第2回演習課題

統計解析 I (鈴木)

出題: 2018 年 4 月 9 日
期限: 2018 年 5 月 21 日

25. $0 < \alpha < 1$, $\beta, \gamma \in \mathbb{R}^p$ について、

$$||y - X(\alpha\beta + (1-\alpha)\gamma)||_2^2 \le \alpha||y - X\beta||_2^2 + (1-\alpha)||y - X\gamma)||_2^2$$

の等号が成立する条件は何か。次に、ある $\lambda \ge 0$ における Lasso で、係数 $\hat{\beta}, \hat{\gamma}$ が共通の最小値 c^* をもつとき、$0 < \alpha < 1$ について、

$$||y - X(\alpha\hat{\beta} + (1-\alpha)\hat{\gamma})||_2^2 + ||\alpha\hat{\beta} + (1-\alpha)\hat{\gamma}||_1 \le c^*$$

となることを示せ。また、$X\hat{\beta} = X\hat{\gamma}$ となることを示せ。さらに、$\lambda > 0$ のとき、$||\hat{\beta}||_1 = ||\hat{\gamma}||_1$ となることを示せ。

26. 問題 19[1] の関数 linear.lasso は、glmnet と同じオプション `standardize=TRUE`, `standardize=FALSE` を設けている。`standardize=TRUE` では、全体の処理の前に、X の各列 (第 j 列) をその大きさ s_j で割って Lasso の処理を行い、座標降下法が収束したら、最後に β_j に s_j をかけている。`standardize=FALSE` では、そのような処理を行わない。以下の 2 条件のそれぞれで、両者は一致することを確認し、その理由を述べよ。

   (a) $\lambda = 0$ のとき

   (b) X の各列をその大きさで割って (正規化して) から、`linear.lasso` を実行したとき

27. ridge にも、`linear.lasso` や `glmnet` と同様の `standardize=TRUE`、`standardize=FALSE` のオプションをつけよ。ヒント: 問題 19 の関数 linear.lasso の場合と同様に、係数を推定する直前に X の各列をその大きさで割って、推定したら各係数をその大きさでわるという 2 行 (`if(standardize==TRUE` ではじまる) を、ridge の場合にもおく。

28. 座標降下法で、最初に λ の値を大きくしてすべての係数を 0 にし、λ の値を徐々に小さくしていく。そして、毎回、直前の λ での値を初期値として、次の λ の値を計算することを考える (warm start)。下記の (1)(2) をうめよ。また、出力を pdf で提出せよ。

```
warm.start=function(X,y,lambda.max=100,standardize=TRUE){
p=ncol(X); n=nrow(X);
X=as.matrix(X); for(j in 1:p)X[,j]=X[,j]-mean(X[,j]);
y=as.vector(y); y=y-mean(y);
if(standardize==TRUE){
scale=array(dim=p);
for(j in 1:p){scale[j]=sqrt(covar(X[,j],X[,j]));X[,j]=X[,j]/scale[j];}
}
dec=round(lambda.max/50);
```

[1] 第 1 回は、何度か更新されている。最新版をダウンロードせよ。また、問題は通し番号とする。

```
lambda.seq=seq(lambda.max,1,-dec);
r=length(lambda.seq);
coef.seq=matrix(nrow=r,ncol=p);
## X[,3],...,X[,7] のそれぞれについて、係数の列を作る。最初は NULL
beta=array(0, dim=p);
k=0;
for(lambda in lambda.seq){
k=k+1;
beta.old=  ## (1) ここにいれる
eps=1;
while(eps>0.001){
for(j in 1:p){
r= y- X[,-j] %*% beta[-j]
beta[j]=soft.th(lambda,covar(r,X[,j]))/covar(X[,j],X[,j])
}
eps=max(abs(beta-beta.old));
beta.old=beta
}
if(standardize==TRUE)for(j in 1:p)beta[j]=beta[j]/scale[j];
for(j in 1:p)coef.seq[k,j]=  ## (2) ここにいれる
## 各 j=1,...,p に対して、coef.seq[[j]] の最後に係数を追加する。
}
return(coef.seq)
}

crime=read.table("crime.txt"); X=crime[,3:7]; y=crime[,1];
coef.seq=warm.start(X,y,300)
plot(log(lambda.seq),coef.seq[,1], xlab="log(lambda)", ylab="係数",
     ylim=c(min(coef.seq),max(coef.seq)), type="n", col="red")
for(j in 1:p){
par(new=TRUE)
lines(log(lambda.seq),coef.seq[,j], col=j)
}
```

