \frac{N}{2}\log(2\pi\sigma^2 e) + \frac{1}{2}(p+1) \tag{4}$$

となることを示せ。また、

$$E_Y E_Z \sum_{i=1}^{N} \log \frac{f(Z_i|x_i,\beta)}{f(Z_i|x_i,\hat{\beta})}$$

が最小になることを示せ。

46. これまでは、説明変数が p 個の場合の β, σ^2 の最尤推定 (最小二乗推定に等しい) を検討した。この問題では、説明変数の個数が $k(\leq p)$ であり、σ_k^2, $\hat{\sigma}_k^2$ をそのときの分散とその推定値であるとして、、$\log\hat{\sigma^2}$ が平均的に $\log\sigma^2$ より、どれだけ小さくなるかを評価する。以下では、真の変数の個数が k であって、k の値は既知として、σ^2 を推定するものとする。

(a)

$$M := \log(\frac{\hat{\sigma}_k^2}{N-k-1}/\frac{\sigma_k^2}{N}) = \frac{RSS}{(N-k-1)\sigma_k^2} - 1 - \frac{1}{2}\{\frac{RSS}{(N-k-1)\sigma_k^2} - 1\}^2 + \cdots$$

の平均値が、k の値によらず、$-\frac{1}{N} + O(1/N^2)$ となることを示せ。ただし、自由度 m の χ^2 分布にしたがう確率変数 X について、平均が $\mu = EX = m$、モーメントが

$$E(X-\mu)^n = 2^{n-1}(n-1)!m$$

$n = 2, 3, \cdots$, $m = 1, 2, \cdots$ となることは、用いて良いものとする。また、$O(f(N))$ とかいて、$g(N)/f(N)$ が有界な任意の N の関数 $g(N)$ をあらわすものとする。ヒント: $\frac{RSS}{(N-k-1)\sigma_k^2}$ の平均と分散を求めよ。

(b) $E_Y[\log\hat{\sigma}_k^2] = \log\sigma_k^2 - \frac{k+2}{N} + O(1/N^2)$ を示せ。ヒント: $\log\frac{N-k-1}{N} = -\frac{k+1}{N} + O(1/N^2)$ を用いる。

(c) (4) で p を k にかえて、$N\log\sigma^2$ に $N\log\hat{\sigma}_k^2 + k + 2$ を代入した値が、以下の値の単調増加になることを示せ。

$$AIC := N\log\hat{\sigma}_k^2 + 2k \tag{5}$$

47. 下記は、AIC の値を求める処理を記述したものである。空欄をうめて、処理を実行せよ。データセット"crime.txt"は、以下から得られる

https://web.stanford.edu/~hastie/StatLearnSparsity/data.html

```
crime=read.table("crime.txt"); X=as.matrix(crime[,3:7]); y=crime[,1];
p=ncol(X);n=length(y)
AIC.min=Inf
for(k in 1:p){
     T=combn(1:p,k); m=ncol(T)
     S.min=Inf
     for(j in 1:m){
          q=T[,j]; S=sum((lm(y~X[,q])$fitted.values-y)^2)/n
          if(S<S.min){S.min=S; set.q=q}
     }
     AIC= ##空欄 (1)##
```

```
    if(AIC<AIC.min){AIC.min= ##空欄 (2)## ; set.min= ##空欄 (3)##}
}
AIC.min
set.min
```

48. (5) ではなく、以下の値を最小化する別の変数選択を考える (BIC, Bayesian Information Criterion)。

$$N \log \hat{\sigma}^2 + k \log N$$

AIC の処理の該当する行を置き換え、関数名を AIC ではなく BIC とせよ。そして、同じデータについて、BIC を実行せよ。さらに

$$AR^2 := 1 - \frac{RSS/(N-k-1)}{TSS/(N-1)}$$

(調整済み決定係数) を最大にする変数選択について、処理を構成し、関数名を AR2 とせよ。そして、同じデータについて、AR2 を実行せよ。

49. AIC、BIC を最小にする k が何であるかを視覚化させたい。空欄をうめて、処理を実行させよ。

```
X=as.matrix(Boston[,1:13])
y=Boston[[14]]

IC=function(k){
T=combn(1:p,k); m=ncol(T)
S.min=Inf
for(j in 1:m){
   q=T[,j]; S=sum((lm(y~X[,q])$fitted.values-y)^2)/n
   if(S<S.min)S.min=S
}
AIC= ## 空欄 (1) ##
BIC= ## 空欄 (2) ##
return(list(AIC=AIC,BIC=BIC))
}

AIC.seq=NULL; BIC.seq=NULL;
for(k in 1:p){
     AIC.seq=c(AIC.seq, ## 空欄 (3) ##);
     BIC.seq=c(BIC.seq, ## 空欄 (4) ##)
}
plot(1:p, ylim=c(min(AIC.seq),max(BIC.seq)), type="n",
     xlab="# of variables", ylab="IC values")
lines(AIC.seq,col="red"); lines(BIC.seq,col="blue")
legend("topleft",legend=c("AIC","BIC"), col=c("red","blue"), lwd=1, cex =.8)
```

5 正則化

49. $N, p \geq 1$ として、$X \in \mathbb{R}^{N \times p}$, $y \in \mathbb{R}^N$, $\lambda \geq 0$ について、

$$\frac{1}{N}(y - X\beta)^2 + \lambda ||\beta||_2^2$$

を最小にする $\beta \in \mathbb{R}^p$ を求めたい。ただし、$\beta = (\beta_1, \cdots, \beta_p)$ について、$||\beta||_2 := \sqrt{\sum_{j=1}^p \beta_j^2}$ であるものとする。$N < p$ であるとき、そのような解が存在することと、$\lambda > 0$ であることとが同値であることを示せ。

50. 関数 $f : \mathbb{R} \to \mathbb{R}$ が凸で、$x = x_0$ で微分可能であるとき、任意の $x \in \mathbb{R}$ で $f(x) \geq f(x_0) + z(x - x_0)$ となるような z が存在し、$x = x_0$ における微分係数 $f'(x_0)$ と一致することを示せ。また、$p \geq 2$ として、$f : \mathbb{R}^p \to \mathbb{R}$ が凸で微分可能のとき、z はどのような値になるか。

51. (a) $-1 \leq z \leq 1$ であることと、すべての $x \in \mathbb{R}$ に対して $zx \leq |x|$ であることが同値であることを示せ。

    (b) 関数 $f(x) = |x|$ と各 $x_0 \in \mathbb{R}$ に対して、(a) で定義される z の集合を求めよ。ヒント: $x_0 > 0$, $x_0 < 0$, $x_0 = 0$ の場合にわけよ。

    (c) $f(x) = x^2 - 3x + |x|$, $f(x) = x^2 + x + 2|x|$ の各点での劣微分を計算して、それぞれの極大値、極小値をもとめよ。

52.
$$S_\lambda(x) := \begin{cases} x - \lambda, & x > \lambda \\ 0, & |x| \leq \lambda \\ x + \lambda, & x < -\lambda \end{cases}$$
で定義される $S_\lambda(x)$, $\lambda > 0$, $x \in \mathbb{R}$ を求める R 関数 `soft.th(lambda,x)` をかき、
`curve(soft.th(5,x),-10,10)` を実行せよ。ヒント: `max` ではなく、`pmax` を用いる。

53. $(x_i, y_i) \in \mathbb{R} \times \mathbb{R}$, $i = 1, \cdots, N$, $\lambda > 0$ から、
$$L = \frac{1}{2N} \sum_{i=1}^N (y_i - x_i\beta)^2 + \lambda|\beta|$$
を最小にする $\beta \in \mathbb{R}$ を求めたい。ただし、$x_1, \cdots, x_N$ は、$\sum_{i=1}^N x_i^2 = 1$ となるように正規化されているものとする。その解を $z := \frac{1}{N}\sum_{i=1}^N x_i y_i$ および関数 $\mathcal{S}_\lambda(\cdot)$ を用いて表わせ。

