NTNU

Kunnskap for ei betre verd

Institutt for matematiske fag

Eksamensoppgåve i **TMA4268 Statistisk læring**

Fagleg kontakt under eksamen: Mette Langaas

Tlf: 988 47 649

Eksamensdato: 24. mai 2018

Eksamenstid (frå–til): 09:00–13:00

Hjelpemiddelkode/Tillatne hjelpemiddel: C: Ett gult A5-ark med dine eigne handskrivne notat (stempla av Institutt for matematiske fag). Bestemd kalkulator.

Annan informasjon:

- Alle svar skal grunngis og løysinga skal romme naturleg mellomrekning.
- For kvart problem er maksimal score gitt.
- Vær god og start svara dine med Q1–Q28.

Målform/språk: nynorsk

Sidetal: 10

Sidetal vedlegg: 0

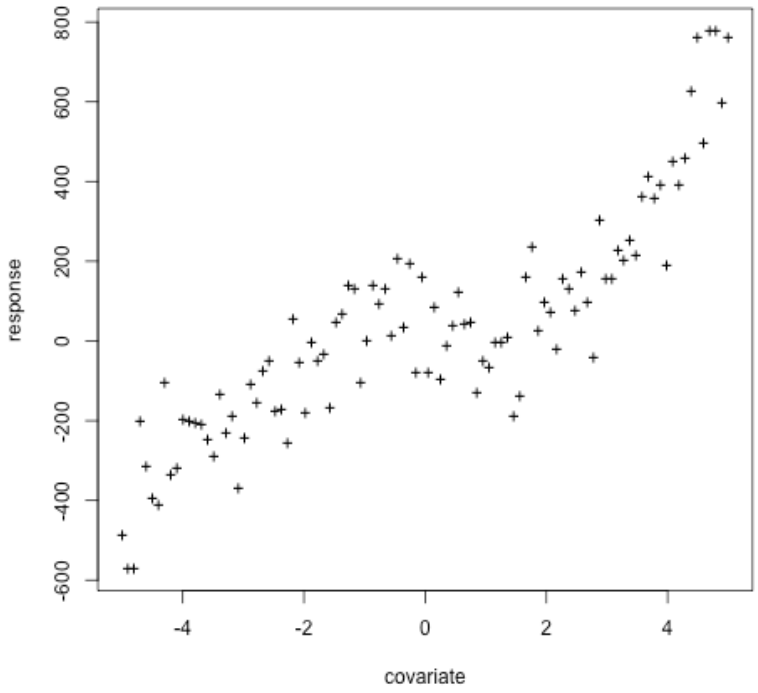
Kontrollert av:

Informasjon om trykking av eksamensoppgåve	
Originalen er:	
1-sidig ☐	**2-sidig** ☒
svart/kvit ☒	**fargar** ☐
skal ha fleirvalskjema ☐	

Dato Sign

Oppgåve 1 ***K*-næraste-nabo-regresjon** [10 points]

Vi har ein univariat kontinuerleg stokastisk variabel Y og ein kovariat x. Vidare har vi observert eit treningssett med uavhengige observasjonspar $\{x_i, y_i\}$ for $i = 1, \ldots, 100$. Eit spreiingsplott av treningsdataa er gitt i figur 1.



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

Anta følgande regresjonsmodell

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

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

No er målet vårt å finne eit estimat for den sanne regresjonskurva ved å bruke K-næraste-nabo-regresjon.

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

I figur 2 har vi brukt K-næraste-nabo-regresjon med K lik 3, 15, 50 og 100, til å estimere regresjonskurva fra treningsdataa. Dei fire panela A–D samsvarer med K lik 3, 15, 50 og 100, men ikkje nødvendigvis i den rekkefølga.

Q2: Koble saman panel A–D med rett verdi av K (3,15,50,100). Grunngi valet ditt.

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

Q3: Forklar korleis 5-fold kryssvalidering utførast, og spesifiser kva for avviksmål du vil bruke. Svaret ditt må innehalde ein formel for korleis avviksmålet reknast ut under kryssvalideringa. Inkluder ei skisse.

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

Q4: Vil du velje «leave-one-out» for 5-fold kryssvalidering i denne situasjonen? Grunngi valet ditt.