29. λ のときの係数を $\beta(\lambda)$ として、上記の横軸の `log(lambda.seq)` を、$||\beta(\lambda)||_1/||\beta(0)||_1$ におきかえ、その出力 (pdf) を提出せよ。ヒント: coef.seq の初期化と更新の後に、`L1.seq=NULL`(初期化) および `L1.seq=c(L1.seq,sum(abs(beta)))`(更新) をおく。また、`lambda.seq=seq(200,10,-10)` は `lambda.seq=seq(200,0,-` とする。そして、最後の plot の行を削除して、以下で置きかえる。

```
L1=sum(abs(linear(X,y)$beta))
L1.seq=L1.seq/L1
plot(L1.seq,beta, xlim=c(0,1), ylim=c(-12,12), type="n", col="red")
```

30. 下記は、データセットから AIC で線形回帰の説明変数を選択する処理である。一般に説明変数が p 個ある場合に、AIC の値を何回計算して比較する必要があるか。また、実際に、犯罪率のデータに適用 (`crime=read.table("crime.txt"); X=crime[,3:7]; y=crime[,1]`) して、最適な変数の組を選択せよ。

```
X=as.matrix(X); y=as.vector(y); X=scale(X); y=y-mean(y); p=ncol(X); beta=array(dim=p)
```

```
AIC.min=Inf
for(k in 1:p){
        A=combn(1:p,k)
        q=ncol(A)
        for(h in 1:q){
                T=A[,h]
                beta[T]=solve(t(X[,T])%*%X[,T])%*%t(X[,T])%*%y     #傾きの最尤値
                Z=0;for(j in T)Z=Z+X[,j]*beta[j]
                S=sum((y-Z)^2)/n
                AIC=log(S)+2*k                 #AICの値の計算
                print(c(S,k,AIC))              #可視化してわかりやすくする
                if(AIC<AIC.min){AIC.min=AIC; k.min=k; T.min=T}
        }
}
k.min; T.min; L.min
```

31. Lassoで実行すると、特に $\lambda > 0$ であれば、一部の変数を選択する。下記 ($\lambda = 30$) の場合、何個の変数を選択するか。そして、選択した変数は、Lassoを用いずにその変数だけで線形回帰を行った場合と比べて、推定した係数の絶対値が小さくなることを確認せよ (テキスト p12 と同じ値が得られればよい)。

```
crime=read.table("crime.txt")
lm.fit=lm(V1~V3+V4+V5+V6+V7, data=crime)
summary(lm.fit)
X=as.matrix(crime[,3:7])
y=as.vector(crime[,1])
p=ncol(X)
for(j in 1:p)X[,j]=X[,j]-mean(X[,j])
y=y-mean(y)
glmnet(X,y,lambda=30)$beta
lm.fit=lm(V1~ ##ここをうめる
      , data=crime)
summary(lm.fit)
```

32. Lassoの最適な λ の値を、交差検証法 (CrossValidation) によって求めたい。また、CVをしない訓練データすべてを用いて係数を推定して、それをテストデータにも用いて2乗誤差を計算したグラフを加えたい。前者を赤で、後者を青の線で描くとして、下記の関数の出力をどのように利用すればよいか。(1)-(4) をうめよ。また、出力をpdfで提出せよ。

```
cv.linear.lasso=function(X,y,lambda.max=100,K=10){
X=as.matrix(X); y=as.vector(y);
p=ncol(X); n=nrow(X);
for(j in 1:p)X[,j]=X[,j]-mean(X[,j]);
y=y-mean(y);
dec=round(lambda.max/50);
lambda.seq=seq(lambda.max,1,-dec);
r=length(lambda.seq);
S=array(0,dim=r);
        m=round(n/K)
```

```
    for(i in 1:K){
    test=seq(i,n,m)
train=setdiff(1:n,test);
coef.seq=warm.start(X[train,],y[train],lambda.max);
for(h in 1:r)S[h]=S[h]+sum((y[test]-X[test,]%*%coef.seq[h,])^2)/m;
    }
S=S/K;
S.min=Inf; for(h in 1:r)if(S[h]<S.min){S.min=S[h]; k=h};
coef.seq=warm.start(X,y,lambda.max);
T=array(0,dim=r);for(h in 1:r)T[h]=sum((y-X%*%coef.seq[h,])^2)/n;
return(list(lambda.min=lambda.seq[k], S.seq=S, T.seq=T, lambda.seq=lambda.seq));
}

result=cv.linear.lasso(X,y)
S.seq=  ## (1) ここをうめる
T.seq=  ## (2) ここをうめる
lambda.seq=result$lambda.seq

plot(lambda.seq, S.seq, xlab="lambda", ylab="テスト時の2乗誤差",
     ylim=c(min(S.seq, T.seq),max(S.seq, T.seq)), col="red", type="n")
lines( ## (3) ここをうめる
lines( ## (4) ここをうめる
```

