NTNU

Kunnskap for en bedre verden

Institutt for matematiske fag

Eksamensoppgave i **TMA4268 Statistisk læring**

Faglig kontakt under eksamen: Mette Langaas

Tlf: 988 47 649

Eksamensdato: 24. mai 2018

Eksamenstid (fra–til): 09:00–13:00

Hjelpemiddelkode/Tillatte hjelpemidler: C: Ett gult A5-ark med dine egne håndskrevne notater (stemplet av Institutt for matematiske fag). Bestemt kalkulator.

Annen informasjon:

- Alle svar skal begrunnes og besvarelsen skal inneholde naturlig mellomregning.
- For hvert problem er maksimal score gitt.
- Vær god og start svarene dine med Q1–Q28.

Målform/språk: bokmål

Antall sider: 10

Antall sider vedlegg: 0

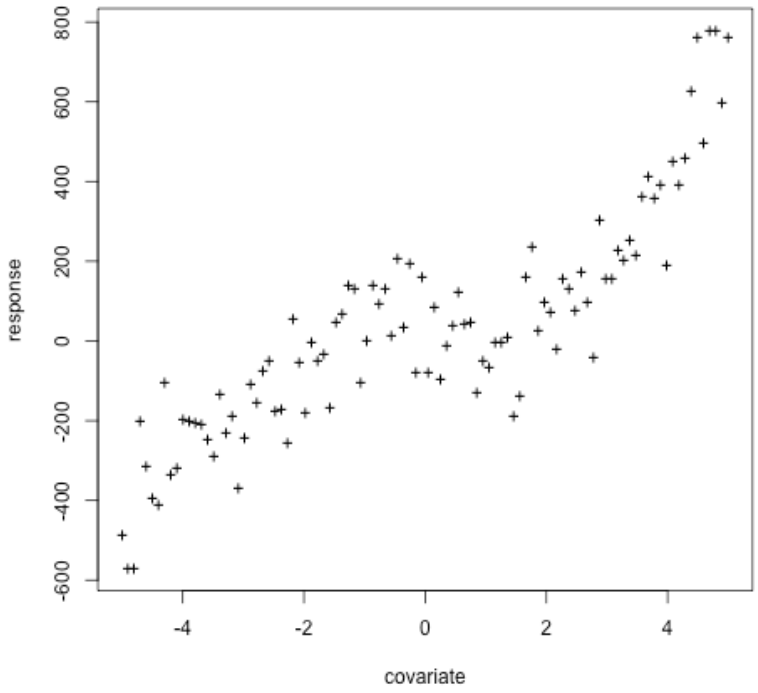
Kontrollert av:

Informasjon om trykking av eksamensoppgave
Originalen er:
1-sidig ☐ **2-sidig** ☒
sort/hvit ☒ **farger** ☐
skal ha flervalgskjema ☐

Dato Sign

Oppgave 1 ***K*-nærmeste-nabo-regresjon** [10 points]

Vi har en univariat kontinuerlig stokastisk variabel Y og en kovariat x. Videre har vi observert et treningssett med uavhengige observasjonspar $\{x_i, y_i\}$ for $i = 1, \ldots, 100$. Et spredningsplott av treningsdataene er gitt i figur 1.



Figur 1: Treningsdata. Observasjoner er vist med '+'.

Anta følgende regresjonsmodell

$$Y_i = f(x_i) + \varepsilon_i$$

der f er den sanne regresjonskurven, og ε_i er en uobservert stokastisk variabel med forventningsverdi lik 0 og konstant varians σ^2 (ikke avhengig av kovariaten).

Nå er målet vårt å finne et estimat for den sanne regresjonskurven ved bruk av K-nærmeste-nabo-regresjon.

Q1: Skriv ned formelen for K-nærmeste-nabo-regresjonskurven i kovariatverdien x_0, og forklar notasjonen du har brukt.