54. $p > 1$, $\lambda > 0$ に対して、係数 $\beta_0 \in \mathbb{R}$, $\beta \in \mathbb{R}^p$ を以下のように得るものとする。初期段階で $\beta \in \mathbb{R}^p$ は、ランダムに決める。次に、β_j を $r_{i,j} := y_i - \sum_{k \neq j} x_{i,j}\beta_j$ として、$\mathcal{S}_\lambda(\sum_{i=1}^N x_{i,j} r_{i,j}/N)$ に更新する。これを $j = 1, \cdots, p$ に対して行い、さらに収束するまでそのサイクルを繰り返す。下記の関数 `lasso` は、p 変数のサンプル分散を 1 におきかえ、(β_0, β) をえるものである。空欄をうめて、処理を実行せよ。

```
lasso=function(X, y, lambda=0){
    X=as.matrix(X); X=scale(X); p=ncol(X); n=length(y); X.bar=array(dim=p);
    for(j in 1:p){X.bar[j]=mean(X[,j]);X[,j]=X[,j]-X.bar[j];};
    y.bar=mean(y); y=y-y.bar
    eps=1; beta=array(0, dim=p); beta.old=array(0, dim=p)
    while(eps>0.001){
        for(j in 1:p){
            r= ## 空欄 (1) ##
            beta[j]= ## 空欄 (2) ##
        }
        eps=max(abs(beta-beta.old)); beta.old=beta
    }
    beta.0= ## 空欄 (3) ##
    return(list(beta=beta, beta.0=beta.0))
}
```

```
df=read.table("crime.txt"); x=df[,3:7]; y=df[,1]; p=ncol(x);
lambda.seq=seq(0,100,0.1); coef.seq=lambda.seq
plot(lambda.seq, coef.seq, xlim=c(0,100), ylim=c(-12,12),
     xlab="lambda",ylab="beta",main="各 lambda についての各係数の値", type="n", col="red")
for(j in 1:p){
     coef.seq=NULL; for(lambda in lambda.seq)coef.seq=c(coef.seq,## 空欄 (4) ##)
     par(new=TRUE); lines(lambda.seq,coef.seq, col=j)
}
```

55. 前問 (Lasso) を問題 49 のような定式化 (Ridge) として、処理を変更し、実行せよ。ヒント: 関数 `lasso` の eps の行と while ループを、以下でおきかえ、関数名も `ridge` とする。

```
beta=drop(solve(t(X)%*%X+n*lambda*diag(p))%*%t(X)%*%y)
```

グラフを描く部分、特に空欄 (4) は、lasso ではなく ridge とする。

56. 関数 `glmnet` および `cv.glmnet` の意味をしらべ、最適な λ とそのときの β を求めよ。5 個の変数のうちのどの変数が選択されるか。

```
library(glmnet)
cv.fit=cv.glmnet(X,y)
lambda.min=cv.fit$lambda.min
fit=glmnet(X,y,lambda=lambda.min)
```

ヒント: `fit$beta` で係数の値が表示される。非ゼロの値をもてば、その変数が選択されたことになる。

6 非線形

57. (a) データ $(x_1, y_1), \cdots, (x_N, y_N) \in \{-1, 1\} \times \mathbb{R}$ から、$u_i = (x_i + 1)/2$ として $\sum_{i=1}^N (y_i - \beta_0 - \beta_1 u_i)^2$ を最小にする (β_0, β_1) と、$v_i = (-x_i + 1)/2$ として $\sum_{i=1}^N (y_i - \gamma_0 - \gamma_1 v_i)^2$ を最小にする (γ_0, γ_1) を求めたい。$\beta_0, \beta_1, \gamma_0, \gamma_1$ の間にどのような関係があるか。

    (b) データ $(x_1, y_1), \cdots, (x_N, y_N) \in \{\pm 1, \pm 2\} \times \mathbb{R}$ から、$u_i = I[x_i > 0], w_i = I[|x_i| = 1]$ として $\sum_{i=1}^N (y_i - \beta_0 - \beta_1 u_i - \beta_2 w_i)^2$ を最小にする $(\beta_0, \beta_1, \beta_2)$、$v_i = I[x_i < 0], z_i = I[|x_i| = 2]$ として $\sum_{i=1}^N (y_i - \gamma_0 - \gamma v_i - \gamma z_i)^2$ を最小にする $(\gamma_0, \gamma_1, \gamma_2)$ を求めたい。$\beta_0, \beta_1, \beta_2, \gamma_0, \gamma_1, \gamma_2$ の間にどのような関係があるか。

58. データ $(x_1, y_1), \cdots, (x_N, y_N) \in \mathbb{R} \times \mathbb{R}$ から、以下の各値を最小にする $\beta_0, \beta_1, \cdots, \beta_p$ が一意的に求まる条件を求め、そのもとでの解を導出せよ。

    (a) $\sum_{i=1}^N (y_i - \sum_{j=0}^p \beta_j x_i^j)^2$

    (b) $\sum_{i=1}^N (y_i - \sum_{j=0}^p \beta_j f_j(x_i))^2$, $f_0(x) = 1$, $x \in \mathbb{R}$, $f_j : \mathbb{R} \to \mathbb{R}$, $j = 1, \cdots, p$

59. $K \geq 1$、$x_0 = -\infty, x_{K+1} = \infty$ として、

    (a) f, g を次数 p の多項式として、$f^{(j)}(x_*) = g^{(j)}(x_*)$, $x_* \in \mathbb{R}$, $j = 0, 1, \cdots, m$, $0 \leq m \leq p$ が成立するとき、

$$\begin{cases} f(x) = \sum_{j=0}^m \beta_j (x - x_*)^p \\ g(x) = \sum_{j=0}^m \gamma_j (x - x_*)^p \end{cases} \Longrightarrow \beta_j = \gamma_j \ , \ j = 0, 1, \cdots, m$$

    を示せ。

(b) f, g を 3 次の多項式として、 $f^{(j)}(x_*) = g^{(j)}(x_*)$, $j = 0, 1, 2$ であれば、$f(x) - g(x) = \gamma(x - x_*)^3$ となるような γ が存在することを示せ。

(c) $x_i \le x \le x_{i+1}$, $i = 0, 1, \cdots, K$ で定義され、$f_{i-1}^{(j)}(x_i) = f_i^{(j)}(x_i)$, $j = 0, 1, 2$, $i = 1, \cdots, K$ を満足する 3 次の多項式 $f_i(x)$ について、$f_i(x) = f_{i-1}(x) + \gamma_i(x - x_i)^3$ を満足する γ_i が存在することを示せ。

(d) f_i, $i = 0, 1, \cdots, K$ を (c) で定義された関数とする。$x_i \le x \le x_{i+1}$ について $f(x) = f_i(x)$, $i = 0, 1, \cdots, K$ となるような区分的に 3 次の多項式 f(3 次のスプライン関数) について、

$$f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \sum_{i=1}^{K} \beta_{i+3}(x - x_i)_+^3 \ ,$$

となるような $\beta_0, \beta_1, \cdots, \beta_{K+3}$ が存在することを示せ。ただし、$(x - x_i)_+$ は、$x > x_i$ で $x - x_i$、$x \le x_i$ で 0 の値をとる関数である。