Oppgåve 2 Ei viktig dekomponering innan regresjon [10 points]

Vi har ein univariat kontinuerleg stokastisk variabel Y og ein kovariat x. Vidare har vi eit treningssett av uavhengige observasjonspar $\{x_i, y_i\}$ for $i = 1, \ldots, n$. Anta ein regresjonsmodell

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

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

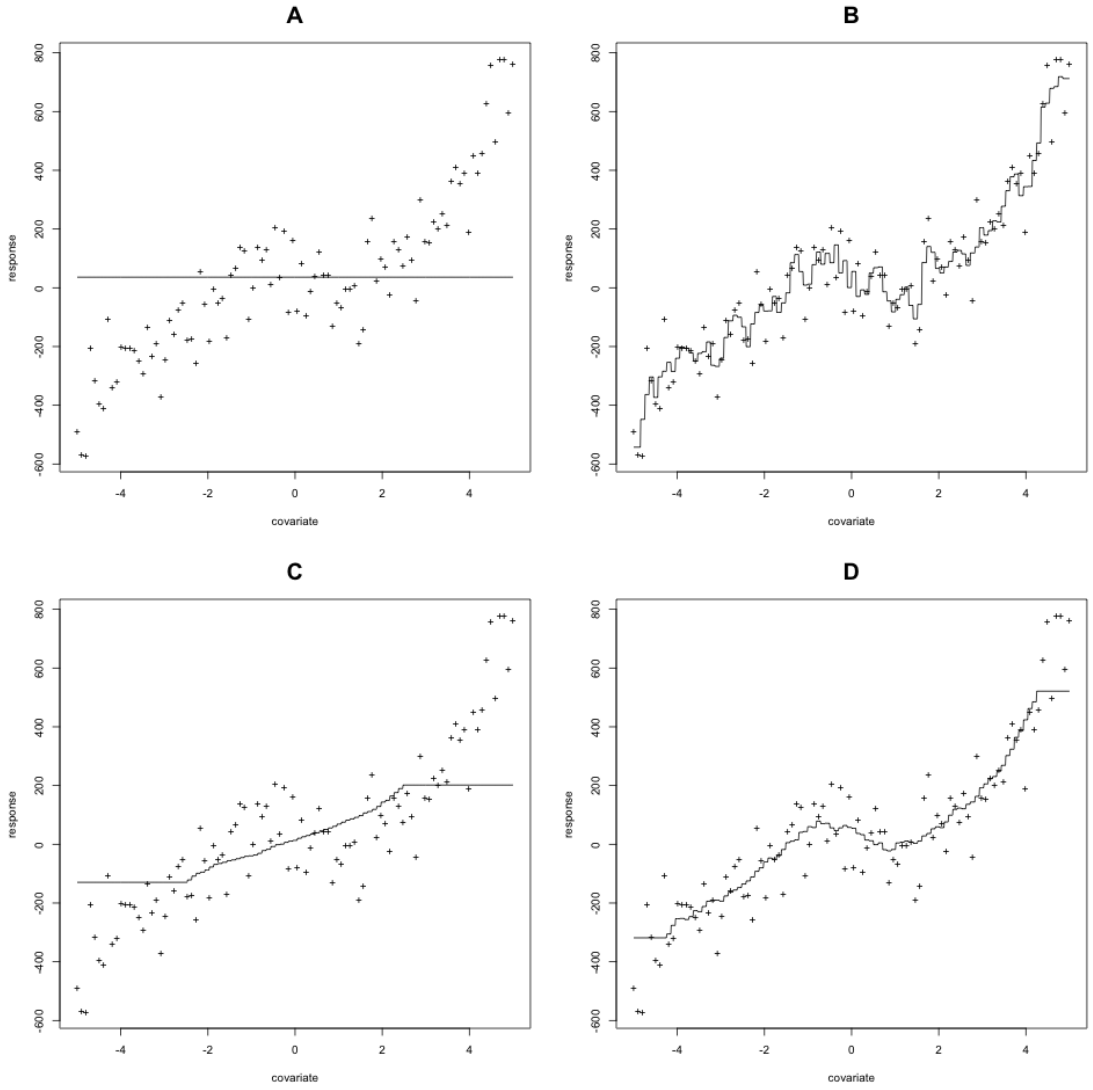
Q5: Skriv ned definisjonen av den forventa testmiddelkvadratfeilen («expected test mean squared error») (MSE) i x_0.
Q6: Utlei dekomponering av den forventa testmiddelkvadratfeilen MSE i tre ledd.
Q7: Forklar med ord korleis vi kan tolke desse tre ledda.

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

Q8: Lag ei skisse av korleis den forventa testmiddelkvadratfeilen MSE (i x_0) utviklar seg som ein funksjon av kompleksitetsparameteren. Gjer det same for dei tre ledda.

Denne dekomponeringa har vore sentral i dette emnet.