I figur 2 har vi brukt K-nærmeste-nabo-regresjon med K lik 3, 15, 50 og 100, til å estimere regresjonskurven fra treningsdataene. De fire panelene A–D samsvarer med K lik 3, 15, 50 og 100, men ikke nødvendigvis i den rekkefølgen.

Q2: Koble sammen panel A–D med riktig verdi av K (3,15,50,100). Begrunn valget ditt.

For å bruke denne metoden må vi velge verdi for parameteren K. Det kan gjøres ved 5-fold kryssvalidering.

Q3: Forklar hvordan 5-fold kryssvalidering utføres, og spesifiser hvilket avviksmål du vil bruke. Svaret ditt må inneholde en formel for hvordan avviksmålet beregnes under kryssvalideringen. Inkluder en skisse.

Et alternativ til 5-fold kryssvalidering er «leave-one-out» kryssvalidering.

Q4: Vil du foretrekke «leave-one-out» fremfor 5-fold kryssvalidering for denne situasjonen? Begrunn valget ditt.

Oppgave 2 En viktig dekomponering innen regresjon [10 points]

Vi har en univariat kontinuerlig stokastisk variabel Y og en kovariat x. Videre har vi et treningssett av uavhengige observasjonspar $\{x_i, y_i\}$ for $i = 1, \ldots, n$. Anta en regresjonsmodell

$$Y_i = f(x_i) + \varepsilon_i$$

der f er den sanne regresjonsfunksjonen, og ε_i er en uobserverbar stokastisk variabel med forventningsverdi 0 og konstant varians σ^2 (ikke avhengig av kovariaten). Ved bruk av treningssettet kan vi finne et estimat av regresjonsfunksjonen f, som vi kaller $\hat{f}$. Vi ønsker å bruke $\hat{f}$ til å predikere en ny observasjon i kovariatverdien x_0 (denne nye observasjonen er ikke avhengig av observasjonene i treningssettet). Den predikerte responsverdien er da $\hat{f}(x_0)$. Vi er interessert i feilen i denne prediksjonen.