60. 以下は、入力を $(x_1, y_1), \cdots, (x_N, y_N) \in \mathbb{R} \times \mathbb{R}$ をとして、3 次元のスプライン関数を出力する処理である。空欄をうめて、処理を実行せよ。

```
# データ生成
n=100; x=rnorm(n); y=sin(x)+0.2*rnorm(n); knots=c(-pi/2,0,pi/2)
## 係数の計算
K=length(knots); n=length(x); X=matrix(nrow=n,ncol=K+3)
for(i in 1:n){
X[i,1]= ##空欄 (1)##
X[i,2]= ##空欄 (2)##
X[i,3]= ##空欄 (3)##
for(j in 1:K)if(x[i]>knots[j]) X[i,j+3]= ##空欄 (4)## else X[i,j+3]= ##空欄 (5)## }
beta=as.numeric(lm(y~X)$coefficients)
# グラフを書く
par(mfrow=c(2,2))
f=function(x) as.numeric(beta[1]+beta[2]*x+beta[3]*x^2+beta[4]*x^3);
curve(f(x),-pi,knots[1])
g=function(x) f(x)+beta[5]*(x-knots[1])^3; curve(g(x),knots[1],knots[2])
h=function(x) g(x)+beta[6]*(x-knots[2])^3; curve(h(x),knots[2],knots[3])
k=function(x) h(x)+beta[7]*(x-knots[3])^3; curve(k(x),knots[3],pi)
```

61. $K \ge 2$ として、以下のような高々3 次のスプライン曲線を定義する (自然な 3 次のスプライン曲線): $x \le x_1$ および $x_K \le x$ で直線、$[x_i, x_{i+1}]$, $i = 1, \cdots, K-1$ のそれぞれで 3 次の多項式であって、K 個の境界 $x_1, \cdots, x_K$ で、関数 g とその 1 次、2 次微分が一致する。

(a) $2 * 2 + 4 * (K - 1) = 4K$ 変数の中に何個の制約式があるか。ヒント: K 個の境界のそれぞれで、3 個の制約がある。

(b) 関数 $g(x)$ は、$\gamma_1, \cdots, \gamma_K \in \mathbb{R}$, $h_1(x) = 1$, $h_2(x) = x$, $h_{j+2}(x) = d_j(x) - d_{K-1}(x)$, $j = 1, \cdots, K-2$ として、$\sum_{i=1}^{K} \gamma_i h_i(x)$ とかけることが知られている。ただし、

$$d_j(x) = \frac{(x - x_j)_+^3 - (x - x_K)_+^3}{x_K - x_j} \ ,$$

$j = 1, \cdots, K-1$ とおいている。このとき、$x_K \le x$ について、

$$h_{j+2}(x) = (x_{K-1} - x_j)(3x - x_j - x_{K-1} - x_K) \ ,$$

$j = 1, \cdots, K-2$ が成立することを示せ。

(c) 関数 $g(x)$ が、$x \leq x_1$ および $x_K \leq x$ で x の線形の関数になることを示せ。

62. 以下の空欄をうめて、n 個の点に適合するようなスプライン曲線を描け。

```
# データ生成
n=100; x=1:n; y=sin(2*pi*x/n)+0.5*rnorm(n)
# Set up knots
K=5; knots=c(10,30,50,70,90)
# 自然なスプライン関数を構成する
d= ##空欄 (1)##
h= ##空欄 (2)##
X=matrix(nrow=n, ncol=K); for(i in 1:n)for(j in 1:K) X[i,j]= ##空欄 (3)##
beta=drop(solve(t(X)%*%X)%*%t(X)%*%y)
g1=function(x){gg=0;for(j in 1:K)gg=gg+beta[j]*h(j,x); return(gg)}
# 通常のスプライン関数を構成する
X=matrix(nrow=n, ncol=K+4)
for(i in 1:n){
    for(j in 1:4) X[i,j]=x[i]^(j-1)
    for(j in 1:K) X[i,j+4]=(max(x[i]-knots[j],0))^3
}
gamma=drop(solve(t(X)%*%X)%*%t(X)%*%y)
g2=function(x){
    gg=0;
    for(j in 1:4)gg=gg+gamma[j]*x^(j-1)
    for(j in 1:K)gg=gg+gamma[j+4]*max(x-knots[j],0)^3
    return(gg)
}
# グラフを描く
plot(x,y)
x.seq=1:n; y.seq=NULL; for(z in x.seq)y.seq=c(y.seq,g1(z)); lines(x.seq,y.seq, col="red")
x.seq=1:n; y.seq=NULL; for(z in x.seq)y.seq=c(y.seq,g2(z)); lines(x.seq,y.seq, col="blue")
legend("bottomleft",legend=c("natural spline","spline"),col=c("red","blue"),lwd=1,cex =.8)
```

ヒント: 関数 h の定義は、$j = 1, j = 2, j > 2$ の場合分けが必要である。また、`h(j+2,x)` ではなく、`h(j,x)` の定義になっている点に注意する。

63. ISLR パッケージの Wage データセットを用いて、年齢から給与を予測したい。

```
library(ISLR)
attach(Wage)
plot(age, wage, col="gray")
```

(a) age の 25%, 50%, 75%のパーセンタイルを quantile 関数を用いて求めよ。

(b) 通常の 3 次元スプライン、自然な 3 次元スプライン曲線は、それぞれ $K+4$ 個、K 個のパラメータを用いて記述される。他方、それぞれの自由度 (df) は $K+3$、$K+1$ として定義される。また、R 言語では、df の値を指定しただけで、境界点が自動的に設定される。たとえば、境界が 3 個であれば、25%, 50%, 75%のパーセンタイルの値が用いられる。下記で最初にあらわれる bs 関数の境界点をかえて、2 個の曲線が重なるようにせよ。

```
library(splines)
```

```
agelims =range(age); age.grid=seq(from=agelims [1],to=agelims [2])
fit=lm(wage~bs(age, knots=c(25,40,60)),data=Wage)
pred=predict(fit, newdata=list(age=age.grid))
lines(age.grid, pred, lwd=2)
fit=lm(wage~bs(age, df=6),data=Wage)
pred=predict(fit, newdata=list(age=age.grid))
lines(age.grid, pred, lwd=2, col="blue")
```

(c) 関数 `bs` (b-spline) を `ns` (natural spline) に、自由度 `df=6` を `df=4` に、色 `col="blue"` を `col="red"` にかえて、同じグラフに通常の 3 次元スプライン、自然な 3 次元スプライン曲線を描け。

64. $(x_1, y_1), \cdots, (x_N, y_N) \in \mathbb{R} \times \mathbb{R}$ とする。任意の $\lambda \geq 0$ に対して、$x_1 < \cdots < x_N$ を境界点とする自然な 3 次のスプライン関数 g が、

$$RSS(f, \lambda) := \sum_{i=1}^{N} (y_i - f(x_i))^2 + \lambda \int_{-\infty}^{\infty} \{f''(t)\}^2 dt \tag{6}$$

を最小にする $f : \mathbb{R} \to \mathbb{R}$ であることを証明したい (平滑化スプライン関数)。

(a) 関数 $h(x) := f(x) - g(x)$ が

$$\int_{x_1}^{x_N} g''(x)h''(x)dx = 0 \tag{7}$$

を満足するとき、任意の $f : \mathbb{R} \to \mathbb{R}$ について、

$$\int_{-\infty}^{\infty} \{g''(x)\}^2 dx \leq \int_{-\infty}^{\infty} \{f''(x)\}^2 dx \ . \tag{8}$$

となることを示せ。ヒント: $x \leq x_1$ および $x_N \leq x$ では、$g(x)$ は 1 次式であって、$g''(x) = 0$ となる。また、(7) は以下を意味する。

$$\int_{x_1}^{x_N} \{g''(x) + h''(x)\}^2 dx = \int_{x_1}^{x_N} \{g''(x)\}^2 dx + \int_{x_1}^{x_N} \{h''(x)\}^2 dx \ .$$