33. 下記は、Bootstrap といって、データセットからランダムに n 個の行を抜き出して、実行して母数を推定する方法である。何度か実行して標本平均、標本分散をとって、推定することができる。どのような処理をしているか、説明せよ。また、出力を pdf で提出せよ (実行に数分程度かかる)。

```
lambda.max=100
M=100
dec=round(lambda.max/50);
lambda.seq=seq(lambda.max,1,-dec);
r=length(lambda.seq);
SS.seq=array(0,dim=r)
SS2.seq=array(0,dim=r)
for(i in 1:M){
index=sample(1:n,n)
S.seq=cv.linear.lasso(X[index,],y[index])$S.seq
SS.seq=SS.seq+S.seq
SS2.seq=SS2.seq+S.seq^2
}
mid=SS.seq/M
sgm=sqrt(SS2.seq/(M-1)-mid^2)
plot(log(lambda.seq), mid, xlab="log(lambda)", ylab="テスト時の2乗誤差",
     ylim=c(min(mid-sgm),max(mid+sgm)), type="n")
lines(log(lambda.seq),mid+sgm,col="blue")
lines(log(lambda.seq),mid-sgm,col="blue")
lines(log(lambda.seq),mid,col="red")
```

34. `X=as.matrix(X);y=vector(y);cv=cv.glmnet(X,y,grouped=FALSE);plot(cv)` を実行して [2]、出力を pdf でせよ。最上部 n の 0 から 5 までの数値は、どういう意味か。

35. p 個の説明変数 $X_1, \cdots, X_6$ のうちの最初の 2 変数が同じものであった場合 ($X_1 = X_2$)、ridge 回帰では $\hat{\beta}_1 = \hat{\beta}_2$ となることを示せ。

36. 下記は、変数 X_1, X_2, X_3 および X_4, X_5, X_6 のそれぞれで、強い相関をもつ乱数を発生させ、変数 Y への線形回帰を Lasso で行うものである (テキスト p56 と同じ図が表示される)。$X_1, \cdots, X_6$ の係数に対応した色に凡例をつけて (X_i が `col=i`)、その出力を提出せよ。問題 21 を参考にし、`c("X1","X2","X3","X4","X5","X6")`, `col=1:6` として、左上におくとよい

```
n=500; x=array(dim=c(6,n)); z=array(dim=c(2,n))
for(i in 1:2)z[i,]=rnorm(n)
y=3*z[1,]-1.5*z[2,]+2*rnorm(n)
for(j in 1:3)x[j,]=z[1,]+rnorm(n)/5
for(j in 4:6)x[j,]=z[2,]+rnorm(n)/5
glm.fit=glmnet(t(x),y); plot(glm.fit)
```

37. 通常の lasso や ridge ではなく、

$$\frac{1}{2N}||y - \beta_0 - X\beta||_2^2 + \lambda\{(1-\alpha)||\beta||_2^2 + \alpha||\beta||_1\} \tag{1}$$

を最小にする β_0, β を求めたい。ridge は $\alpha = 0$、lasso は $\alpha = 1$ の場合に相当する。すなわち、正則化項を各係数 β_j について、$(1-\alpha)\frac{1}{2}\beta_j^2 + \alpha|\beta_j|$ とする。(1) を最小にする β_j は、lasso の場合 $\hat{\beta}_j = \dfrac{S_\lambda(\frac{1}{N}\sum_{i=1}^N r_{i,j}x_{i,j})}{\frac{1}{N}\sum_{i=1}^N x_{i,j}^2}$、ridge の場合 $\hat{\beta}_j = \dfrac{\frac{1}{N}\sum_{i=1}^N r_{i,j}x_{i,j}}{\frac{1}{N}\sum_{i=1}^N x_{i,j}^2 + \lambda}$ となる。一般には、

$$\hat{\beta}_j = \frac{S_{\lambda\alpha}(\frac{1}{N}\sum_{i=1}^N r_{i,j}x_{i,j})}{\frac{1}{N}\sum_{i=1}^N x_{i,j}^2 + \lambda(1-\alpha)}$$

となる (elastic net) ことを示せ。

38. 問題 19 の linear.lasso を修正して、関数の引数として `alpha=1` を設定し、問題 37 の公式で置き換えて、関数を一般化せよ。

39. 問題 36 の `glm.fit=glmnet(t(x),y)` に、`alpha=0.3` のオプションを含めて、出力を pdf で提出せよ。また、`alpha=0` (ridge)、`alpha=1`(lasso) とした場合でどのような差異があるか。

40. 問題 28 の関数 warm.start に `alpha=1` の引数を入れて一般化して、$\alpha = 0.3, 1$ のそれぞれで、3 変数ずつの 2 グループにわかれるかどうか確認せよ (問題 39 とは横軸が異なる)。

[2]"grouped=TRUE" にすると、K 個の各グループの 2 乗誤差の和を $K-1$ で割ることになる。"grouped=FALSE" だと K でわる