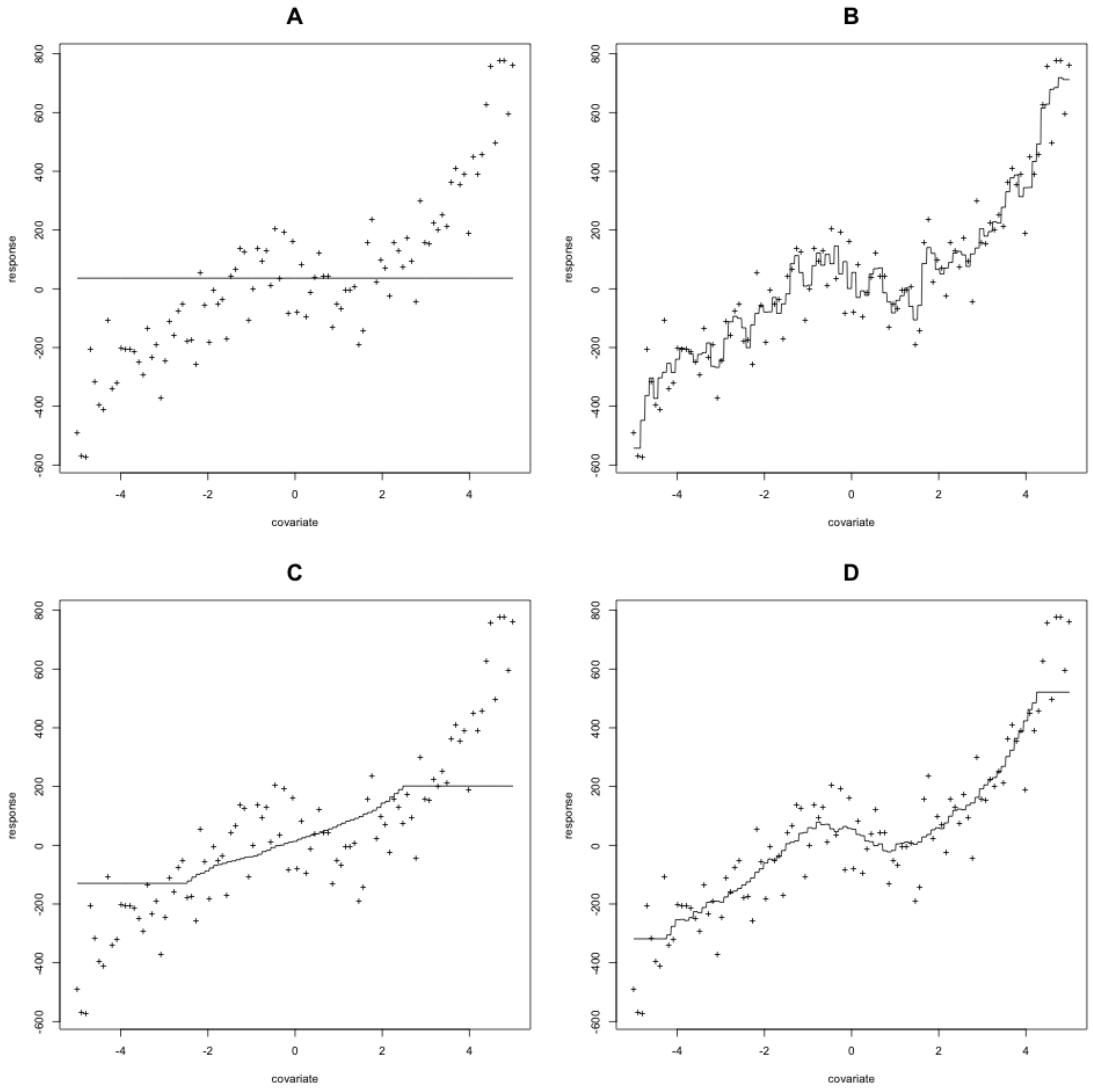
Q5: Skriv ned definisjonen av den forventede testmiddelkvadratfeilen («expected test mean squared error») (MSE) i x_0.
Q6: Utled dekomponering av den forventede testmiddelkvadratfeilen MSE i tre ledd.
Q7: Forklar med ord hvordan vi kan tolke disse tre leddene.

Anta at vi har en metode for å estimere regresjonsfunksjonen, der metoden har en kompleksitetsparameter som kontrollerer kompleksiteten til modellen, og en stor verdi av parameteren gir høy modelkompleksitet.

Q8: Lag en skisse av hvordan den forventede testmiddelkvadratfeilen MSE (i x_0) utvikler seg som en funksjon av kompleksitetsparameteren. Gjør det samme for de tre leddene.

Denne dekomponeringen har vært sentral i dette emnet.

Q9: Hva synes du er den viktigste følgen (implikasjonen) av denne dekomponeringen? Svar med *bare en* setning.



Figur 2: Observasjoner er vist som '+' og heltrukne kurver gir løsning av K-nærmeste nabo-regresjon med K lik 3,15, 50 og 100 (i en eller annen rekkefølge).

Oppgave 3 Lineær diskriminantanalyse [10 points]

Her kan du bruke at sannsynlighetstettheten til en p-dimensjonal multivariat normalfordelt stokastisk variabel $\mathbf{X}$ med forventningsverdi $\boldsymbol{\mu}$ og varians-kovariansmatrise $\boldsymbol{\Sigma}$ er gitt som

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}\det(\boldsymbol{\Sigma})^{1/2}} \exp\Big(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\Big).$$

Vi vil ikke diskutere hvordan man estimerer forventningsverdien og varians-kovariansmatrisen her, så du kan anta at de er kjente.

Q10: Skriv ned de matematiske modellantagelsene til en lineær diskriminantanalyse med to klasser (kodet som 0 og 1) og p forklaringsvariabler, og forklar hva de ulike ingrediensene er.
Q11: Forklar hvordan du utleder den matematiske formelen for aposteriorisannsynligheten til klasse 1.
Q12: Utled den matematiske formelen til klassegrensen mellom de to klassene, gitt at klassifikasjonsregelen er å klassifisere til klassen med høyest aposteriorisannsynlighet.
Q13: Er denne klassegrensen lineær eller ikke-lineær i rommet til kovariatene?