(b) 以下を満足する $\gamma_1, \cdots, \gamma_{N-1} \in \mathbb{R}$ が存在することを示せ。

$$\int_{x_1}^{x_N} g''(x)h''(x)dx = -\sum_{i=1}^{N-1} \gamma_i \{h(x_{i+1}) - h(x_i)\}$$

ヒント: $g''(x_1) = g''(x_N) = 0$、および各 $x_i \leq x \leq x_{i+1}$ で 3 次微分係数は一定であることを用いる。

(c) (6) を最小にする $f : \mathbb{R} \to \mathbb{R}$ の集合の中に、自然な 3 次スプライン曲線 g が含まれていることを示せ。ヒント: $RSS(f, \lambda)$ が最小値であれとき、$g(x_i) = f(x_i)$, $i = 1, \cdots, N$ を満足する自然な 3 次スプライン g について、$RSS(g, \lambda) \leq RSS(f, \lambda)$ を示す。

65. $x_1 < \cdots < x_N$ を境界点にもつ平滑化スプライン関数 g について $g(x) = \sum_{j=1}^{N} g_j(x)\gamma_j$ および $g''(x) = \sum_{j=1}^{N} g_j''(x)\gamma_j$ なる $\gamma_1, \cdots, \gamma_N \in \mathbb{R}$ が存在することを仮定する。ただし、g_j, $j = 1, \cdots, N$ は高々3 次の多項式である。このとき、係数 $\gamma = [\gamma_1, \cdots, \gamma_N]^T \in \mathbb{R}^N$ が $G = (g_j(x_i)) \in \mathbb{R}^{N \times N}$ および $G'' = (\int_{-\infty}^{\infty} g_j''(x)g_k''(x)dx) \in \mathbb{R}^{N \times N}$ を用いて、$\gamma = (G^T G + \lambda G'')^{-1} G^T y$ とかけることを示せ。

66. 平滑化スプライン関数は、パラメータ λ で指定できるが、行列 $(G^T G + \lambda G'')^{-1} G^T$ のトレース (有効自由度) もよく用いられる。以下の処理は、クロスバリデーションで最適な有効自由度を求めている。そして、$df = 16$ と $df =$ 最適値について、平滑化スプライン曲線を描いている。

```
plot(age, wage, xlim=agelims, cex=.5, col="darkgrey")
title("平滑化スプライン")
fit=smooth.spline(age, wage, df=16)
fit2=smooth.spline(age, wage, cv=TRUE)
fit2$df
lines(fit,col="red",lwd =2); lines(fit2,col="blue",lwd=2)
legend("topright",legend=c("16 DF","最適な DF"), col=c("red","blue"),lty=1,lwd=2, cex =.8)
```

(a) 最適な λ の値を求めよ。ヒント: `names(fit2)` を実行して、その答えに必要な属性を見い出す。

(b) 同じグラフで、有効自由度 $df = 4, 16, 32$ の平滑化スプラインを異なる色で描け。

67. データ $(x_1, y_1), \cdots, (x_N, y_N) \in \mathbb{R}^p \times \mathbb{R}$ から

$$RSS := \{y_i - \beta_0 - \sum_{j=1}^{q} f_j(x_{i,j})\}^2 .$$

を最小にする $f_1 : \mathbb{R} \to \mathbb{R}, \cdots, f_q : \mathbb{R} \to \mathbb{R}$、を求めたい。各 f_j, $j = 1, \cdots, q$ は、多項式でも、通常のスプラインでも、自然な 3 次のスプラインでもよいが、平滑化スプラインがはいると、通常の最小 2 乗法が適用できない。以下では、GAM(generalized additive model) とよばれる方法を適用して、各関数を求める。(2 値) 分類の場合、データ $(x_1, y_1), \cdots, (x_N, y_N) \in \mathbb{R}^p \times \{-1, 1\}$ から、

$$\sum_{i=1}^{N} \log[1 + \exp\{-y_i(\beta_0 + \sum_{j=1}^{q} f_j(x_{i,j}))\}]$$

を最小にする $f_1 : \mathbb{R} \to \mathbb{R}, \cdots, f_q : \mathbb{R} \to \mathbb{R}$ を求めることになる。下記 (a)(b) それぞれで、どのような関数の和になっているか。ただし、`df=` オプションは、平滑化スプラインにおける有効自由度を指定したもの、また `se=TRUE` オプションは、観測値に $\pm$ 標準偏差の変動を加えたものである。

```
library (gam); library(ISLR); attach(Wage); par(mfrow=c(1,3))
```

(a)
```
gam.m1=gam(wage~s(year, df=4)+s(age ,df=5)+education);
plot(gam.m1, se=TRUE, col ="blue")
```

(b)
```
gam.m2=gam(I(wage>250)~year+s(age, df=5)+education, family=binomial,
     subset =(education !="1. < HS Grad"))
plot(gam.m2, se=TRUE, col="green")
```

(`table(education, I(wage>250))` を実行すると、"< HS Grad" で wage>250 のサンプルがないことがわかる。そのため、そうしたサンプルは除いている)

7 決定木

68. データセット `Carseats` から、何が売上に影響するかの決定木を構成し、予測性能を評価したい。空欄をうめ、訓練例 200 サンプルから得られた木を描け。

```
library(tree); library(ISLR); attach(Carseats)
High=ifelse(Sales<=8, "No","Yes")    # 売上が"No"または"Yes"に分類される
Carseats =data.frame(Carseats ,High)  # High という変数がデータフレーム Carseats に加わる
tree.carseats =tree(High~.-Sales, Carseats )
# Sales のデータを除いて、決定木を構成する
plot(tree.carseats); text(tree.carseats, pretty=0)
```

```
tree.carseats
train=sample(1: nrow(Carseats), 200)     # 訓練例として用いる 200 サンプルをランダムに選ぶ
Carseats.test=Carseats[-train,]
High.test=High[-train]                   # 正解
tree.carseats =tree(High~.-Sales, Carseats, subset=##空欄 (1)##)
tree.pred=predict(tree.carseats, Carseats.test, type="class")
     # 決定木を用いて得られた予測値
table(##空欄 (1)##, ##空欄 (2)##)
```

69. ある決定木で、各葉 $m = 1, \cdots, M (M \geq 1)$ がクラス $k = 1, \cdots, K$ のサンプルを $\alpha_{m.k}$ 個保持しているものとする (サンプルは全部で $N = \sum_{m=1}^{M} \sum_{k=1}^{K} \alpha_{m,k}$ ある)。この決定木の評価として、エントロピー

$$H := \sum_{m=1}^{M} \sum_{k=1}^{K} -\frac{\alpha_{m,k}}{N} \log \frac{\alpha_{m,k}}{\sum_{k'=1}^{K} \alpha_{m,k'}}$$

がよく用いられる。

(a) $H \geq 0$ を示せ。

(b) $\sum_{k=1}^{K} \alpha_{m,k} \geq 1$ のとき、$H = 0$ であることと、各 m で $\alpha_{m,k} = \sum_{k'=1}^{K} \alpha_{m,k'}$ なる k が存在することが必要十分であることを示せ。

70. 以下処理は、Boston データセットについて、クロスバリデーションによって、木の大きさ (葉の数) に関して最適な木を求めている。`names(cv.boston)` でどのような属性があるかを調べ、木の大きさ (size) を横軸、標準偏差 (deviation) を縦軸にしたグラフをかけ。

```
library(MASS)
train = sample(1:nrow(Boston), nrow(Boston)/2)
tree.boston=tree(medv~.,Boston,subset=train)
plot(tree.boston); text(tree.boston, pretty =0)
cv.boston=cv.tree(tree.boston)
```