Q9: Kva synes du er den viktigaste implikasjonen av denne dekomponeringa? Svar med *berre ei* setning.



Figur 2: Observasjonar er vist som '+' og heiltrekte kurver gir løysing av K-næraste nabo-regresjon med K lik 3,15, 50 og 100 (i ein eller annan rekkefølge).

Oppgåve 3 Lineær diskriminantanalyse [10 points]

Her kan du bruke at sannsynstettleiken til ein 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 ikkje diskutere korleis ein estimerer forventningsverdien og varians-kovariansmatrisa her, så du kan anta at dei er kjende.

Q10: Skriv ned dei matematiske modellføresetnadene til ei lineær diskriminantanalyse med to klassar (koda som 0 og 1) og p forklaringsvariablar, og forklar kva dei ulike ingrediensane er.
Q11: Forklar korleis du utleier den matematiske formelen for aposteriorisannsynet til klasse 1.
Q12: Utlei den matematiske formelen til klassegrensa mellom dei to klassene, gitt at klassifikasjonsregelen er å klassifisere til klassen med høgast aposteriorisannsyn.
Q13: Er denne klassegrensa lineær eller ikkje-lineær i rommet til kovariatane?

Oppgåve 4 Klassifikasjon av diabetestilfelle

Vi ser på diabetesdata (`diabetes` er koda '0' dersom ikkje til stades og '1' dersom til stade) frå ein populasjon av kvinner frå Pimaindiansk avstamning i USA, tilgjengeleg frå `MASS` R-pakken. Følgande kovariatar blei samla inn for kvar kvinne:

- `npreg`: talet på graviditetar
- `glu`: plasmaglukosekonsentrasjon i ein oral glukosetoleransetest
- `bp`: diastolisk blodtrykk (mmHg)
- `skin`: triceps hudfoldtjukkleik (mm)
- `bmi`: kroppsmasseindeks (vekt i kg/(høyde i m)^2)
- `ped`: diabetes-arvelighetsfunksjon
- `age`: alder i år

Vi vil bruke eit treningssett (kalt `train`) med 200 observasjonar (132 ikkje-diabetikarar og 68 diabetikarar) og eit testsett (kalt `test`) med 332 observasjonar (223 ikkje-diabetikarar og 109 diabetikarar). Målet vårt er å lage ein klassifikasjonsregel for diabetes (eller ikkje) basert på dei innsamla kovariatane.

I figur 3 finn du R-kode og resultatar frå ein logistisk regresjon på treningssettet `train`.

a) [10 points]

Q14: Skriv ned den statistiske modellen for den logistiske regresjonen.
Q15: Kva er den estimerte effekten av `ped`-kovariaten på diabetesstatus? Forklar.
Q16: Vil du predikere at ein person med følgande verdiar for kovariatane har diabetes?

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

b) [10 points]

Q17: Tekn eit «feedforward neural network» med same arkitektur som den logistiske regresjonen og spesifiser kva for aktiveringsfunksjon(ar) brukast. Bruk same notasjon for nodar og koplingar som i den matematiske modellen for den logistiske regresjonen.

Testsettet `test` blei brukt til å evaluere oppførselen til den logistiske regresjonen. Ei «receiver–operator curve» (ROC) blei laga og er den heiltrekte kurva i figur 4 (dei stipla og prikka kurvene vil vi ser meir på i Q24).

Q18: Forklar kort korleis ei ROC-kurve lagast. Forklaringa di bør innehalde orda grense («cut-off»), forvirringsmatrise, sensitivitet, spesifisitet.