Oppgave 4 Klassifikasjon av diabetestilfeller

Vi ser på diabetesdata (`diabetes` er kodet '0' hvis ikke til stede og '1' hvis til stede) fra en populasjon av kvinner fra Pimaindiansk herkomst i USA, tilgjengelig fra `MASS` R-pakken. Følgende kovariater ble samlet inn for hver kvinne:

- `npreg`: antall graviditeter
- `glu`: plasmaglukosekonsentrasjon i en oral glukosetoleransetest
- `bp`: diastolisk blodtrykk (mmHg)
- `skin`: triceps hudfoldtykkelse (mm)
- `bmi`: kroppsmasseindeks (vekt i kg/(høyde i m)^2)
- `ped`: diabetes-arvelighetsfunksjon
- `age`: alder i år

Vi vil bruke et treningssett (kalt `train`) med 200 observasjoner (132 ikke-diabetikere og 68 diabetikere) og et testsett (kalt `test`) med 332 observasjoner (223 ikke-diabetikere og 109 diabetikere). Målet vårt er å lage en klassifikasjonsregel for diabetes (eller ikke) basert på de innsamlede dataene.

I figur 3 finner du R-kode og resultater fra en logistisk regresjon på treningssettet `train`.

a) [10 points]

Q14: Skriv ned den statistiske modellen for den logistiske regresjonen.
Q15: Hva er den estimerte effekten av `ped`-kovariaten på diabetesstatus? Forklar.
Q16: Vil du predikere at en person med følgende verdier for kovariatene har diabetes?

Person: `npreg`=2, `glu`=145, `bp`=85, `skin`=35, `bmi`=37, `ped`=0.7, `age`=40.

b) [10 points]

Q17: Tegn et «feedforward neural network» med samme arkitektur som den logistiske regresjonen og spesifiser hvilke(n) aktiveringsfunksjon(er) som brukes. Bruk samme notasjon for noder og koblinger som i den matematiske modellen for den logistiske regresjonen.

Testsettet `test` ble brukt til å evaluere oppførselen til den logistiske regresjonen. En «receiver–operator curve» (ROC) ble laget og er den heltrukne kurven i figur 4 (de stiplede og prikkede kurvene vil vi ser mer på i Q24).

Q18: Forklar kort hvordan en ROC-kurve lages. Forklaringen din bør inneholde ordene grense («cut-off»), forvirringsmatrise, sensitivitet, spesifisitet.