ヒント: `plot` のオプション `type="b"`を用いると、点と点が直線で結ばれる。

71. 以下の処理は、ランダムフォーレストの回帰を用いて、他の変数からある変数の値の予測を行うものである。(yhat.bag, boston.test) の対をプロットし、サンプルサイズが十分大きい場合の理論的な直線を引け。ヒント: `plot` および `abline(0,1)` を用いて、点と直線を引け。

```
library(MASS); library(randomForest)
train=sample(1:nrow(Boston), nrow(Boston)/2)
boston.test=Boston[-train,"medv"]
bag.boston=randomForest(medv~.,data=Boston, subset=train, mtry=13)
yhat.bag=predict(bag.boston, newdata=Boston[-train,])
```

72. ランダムフォーレストの性能は、森を生成するための変数の個数 m に依存する。Boston データセットで、m の値を 2 からまで 13 変化させ、12 種類の各 m について、予測エラー `mean((yhat.bag -boston.test)^2)` を求める。そして、その関係をグラフで描け (プロットの間は直線で結べ)。

73. 以下の処理では、Carseats データセットについて、$m = \sqrt{p}$, $m = p/2$, $m = p (p$ は変数の個数) のそれぞれについてサンプル数 n とともに、性能がどのように変化していくかを見ている。空欄をうめて、グラフをえがけ。

```
library(randomForest)
library(ISLR)
data(Carseats)
nn=c(seq(1,9,1),seq(10,50,10))
plot(nn,nn/MAX*0.5, type="n", xlim=c(0,60),ylim=c(0,0.15),xlab="生成した木の数",ylab="
テストデータの誤り率")
m=c(3,5,10); color=c("blue","green","red")
attach(Carseats)
High=ifelse(Sales<=8, "No","Yes")
Carseats =data.frame(Carseats ,High)
train=sample(1: nrow(Carseats), 200)    # 訓練例として用いる 200 サンプルをランダムに選
ぶ
Carseats.test=Carseats[-train,"High"]
for(i in 1:3){
     x=NULL; y=NULL
     for(n in nn){
          bag.carseats=randomForest(High~., ntree=n,
               mtry=##空欄 (1)##, data=Carseats, subset=train)
          yhat.bag=predict(bag.carseats, newdata=Carseats[-train,], type="class")
          tab=table(yhat.bag,carseats.test)
          x=c(x,n); y=##空欄 (2)##)
     }
lines(x,y, col=color[i])
}
legend("topright",legend=c("m=p^0.5","m=p/2","m=p"), col=color, lwd=2, cex =.8)
```

74. 以下の処理は、Boston データセットについて、d=1, d=2, d=3 の 3 個のオプションについて、性能を比較している。空欄をうめて、グラフをえがけ。

```
library(MASS); library(gbm)
train=sample(1:nrow(Boston), nrow(Boston)/2)
MAX=3000; nn=c(seq(1,9,1),seq(10,90,10),seq(100,MAX,50))
plot(nn,nn/MAX*80, type="n", xlab="生成した木の個数", ylab="テストデータに対しての 2 乗
誤差")
d=1:3; color=c("blue","green","red")
for(i in 1:3){
     x=NULL; y=NULL
     for(n in nn){
          boost.boston=gbm(medv~., data=Boston[train,], distribution="gaussian",
               n.trees=n, interaction.depth= ##空欄 (1)## )
          yhat.boost=predict(boost.boston, n.trees=n, newdata= ##空欄 (2)##)
          S=mean((yhat.boost-boston.test)^2)
          x=c(x,n); y=c(y,S)
     }
lines(x,y, col=color[i])
}
legend("topright",legend=c("d=1","d=2","d=3"), col=color, lwd=2, cex =.8)
```

8 サポートベクトルマシン

75. $\beta_0 \in \mathbb{R}$、および $||\beta||_2 = 1$ なる $\beta \in \mathbb{R}^p$ について、関数 $f(z) := \beta_0 + z^T\beta$, $z \in \mathbb{R}^p$ を定義する。このとき、

(a) $(u, v) \in \mathbb{R}^2$ と直線 $aU + bV + c = 0$, $a, b \in \mathbb{R}$ の距離を、

$$\frac{|au + bv + c|}{\sqrt{a^2 + b^2}}$$

であると定義する。$(x_1, y_1), \cdots, (x_N, y_N) \in \mathbb{R}^p \times \{-1, 1\}$ について、$y_1 f(x_1), \cdots, y_N f(x_N) \geq 0$ が成立すると仮定する。各点 x_i と直線 $\beta_0 + X^T\beta = 0$ の距離の最小値 $M \geq 0$ を最大にする (β_0, β) を求めたい。この問題を定式化せよ。

(b) $\gamma \geq 0$, $M \geq 0$ とする。$\sum_{i=1}^{N} \epsilon_i \leq \gamma$ および

$$y_i(\beta_0 + \beta^T x_i) \geq M(1 - \epsilon_i) \ , \ i = 1, \cdots, N$$

を満足する $(\beta_0, \beta) \in \mathbb{R}^p \times \mathbb{R}$ および $\epsilon_i \geq 0, i = 1, \cdots, N$ を動かして、M を最大にしたい。$\gamma = 0$ であれば、この問題が (a) に帰着されることを示せ。

76. $f_j(\beta) \leq 0$, $j = 1, \cdots, m$ のもとで、$f_0(\beta)$ を最小にする $\beta \in \mathbb{R}^p$ を求めたい。そのような解が存在するとき、その最小値が f^* であるとする。$\alpha = (\alpha_1, \cdots, \alpha_m) \in \mathbb{R}^m$ に対して、

$$L(\alpha, \beta) := f_0(\beta) + \sum_{j=1}^{m} \alpha_j f_j(\beta)$$

であるとき、以下の 2 式を示せ。

$$\sup_{\alpha \geq 0} L(\alpha, \beta) = \begin{cases} f_0(\beta), & f_j(\beta) \leq 0 \ , \ j = 1, \cdots, m \\ +\infty & \text{otherwise} \end{cases} \tag{9}$$

$$f^* := \inf_{\beta} \sup_{\alpha \geq 0} L(\alpha, \beta) \geq \sup_{\alpha \geq 0} \inf_{\beta} L(\alpha, \beta) \tag{10}$$

77. $f_0, f_1, \cdots, f_m : \mathbb{R}^p \to \mathbb{R}$ を凸で、$\beta = \beta^*$ で微分可能であるとする。$\beta^* \in \mathbb{R}^p$ が $\min\{f_0(\beta) | f_i(\beta) \leq 0, i = 1, \cdots, m\}$ の最適解であることと、

$$f_i(\beta^*) \leq 0 \ , \ i = 1, \cdots, m \tag{11}$$

であって、以下の 2 条件を満足する $\alpha_i \geq 0, i = 1, \cdots, m$ が存在することが同値であることが知られている (KKT 条件)。

$$\alpha_i f_i(\beta^*) = 0 \ , \ i = 1, \cdots, m \tag{12}$$

$$\nabla f_0(\beta^*) + \sum_{i=1}^{m} \alpha_i \nabla f_i(\beta^*) = 0 \ . \tag{13}$$

本問題では、十分性のみを示す。

(a) $f : \mathbb{R}^p \to \mathbb{R}$ が凸であって、$x = x_0 \in \mathbb{R}$ において微分可能であるとき、各 $x \in \mathbb{R}^p$ について、

$$f(x) \geq f(x_0) + \nabla f(x_0)^T (x - x_0) \tag{14}$$

となることを示せ。ヒント: すでに $p = 1$ の場合は、証明している。

(b) (11) を満足する任意の $\beta \in \mathbb{R}^p$ について、$f_0(\beta^*) \leq f_0(\beta)$ となることを示せ。ヒント: (11)(12)(13) を各 1 回、(14) を 2 回使う。