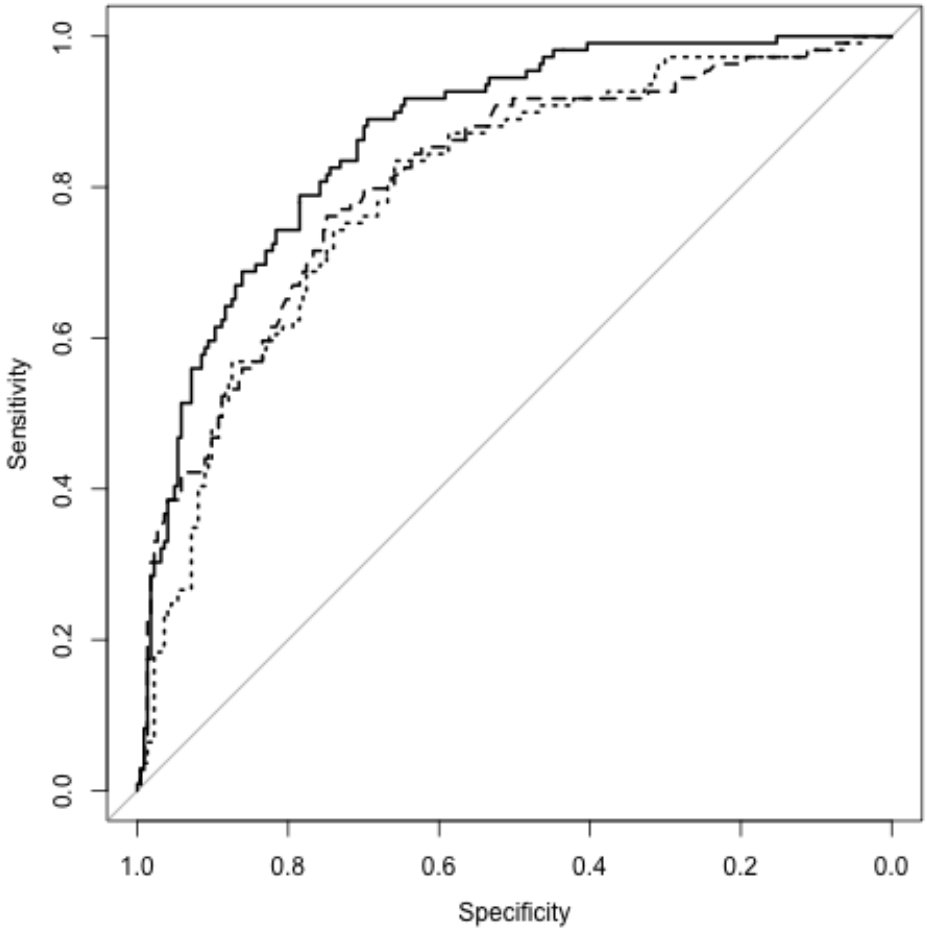
Se på forvirringsmatrisa (sannsynsgrense 0.5) for testsettet i figur 3 (nedre del av utskrifta).

Q19: Kva punkt på ROC-kurva gir denne grensa?

```
# 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 resultat frå den logistiske regresjonen.



Figur 4: ROC-kurver for logistisk regresjon (heiltrekt), «bagged»-trær (stipla) og kvadratisk diskriminantanalyse (prikka) for diabetestestsettet.

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

c) [10 points]

Q20: Forklar korleis vi har laga eit «bagged» sett av trær, og kvifor vi kunne ønske å bruke meir enn ett tre?

For å estimere feilratar (for «bagged»-trær) er det mogeleg å bruke eit såkalla «out-of-bag» (OOB) utval.

Q21: Anta at vi har eit datasett av storleik n. Berekn sannsynet for at ein gitt observasjon er i eit gitt bootstraputval.
Q22: Kva er eit OOB-utval?
Q23: Bruk resultatane i figur 5 til å sammenlikne misklassifikasjonsraten for OOB-utvalet med misklassifikasjonsraten for testsettet `test` og kommenter.

Vi har nå laga klassifikasjonsreglar for diabetes ved help av logistisk regresjon og «bagged»-trær. I tillegg har vi og tilpassa ein kvadratisk diskriminantanalyse (R-kode og resultat i figur 5). Resultata frå alle desse tre metodane er vist som ROC-kurver i figur 4. Her er den heiltrekte kurva frå logistisk regresjon, den stipla frå «bagged»-trær og den prikka frå kvadratisk diskriminantanalyse.

Q24: Kva for metode (av dei tre) 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 resultat frå å tilpasse «bagged»-trær og kvadratisk diskriminantanalyse.

Oppgåve 5 **Klyngeanalyse** [10 points]

Vi ser på eit datasett med 5 observasjonar (merka a–e) av to variablar (x_1, x_2). Dette ser du i den venstre tabellen i figur 6. I tillegg er den euklidske avstandsmatrisa mellom observasjonane delvis gitt i den høgre tabellen.

Observasjonar

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: Observasjonar og euklidsk avstandsmatrise for klyngeanalyse.

Q25: I den euklidske avstandsmatrisa er det tre manglande element, markert med spørretekn (høgre tabell i figur 6). Berekn verdiar for dei manglande elementa.
Q26: Utfør hierarkisk klyngeanalyse med komplett linkage («complete linkage») og tekn dendrogram.
Q27: Vi bruker dendrogrammet og ønsker to klynger, kva for observasjonar er i kvar klynge?

Ein konkurrerande metode til klyngeanalyse er «k-means».

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