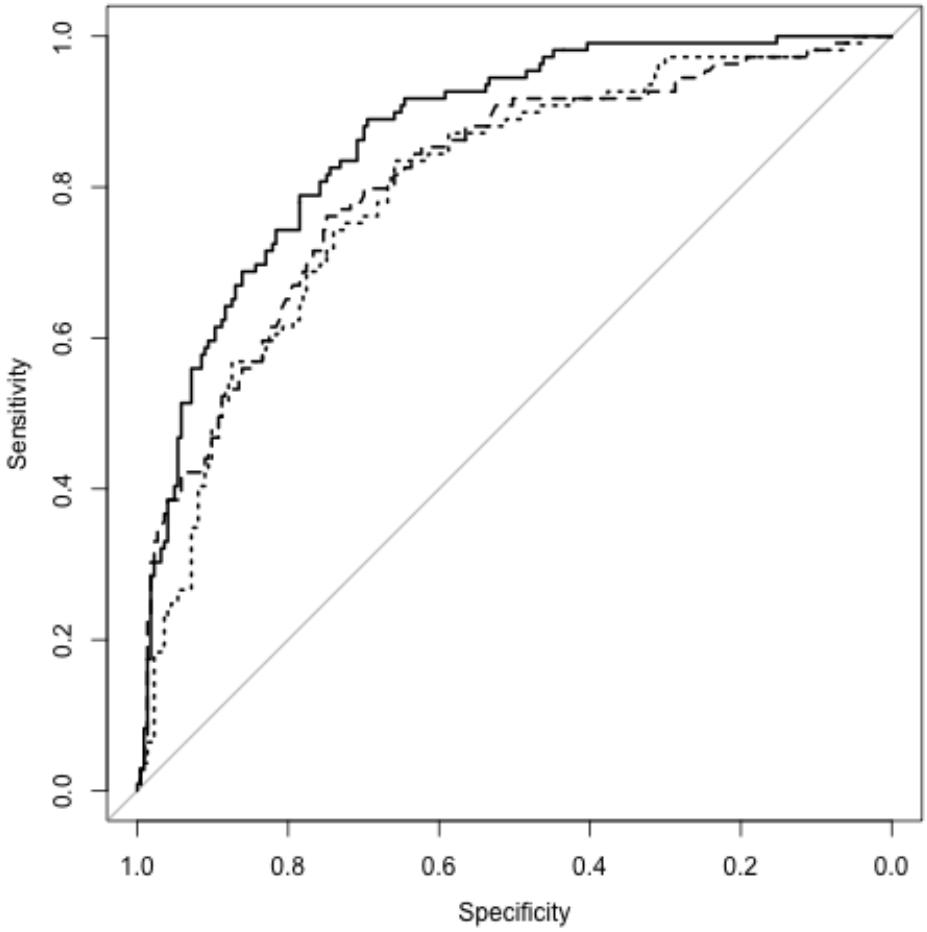
Se på forvirringsmatrisen (sannsynlighetsgrense 0.5) for testsettet i figur 3 (nedre del av utskriften).

Q19: Hvilket punkt på ROC-kurven gir denne grensen?

```
# logistic regression
> fitlogist=glm(diabetes~npreg+glu+bp+skin+bmi+ped+age,data=train,
family=binomial(link="logit"))
> summary(fitlogist)
Call:
glm(formula = diabetes ~ npreg + glu + bp + skin + bmi + ped +
    age, family = binomial(link = "logit"), data = train)
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -9.773062   1.770386  -5.520 3.38e-08 ***
npreg        0.103183   0.064694   1.595  0.11073
glu          0.032117   0.006787   4.732 2.22e-06 ***
bp          -0.004768   0.018541  -0.257  0.79707
skin        -0.001917   0.022500  -0.085  0.93211
bmi          0.083624   0.042827   1.953  0.05087 .
ped          1.820410   0.665514   2.735  0.00623 **
age          0.041184   0.022091   1.864  0.06228 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 256.41  on 199  degrees of freedom
Residual deviance: 178.39  on 192  degrees of freedom
AIC: 194.39
Number of Fisher Scoring iterations: 5

> predlogist=predict(fitlogist,newdata=test,type="response")
> testclasslogist=ifelse(predlogist > 0.5, 1, 0)
> table(test$diabetes, testclasslogist)
   testclasslogist
      0   1
  0 200  23
  1  43  66
>  # note: 223 true non-diabetes and 109 true diabetes cases
> library(pROC)
> roclogist=roc(test$diabetes,predlogist)
> auc(roclogist)
Area under the curve: 0.8659
> plot(roclogist, lty="solid")
```

Figur 3: R-kode og resultater fra den logistiske regresjonen.



Figur 4: ROC-kurver for logistisk regresjon (heltrukket), «bagged»-trær (stiplet) og kvadratisk diskriminantanalyse (prikket) for diabetestestsettet.

Vi har brukt «bagging» i kombinasjon med klassifikasjonstrær for treningssettet `train`, og testet på testsettet `test`. R-kode og resultater finner du i figur 5.

c) [10 points]

Q20: Forklar hvordan vi har laget et «bagged» sett av trær, og hvorfor vi kunne ønske å tilpasse mer enn ett tre?

For å estimere feilrater (for «bagged»-trær) er det mulig å bruke et såkalt «out-of-bag» (OOB) utvalg.