78. 問題 75 の条件 $||\beta||_2 = 1$ をはずし、$\beta_0/M, \beta/M$ を β_0, β とみなすとき、

$$L_P := \frac{1}{2}||\beta||_2^2 + C\sum_{i=1}^{N}\epsilon_i - \sum_{i=1}^{N}\alpha_i\{y_i(\beta_0 + \beta^T x_i) - (1-\epsilon_i)\} - \sum_{i=1}^{N}\mu_i\epsilon_i \tag{15}$$

を最小にする $\beta_0, \beta, \epsilon_i$, $i = 1, \cdots, N$ を見出す問題に帰着される。ただし、$C > 0$(コスト) であるとし、最後の 2 項が制約、$\alpha_i, \mu_i \geq 0$, $i = 1, \cdots, N$ がラグランジェ係数となる。(10) の $L_P = \sup_{\alpha \geq 0} L(\alpha, \beta)$ を最小にする (主問題) のではなく、$L_D := \inf_\beta L(\alpha, \beta)$ を最大にする (双対問題) を考える。ここで、L_P を β_0, β, ϵ_i で偏微分して、

$$\sum_{i=1}^{N}\alpha_i y_i = 0 \tag{16}$$

$$\beta = \sum_{i=1}^{N}\alpha_i y_i x_i \in \mathbb{R}^p \tag{17}$$

$$C - \alpha_i - \mu_i = 0 \tag{18}$$

が得られる。(16)(17)(18) から、双対問題が以下で与えられることを示せ。

$$L_D := \sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j x_i^T x_j \ . \tag{19}$$

79. (16)(17)(18) および

$$\alpha_i[y_i(\beta_0 + \beta^T x_i) - (1-\epsilon_i)] = 0 \tag{20}$$

$$\mu_i\epsilon_i = 0 \tag{21}$$

$$y_i(\beta_0 + \beta^T x_i) - (1-\epsilon_i) \geq 0 \tag{22}$$

$$\epsilon_i \geq 0 \tag{23}$$

が、$\beta_0, \beta, \epsilon_i$, $i = 1, \cdots, N$ が (15) を最小にするための必要十分条件となる (KKT 条件)。それらから、以下を示せ。

(a) $\alpha_i > 0 \Longrightarrow y_i(\beta_0 + \beta^T x_i) \leq 1$

(b) $\alpha_i = 0 \Longrightarrow y_i(\beta_0 + \beta^T x_i) \geq 1$.

(c) $0 < \alpha_i < C \Longrightarrow y_i(\beta_0 + \beta^T x_i) = 1 \Longrightarrow \beta_0 = y_i - \beta^T x_i$.

80. (18) より、 以下の不等式を得る。

$$0 \leq \alpha_i \leq C \ . \tag{24}$$

したがって、制約 (16)(24) のもとで、α に関して (19) を最大にする問題に帰着される。その問題を、2 次計画法の解ソルバーにわたすために、

$$L_D = -\frac{1}{2}\alpha^T D_{mat}\alpha + d_{vec}^T\alpha$$

$$A_{mat}^T\alpha \geq b_{vec} \ ,$$

となるような $D_{mat} \in \mathbb{R}^{N\times N}$, $A_{mat} \in \mathbb{R}^{m\times N}$, $d_{vec}, b_{vec} \in \mathbb{R}^N$ $(m \geq 1)$ を指定する必要がある。ただし、制約式 $A_{mat}^T\alpha \geq b_{vec}$ の中の最初の meq 個は (不等式ではなく) 等式であるものとする。

$$b_{vec} = [0, -C, \cdots, -C, 0, \cdots, 0]^T \ .$$

とするとき、D_{mat}, A_{mat}, d_{vec}, meq はそれぞれ何になるか。

81. 前問にもとづいて、$p = 2$ の場合に最適な直線を求めてみた。特に、`quadprog` パッケージという 2 次計画法のソルバーを使ってみた。 空欄を埋めて、実行せよ。

```
library(quadprog)
# 関数の定義
svm.1=function(X,y,C){
     eps=0.0001; n=nrow(X)
     Dmat=matrix(nrow=n,ncol=n)
     for(i in 1:n)for(j in 1:n)Dmat[i,j]=sum(X[i,]*X[j,])*y[i]*y[j]
     Dmat=Dmat+eps*diag(n)
     dvec= ##空欄 (1)##
     Amat=matrix(nrow=(2*n+1),ncol=n)
     Amat[1,]=y; Amat[2:(n+1),1:n]=-diag(n);
     Amat[(n+2):(2*n+1),1:n]= ##空欄 (2)##
     Amat=t(Amat)
     bvec= ##空欄 (3)##
     meq=1
     alpha=solve.QP(Dmat,dvec,Amat,bvec=bvec,meq=1)$solution
     index=(1:n)[eps<alpha&alpha<C-eps]
     beta=drop((alpha*y)%*%X); beta.0=y[index]-X[index,]%*%beta
     return(list(beta=beta,beta.0=beta.0[1]))
}
# svm.1 を用いた実行
a=rnorm(1); b=rnorm(1)
n=100; X=matrix(rnorm(n*2),ncol=2,nrow=n); y=sign(a*X[,1]+b*X[,2]+0.3*rnorm(n))
plot(-3:3,-3:3,xlab="X[,1]",ylab="X[,2]", type="n")
for(i in 1:n){if(y[i]==1)points(X[i,1],X[i,2],col="red")
     else points(X[i,1],X[i,2],col="blue")}
qq=svm.1(X,y,10)
abline(-qq$beta.0/qq$beta[2],-qq$beta[1]/qq$beta[2])
```

82. V をベクトル空間として、$(x,y) \in \mathbb{R}^p \times \mathbb{R}^p$ の ($h : \mathbb{R}^p \to V$ によって誘発される) カーネル $K(x,y)$ を、$h(x)$ と $h(y)$ の間の内積として定義する。たとえば、d 次元の多項式カーネル $K(x,y) = (1 + x^T y)^d$ については、$d = 1$, $p = 2$ であれば、

$$((x_1, x_2), (y_1, y_2)) \mapsto 1 \cdot 1 + x_1 y_1 + x_2 y_2 = (1, x_1, x_2)^T (1, y_1, y_2) \ ,$$

となる。この場合、h が $(x_1, x_2) \mapsto (1, x_1, x_2)$ となる。$p = 2, d = 2$ の場合、h は何になるか。

83. V を $\mathbb{R}$ 上のベクトル空間とする。

(a) V が $[0,1]$ における連続関数の集合であるとすると、$\int_0^1 f(x)g(x)dx$, $f, g \in V$ が V の内積となることを示せ。

(b) ベクトル空間 $V := \mathbb{R}^p$ について、$(1 + x^T y)^2$, $x, y \in \mathbb{R}^p$ が V の内積ではないことを示せ。

ヒント: 内積の定義を確認せよ:$a, b, c \in V$, $\alpha \in \mathbb{R}$ について、$< a + b, c >=< a, c > + < b, c >$; $< a, b >=< b, a >$; $< \alpha a, b >= \alpha < a, b >$; $< a, a >= ||a||^2 \geq 0$。

84. 通常の内積を実現する R 関数 `K.linear(x,y)`、および $d = 2$ の多項式カーネルを実現する R 関数 `@K.poly(x,y)` をかけ。

