UNIVERSITETET I OSLO

Det matematisk-naturvitenskapelige fakultet

Eksamen i:	STK2100 — Maskinlæring og statistiske metoder for prediksjon og klassifikasjon
Eksamensdag:	Torsdag 15. juni 2017.
Tid for eksamen:	09.00 – 13.00.
Oppgavesettet er på 7 sider.	
Vedlegg:	Ingen
Tillatte hjelpemidler:	Godkjent kalkulator og formelsamlinger for STK1100/STK1110 og STK2100

Kontroller at oppgavesettet er komplett før du begynner å besvare spørsmålene.

Oppgave 1

I denne oppgaven skal vi se på et datasett for boligpriser i forsteder til Boston. Som responsvariabel vil vi ha

MEDV Median verdi av bolig innenfor et område (i $1000)

Det er 11 forklaringsvariable. Det er ikke viktig å forstå hva disse er i de etterfølgende spørsmål, men en beskrivelse av disse er gitt nedenfor. Alle variable, bortsett fra CHAS (binær) og RAD (kategorisk med 9 nivåer) er numeriske.

CRIM Kriminalitetsrate per innbygger

ZN Andelen boligområder over 25 000 kvadrat fot.

CHAS Binær variabel, lik 1 hvis område grenser til Charles River

NOX Nitrogrenoksidkonsentrasjon (del av 10 millioner)

RM Gjennomsnittelig rom per bolig

AGE Andel boliger bygget før 1940

DIS Vektet avstand til fem Boston sysselsetting sentre

RAD Indeks for tilgjengelighet til motorveier (kategorisk)

PTRATIO Elev-lærer forholdet i byen

B $1000(Bk - 0.63)^2$ der Bk er andel av mørkhudede

(Fortsettes på side 2.)

LSTAT % med lavere status i befolkningen

Det er totalt $n = 506$ områder (observasjoner), men all analyse nedenfor er basert på oppdeling i et treningsset og et testsett.
Oppdelingen i trenings- og testsett er gjort ved å trekke tilfeldig 253 observasjoner som brukes til trening og de resterende til test. Som referanse videre vil vi bruke resultater fra en lineær regresjonsmodell. Tilpasning med minste kvadraters metode ga følgende regresjonstabell:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	38.071527	7.841521	4.855	2.20e−06
CRIM	−0.140742	0.040019	−3.517	0.000524
ZN	0.047965	0.019805	2.422	0.016204
CHAS1	5.205026	1.401956	3.713	0.000256
NOX	−16.911410	5.702240	−2.966	0.003332
RM	2.614760	0.617887	4.232	3.33e−05
AGE	0.019537	0.019267	1.014	0.311638
DIS	−1.262763	0.294073	−4.294	2.57e−05
RAD2	4.604395	2.072367	2.222	0.027254
RAD3	6.099546	1.914405	3.186	0.001638
RAD4	2.636360	1.776669	1.484	0.139187
RAD5	4.019631	1.781579	2.256	0.024981
RAD6	2.250257	2.255377	0.998	0.319441
RAD7	6.901910	2.114468	3.264	0.001262
RAD8	5.704424	2.151248	2.652	0.008557
RAD24	7.544125	1.989127	3.793	0.000190
PTRATIO	−0.983504	0.213207	−4.613	6.54e−06
B	0.007554	0.003570	2.116	0.035400
LSTAT	−0.695442	0.073822	−9.421	< 2e−16

Log-likelihood verdien for denne lineære modellen er -742.34. Gjennomsnittlig feilrate for testdata basert på denne modellen var 25.14.

(a) Hvorfor er det lurt å dele opp i trenings- og testsett tilfeldig i forhold til andre strategier?

Gitt denne tabellen, argumenter for hvorfor det kan være rimelig å fjerne AGE fra modellen.

(b) Nedenfor er en regresjonstabell gitt der AGE er fjernet

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	37.268866	7.801931	4.777	3.14e−06
CRIM	−0.141442	0.040016	−3.535	0.000492
ZN	0.046455	0.019750	2.352	0.019493
CHAS1	5.339371	1.395765	3.825	0.000167
NOX	−15.305548	5.478227	−2.794	0.005637
RM	2.719095	0.609295	4.463	1.26e−05
DIS	−1.359251	0.278268	−4.885	1.92e−06

(Fortsettes på side 3.)

RAD2	4.647958	2.072045	2.243	0.025818
RAD3	6.042215	1.913684	3.157	0.001800
RAD4	2.556993	1.775050	1.441	0.151052
RAD5	4.009193	1.781656	2.250	0.025358
RAD6	2.223959	2.255363	0.986	0.325110
RAD7	6.935950	2.114328	3.280	0.001194
RAD8	5.818663	2.148425	2.708	0.007259
RAD24	7.358077	1.980765	3.715	0.000254
PTRATIO	−0.955146	0.211378	−4.519	9.86e−06
B	0.007894	0.003554	2.221	0.027306
LSTAT	−0.665377	0.067610	−9.841	< 2e−16

Log-likelihood verdien for denne modellen er -742.90 mens gjennomsnittlig feilrate for testdata basert på denne modellen var 24.86.