Q21: Anta at vi har et datasett av størrelse n. Beregn sannsynligheten for at en gitt observasjon er i et gitt bootstraputvalg.
Q22: Hva er et OOB-utvalg?
Q23: Bruk resultatene i figur 5 til å sammenligne misklassifikasjonsraten for OOB-utvalget med misklassifikasjonsraten for testsettet `test` og kommenter.

Vi har nå laget klassifikasjonsregler for diabetes ved help av logistisk regresjon og «bagged»-trær. I tillegg har vi også tilpasset en kvadratisk diskriminantanalyse (R-kode og resultater i figur 5). Resultatet av alle disse tre metodene er presentert som ROC-kurver i figur 4. Her er den heltrukne kurven fra logistisk regresjon, den stiplede fra «bagged»-trær og den prikkete fra kvadratisk diskriminantanalyse.

Q24: Hvilken av de tre metodene ville du brukt hvis målet var å predikere diabetesstatus?

```
> # bagged trees with call to randomForest
> library(randomForest)
> library(pROC)
> set.seed(4268)
> rf=randomForest(factor(diabetes)~npreg+glu+bp+skin+bmi+ped+age,
+                 data=train,
+                 mtry=7,ntree=1000,importance=TRUE)
> rf$confusion #error rates based on OOB data
    0  1 class.error
0 108 24   0.1818182
1  31 37   0.4558824
> yrf=predict(rf,newdata=test)
# note: test data 332 observations with
# 223 non-diabetes and 109 diabetes cases
> table(test$diabetes, yrf)
   yrf
      0   1
  0 184  39
  1  42  67
> predrf = predict(rf,test, type = "prob")
> rocrf=roc(test$diabetes, predrf[,2])
> auc(rocrf)
Area under the curve: 0.8094

> # fitting and ROC for QDA
> fitqda=qda(diabetes~npreg+glu+bp+skin+bmi+ped+age,data=train)
> predqda = predict(fitqda,newdata=test)
> testclassqda=ifelse(predqda$posterior[,2] > 0.5, 1, 0)
> table(test$diabetes, testclassqda)
   testclassqda
      0   1
  0 194  29
  1  47  62
> rocqda=roc(test$diabetes,predqda$posterior[,2])
> auc(rocqda)
Area under the curve: 0.7962

> # plotting all tree methods with ROC
> plot(roclogist) #logistic regression solid line
> plot(rocrf,add=TRUE,lty="dashed") #bagged trees dashed
> plot(rocqda,add=TRUE,lty="dotted") #QDA dotted line
```

Figur 5: R-kode og resultater fra «bagged»-trær og kvadratisk diskriminantanalyse.

Oppgave 5 **Klyngeanalyse** [10 points]

Vi ser på et datasett med 5 observasjoner (merket a–e) av to variabler (x_1, x_2). Dette vises i den venstre tabellen i figur 6. I tillegg er den euklidske avstandsmatrisen mellom observasjonene delvis gitt i den høyre tabellen.

Observasjoner

label	x_1	x_2
a	1	1
b	2	1
c	4	5
d	7	7
e	5	7

Avstandsmatrise

	a	b	c	d	e
a	0.0				
b	?	0.0			
c	5.0	4.5	0.0		
d	8.5	?	3.6	0.0	
e	7.2	6.7	2.2	?	0.0

Figur 6: Observasjoner og euklidsk avstandsmatrise for klyngeanalyse.

Q25: I den euklidske avstandsmatrisen er det tre manglende elementer, markert med spørsmålstegn (høyre tabell i figur 6). Beregn verdier for de manglende elementene.
Q26: Utfør hierarkisk klyngeanalyse med komplett linkage («complete linkage») og tegn dendrogram.
Q27: Vi bruker dendrogrammet og ønsker to klynger, hvilke observasjoner er i hver klynge?

En konkurrerende metode for klyngeanalyse er «k-means».

Q28: Diskuter kort *to forskjeller* mellom hierarkisk og «k-means»-klyngeanalyse.