85. 以下では、ある $h : \mathbb{R}^p \to V$ を用いて、$x_i \in \mathbb{R}^p$, $i = 1, \cdots, N$ をすべて $h(x_i) \in V$ におきかえるものとする。したがって、$\beta \in \mathbb{R}^p$ は $\beta = \sum_{i=1}^N \alpha_i y_i h(x_i) \in V$ となり、L_D の定義の中の内積 $x_i^T x_j$ は $h(x_i)$ と $h(x_j)$ の間の内積、すなわちカーネル $K(x_i, x_j)$ となる。このように拡張すると、境界線

$\beta^T h(X) + \beta_0 = 0$、すなわち $\sum_{i=1}^{N} \alpha_i y_i K(X, x_i) + \beta_0 = 0$ が、必ずしも直線 $\beta^T X + \beta_0 = 0$ ではなくなる。問題 81 の `svm.1` を以下のように変更せよ。

(a) 引数 `K` を関数の定義におき、

(b) `sum(X[,i]*X[,j])` を `K(X[i,],X[j,])` におきかえ、

(c) `return()` における `beta` を `alpha` におきかえる。

そして、関数名を `svm.2` として、以下の空欄を埋めて処理を実行せよ。

```
# 関数の定義
plot.kernel=function(K, lty){
qq=svm.2(X,y,1,K); alpha=qq$alpha; beta.0=qq$beta.0
f=function(u,v){x=c(u,v);S=beta.0; for(i in 1:n)S=S+ ## 空欄 ##; return(S)}
u=seq(-2,2,.1);v=seq(-2,2,.1);w=array(dim=c(41,41))
for(i in 1:41)for(j in 1:41)w[i,j]=f(u[i],v[j])
contour(u,v,w,level=0,add=TRUE,lty=lty)
}
# 実行
a=rnorm(1); b=rnorm(1)
n=100; X=matrix(rnorm(n*2),ncol=2,nrow=n); y=sign(a*X[,1]+b*X[,2]+0.3*rnorm(n))
plot(-3:3,-3:3,xlab="X[,1]",ylab="X[,2]", type="n")
for(i in 1:n){
     if(y[i]==1)points(X[i,1],X[i,2],col="red")
     else points(X[i,1],X[i,2],col="blue")
     }
plot.kernel(K.linear,1)
plot.kernel(K.poly,2)
```

86. 以下の手続きは、人工データに対して、$\gamma = 1$ のラジカルカーネルとコスト $C = 1$ のサポートベクトルマシンを実行させたものである。

```
library(e1071)
x=matrix(rnorm(200*2), ncol=2); x[1:100,]=x[1:100,]+2; x[101:150,]=x[101:150,]-2
y=c(rep(1,150), rep(2,50)); dat=data.frame(x=x,y=as.factor(y))
train=sample(200,100)
svmfit=svm(y~., data=dat[train,], kernel ="radial", gamma=1, cost=1)
plot(svmfit, dat[train,])
```

(a) $\gamma = 1, C = 100$ に対して、サポートベクトルマシンを実行せよ。

(b) `tune` コマンドを用いて、クロスバリデーションによって、最適な C および γ を $C = 0.1, 1, 10, 100, 1000$ および $\gamma = 0.5, 1, 2, 3, 4$ の中から選べ。

```
tune.out=tune(svm, y~., data=dat[train ,], kernel="radial",
     ranges=list(cost=c(0.1,1,10,100,1000), gamma=c(0.5,1,2,3,4)))
summary (tune.out)
```

87. サポートベクトルマシンは、`e1071` などのパッケージでは、クラスが 2 個以上の場合でも、クラス数を指定しなくても、実行可能である。空欄を埋めて、処理を実行せよ。

```
library(e1071)
df=read.table("iris.txt",sep=","); x=df[,1:4]; y=as.factor(df[,5])
m=30; train=sample(1:150,150-m); x.train=x[train,]; y.train=y[train]
dat=data.frame(##空欄 (1)##, y.train)
svmfit=svm(##空欄 (2)##~., data=dat,kernel="radial",cost=10,gamma=1)
x.test=x[-train,]; y.test=y[-train]; table(y.test,predict(svmfit,x.test))
```

9 教師なし学習

88. 以下の処理は、K-means クラスタリングという方法で、p 変数の値をもった N を交わりのない K 個の集合 (クラスタ) に分割している。最初に $1, \cdots, K$ のいずれかを N データのそれぞれに割当て、以下の 2 ステップを繰り返す。

    (a) クラスタ $k = 1, \cdots, K$ のそれぞれで、中心 (平均ベクトル) を求める。

    (b) N データのそれぞれに、K クラスタの中で最も近いクラスタを割り当てる

    空欄を埋めて、処理を実行せよ

```
# データ生成
K=10; p=2; n=1000; X=matrix(rnorm(p*n),nrow=n,ncol=p)
# K-means クラスタリング
y=sample(1:K, n, replace=TRUE); center=array(dim=c(K,p))
for(h in 1:10){
for(k in 1:K){
   if(sum(y[]==k)==0)center[k,]=Inf else for(j in 1:p)center[k,j]= ## 空欄 (1) ##)
}
S.total=0
for(i in 1:n){
   S.min=Inf;
    for(k in 1:K){
          S=sum((X[i,]-center[k,])^2); if(S<S.min){S.min=S; ## 空欄 (2)## }
   }
   S.total=S.total+S.min
}
print(S.total)
}

# クラスタごとに色を変えて、点を描く
plot(-3:3, -3:3, xlab="x", ylab="y", type="n"); points(X[,1],X[,2],col=y+1)
```

89. K-means クラスタリングの結果は、ランダムに選ばえた初期値に大きく依存する。現在の処理の外側にループを設定し (10 回程度まわす)、異なる初期値で実行した結果の中の最良のものを選択せよ。前問の `print(S.total)` の位置を変えて、実行のスコアの推移を見よ。

90. K-means クラスタリングは、 データ $X = (x_{i,j})$ から、クラスタ $C_1, \cdots, C_K$ をかえて、

$$S := \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{i,j} - x_{i',j})^2$$

    を最小化している。

(a) 以下の等式を示せ。

$$\frac{1}{|C_k|}\sum_{i,i'\in C_k}\sum_{j=1}^{p}(x_{i,j}-x_{i',j})^2 = 2\sum_{i\in C_k}\sum_{j=1}^{p}(x_{i,j}-\bar{x}_{k,j})^2$$

(b) 問題 88 の 2 ステップを毎回実行するごとに、スコア S が単調に減少することを示せ。

91. 以下の処理は、関数 kmeans を用いて、K-means クラスタリングを行っている。その関数の意味を調べて、空欄をうめて、処理を実行せよ。

```
n=100; K=5; x=matrix(rnorm(n*2),ncol=2); y= ## 空欄 ##
plot(x,col=y, xlab="", ylab="", main=paste("K=",K), pch=20,cex=2)
```