Hvis du ser bort i fra testdataene, utfør en prosdyre for modelvalg mellom de to lineære regresjonsmodeller. Hva blir konklusjonen av dette? Virker det rimelig i forhold til hva som kom ut av testdataene?

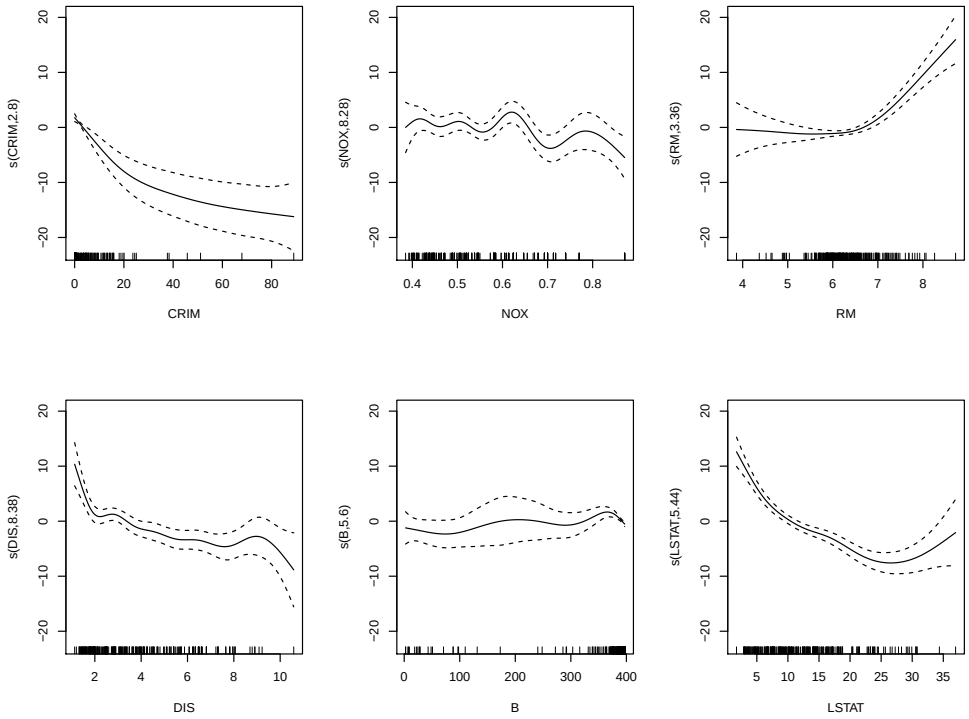
Basert på utskriften, kunne du tenke deg en ytterligere forenkling av modellen?

(c) En alternativ modell er GAM. Plottene nedenfor viser de ikke-lineære funksjoner som inngikk i denne modellen. Log-likelihood verdien for denne modellen er -615.79 mens estimert antall frihetsgrader er 46.12.

Forklar hvordan antall frihetsgrader blir beregnet i dette tilfellet. Bruk dette til å sammenlikne denne modellen med de tidligere modeller. Kommenter resultatet.

Gjennomsnittlig feilrate for testdata basert på denne modellen var 15.52. Er dette i samsvar med de modell sammenlikninger du har gjort?

(Fortsettes på side 4.)



(d) Nok en alternativ modell kan fås ved å bruke regresjonstrær. Nedenfor er et plot av et regresjonstre basert på 9 endenoder.

Diskuter hvorfor regresjonstrær gir mulighet for å inkludere *interaksjoner* mellom forklaringsvariable.

Forklar hvorfor en rimelig likelihood funksjon i dette tilfellet er

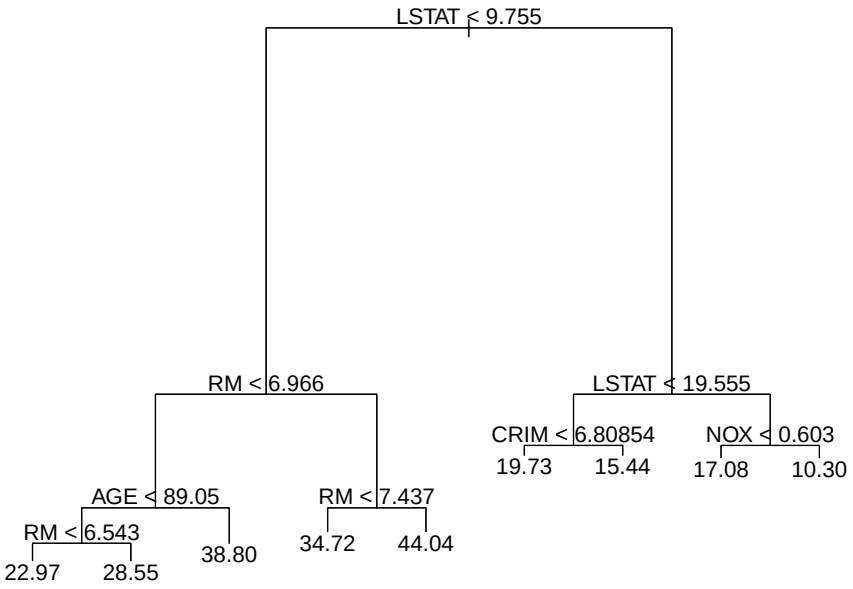
$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(y_i - \mu_i)^2}$$