92. 以下の処理は、データ $x_1,\cdots,x_N \in \mathbb{R}^p$ について、階層的クラスタリングを行うものである。最初は、各クラスタが 1 データのみからなっていて、クラスタをマージしていながら、任意のクラスタ数 K でのクラスタリングを行うものである。要素 (データ) 間の距離と、クラスタ (一般には複数の要素を含む) 間の距離を指定する必要がある。空欄をうめて、処理を実行せよ。

```
# クラスタ間の距離
dist.1=function(x,y){
dist.max=0
for(u in x)for(v in y){
u=unlist(u);v=unlist(v);dist=dist.2(u,v);
if(sum(dist)>dist.max)dist.max=dist
}
return(dist.max)
}
# 要素の間の距離
dist.2=function(x,y)sum((x-y)^2)
# 階層的クラスタリング
h.cluster=function(x,K){ # K はクラスタの個数
for(k in n: ##空欄 (1)##){
dist.min=Inf
for(i in 1:(k-1))for(j in (i+1):k){
dist= ##空欄 (2)##
if(dist<dist.min){dist.min=dist; pair.1=i; pair.2=j}
}
i=pair.1; j=pair.2;
x[[i]]= append(x[[i]],x[[j]]);
if(j<k)for(h in (j+1):k)x[[h-1]]= ##空欄 (3)##
x[[k]]=NULL
}
return(x)
}
# データ生成 N=100 および p=2 を仮定
n=100; x=list(dim=n); for(i in 1:n)x[[i]]=list(list(rnorm(1),rnorm(1)))
# 色でクラスタリングをあらわす
par(mfrow=c(3,2))
for(K in 2:7){
     y=h.cluster(x,K)
```

```
    plot(-5:5,-5:5,type="n", xlab="",ylab="",main=paste("K=",K))
    for(k in 1:K){
        w=y[[k]];m=length(w)
        for(h in 1:m){
            u=w[[h]][[1]]; v=w[[h]][[2]];
            points(u,v, col=k+1, pch=20,cex=2)
        }
    }
}
```

93. 前問の関数 `dist.1` は、2 クラスタ間の対となるデータ間の距離の最大値でクラスタ間の距離を定義している (Complete Linkage)。その関数を、2 クラスタのそれぞれの中心の間の距離でクラスタ間の距離を定義する方法 (Centroid Linkage) に置き換えて実行せよ。

94. 関数 `hcluster` では、Complete と Centroid 以外に、Average と Single が用意されている。それぞれの意味を調べ、空欄をうめて、実行せよ。ただし、関数 `dist` は、要素間の距離が下三角になる行列を求めている。このようにして得られたグラフをデンドログラム (dendrogram) と呼ばれる。

```
x=matrix(rnorm(n*2),ncol=2); par(mfrow=c(2,2))
hc.complete=hclust(dist(x),method="complete");plot(hc.complete)
hc.centroid=hclust(dist(x),method="centroid");plot(hc.centroid)
hc.average=hclust(dist(x),method="##空欄 (1)##");plot(hc.average)
hc.single=hclust(dist(x),method="##空欄 (2)##");plot(hc.single)
```

95. 行列 $X \in \mathbb{R}^{N\times p}$ について、

    (a) $||\phi_1|| = 1$ であって、$||X\phi_1||^2$ を最大にするベクトル $\phi_1 \in \mathbb{R}^p$ が、ある $\lambda \geq 0$ について以下を満足することを示せ。ただし、$\Sigma := \dfrac{X^TX}{N}$ である。

    $$\Sigma\phi_1 = \lambda\phi_1 \tag{25}$$

    ヒント: 次式を ϕ_1 で微分せよ。

    $$L := ||X\phi_1||^2 - \lambda(||\phi_1||^2 - 1)$$

    (b) 各 $i = 1, \cdots, p$ について、$\phi_j^T\phi_i = \delta_{i,j} := \begin{cases} 1, & j = i \\ 0, & j = 1, \cdots, i-1 \end{cases}$ であって、$||X\phi_i||^2$ を最大にする $\phi_i \in \mathbb{R}^p$ が、ある $\lambda_1, \cdots, \lambda_p$ について、以下を満足することを示したい。

    $$\Sigma\phi_i = \lambda\phi_i \ . \tag{26}$$

    $L := ||X\phi_i||^2 - \mu_i(||\phi_i||^2 - 1)$ について、$\dfrac{\partial L}{\partial \phi_i} = 0, i = 1, \cdots, p$ であれば、ある $\lambda_1, \cdots, \lambda_p \geq 0$ について、$\Sigma\phi_i = \lambda_i\phi_i$ を満足する (主成分分析) ことを示せ。

96. R 言語の `eigen` 関数を用いて、行列 $X \in \mathbb{R}^{N\times p}$ を入力とし、`pca(X)$values` で固有値 $\lambda_1, \cdots, \lambda_p$ を要素にもつベクトル、`pca(X)$vectors` で $\phi_1, \cdots, \phi_p$ を各列にもつ行列を出力する関数 `pca` を R 言語でかけ。

97. 以下の処理は、N 個の点 $(x_1, y_1), \cdots, (x_N, y_N)$ から、主成分の方向ベクトル ϕ_1 および ϕ_2 を求めている。空欄をうめて処理を実行せよ。

```
#データ生成
n=100; a=0.7; b=sqrt(1-a^2); u=rnorm(n); v=rnorm(n); x=u; y=u*a+v*b;plot(x,y)
#固有ベクトルを求める
x=x-mean(x); y=y-mean(y)
X2=sum(x^2);Y2=sum(y^2); XY=sum(x*y);S=matrix(c(X2,XY,XY,Y2),ncol=2)/n;
T=eigen(S)$vectors
#直交する 2 直線をひく
abline(##空欄 (1)##, ##空欄 (2)##,col="red");
abline(##空欄 (3)##, ##空欄 (4)##,col="blue")
```

さらに 2 直線の傾きの積が -1 であることを確認せよ。

98. $0 \leq K \leq p$ とし、$\Phi \in \mathbb{R}^{p \times K}$ を、大きさが 1 で相互に直交するベクトルを各列にもつ行列とする。x_i を行列 $X \in \mathbb{R}^{N \times p}$ の第 i 行 (行ベクトル) として、Φ による射影 $\mathbb{R}^p \ni x_i \mapsto x_i \Phi \Phi^T \in \mathbb{R}^p$, $i = 1, \cdots, N$ をスコア $\sum_{i=1}^N ||x_i - x_i \Phi \Phi^T||^2$ に対応付けることを考える。

    (a) ϕ_j を Φ の第 j 列として、$\sum_{i=1}^N ||x_i - x_i \Phi \Phi^T||^2 = \sum_{i=1}^N ||x_i||^2 - \sum_{i=1}^N ||x_i \Phi||^2$ および $\sum_{i=1}^N ||x_i \Phi||^2 = \sum_{j=1}^K ||X \phi_j||^2$ を示せ。

    (b) $\lambda_1, \cdots, \lambda_K$ を $\Sigma = X^T X / N$ の最大の K 固有値として、$\sum_{i=1}^N ||x_i - x_i \Phi \Phi^T||^2$ の値は、$\phi_1, \cdots, \phi_K$ として、その対応する K 固有ベクトルをとることによって、最小値 $\sum_{i=1}^N ||x_i||^2 - \sum_{j=1}^K \lambda_j$ をとることを示せ。

99. 以下の処理は、4 種類の犯罪について 50 州での逮捕者のデータセットである。

```
library(ISLR); pr.out=prcomp(USArrests,scale=TRUE)
```

    (a) `names(pr.out)` によって属性を調べ、 それらの結果を表示せよ。

    (b) `biplot(pr.out)` を実行し、`pr.out$x` および `pr.out$rotation` の値に -1 をかけて、再度 `biplot(pr.out)` を実行せよ。

100. 各 $1 \leq K \leq p$ について、寄与率および累積寄与率は、それぞれ $\lambda_k / \sum_{j=1}^L \lambda_j$ および $\sum_{k=1}^K \lambda_k / \sum_{j=1}^L \lambda_j$ で定義される。空欄をうめて、処理を実行せよ。

```
pr.var=##空欄 (1)##
pve=##空欄 (2)##
par(mfrow=c(1,2))
plot(pve, xlab="主成分",
     ylab="寄与率", ylim=c(0,1) ,type="b")
plot(cumsum(pve), xlab="主成分",
     ylab="累積の寄与率", ylim=c(0,1), type="b")
```