der $\mu_i = c_m$ hvis $\boldsymbol{x}_i \in R_m$. Hva slags antagelser bygger en slik likelihood på?

Hvor mange parametre må spesifiseres for å tilpasse et tre med 9 endenoder?

Vurdér denne modellen mot tidligere modeller når du får oppgitt at log-likelihood verdien (med estimerte verdier for $\boldsymbol{\theta}$) i dette tilfellet blir -697.49.

(Fortsettes på side 5.)



(e) Alternative metoder som Bagging, Random Forest og Boosting ga følgende resultater (der Feil er estimert kvadratisk feil på testsett):

Metode	Feil
Regresjonstre	17.16
Bagging	11.36
Random Forrest	11.28
Boosting	11.58

Beskriv **kort** disse tre metodene og kommenter resultatene. Diskuter spesielt forbedringene i forhold til GAM modellen.

Oppgave 2

Vi vil her se på en klassifikasjonssetting. La $Y \in \{1, ..., G\}$ være variabelen av interesse og anta at vi observerer $\boldsymbol{x} \in \mathcal{R}^p$. Vi har som vanlig data $\{(y_i, \boldsymbol{x}_i), i = 1, ..., n\}$ der $y_i \in \{1, ..., G\}$ mens $\boldsymbol{x}_i \in \mathcal{R}_i$. Vi ønsker å predikere Y basert på $\boldsymbol{x}$.

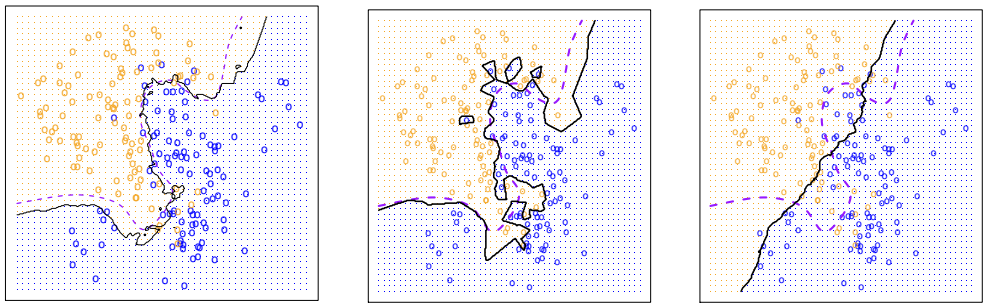
(a) Anta vi ønsker å bruke K-nærmeste nabo metoden for klassifikasjon.

Forklar hvordan denne metoden fungerer. Hva er styrker og svakheter med denne metoden?

(Fortsettes på side 6.)

(b) Nedenfor vises 3 plot av nærmeste nabo metoden for $G = 2$ og $K = 1, 10$ eller 100 (men ikke nødvendigvis i denne rekkefølgen). Her gir den heltrukne linjen klassifikasjonsgrensen mellom de to klassene mens den stiplede linjen gir den optimale grensen (data her er simulerte slik at vi vet den underliggende sanne modell). Fargene på punktene og områdene angir klasseverdier for observasjoner og klassifikasjoner, henholdsvis.

Spesifiser hvilke figurer som tilhører de ulike K verdier. Hvilken verdi av K vil du foretrekke? Begrunn svaret.



a b c

Anta nå vi innfører en tapsfunksjon

$$L(y, \hat{y}) = \begin{cases} 1 & \text{hvis } \hat{y} \neq y; \\ 0 & \text{ellers.} \end{cases}$$

som sier noe om hvor alvorlig vi måler feil som gjøres.

(c) Vis at den optimale prediktor i denne situasjonen er

$$\widehat{Y}(\boldsymbol{x}) = \arg\max_g \Pr(Y = g|\boldsymbol{x}).$$

Forklar hvorfor det dermed er viktig å estimere $f_g(\boldsymbol{x}) = E[I(Y = g)|\boldsymbol{x}]$ for $g = 1, ..., G$ der $I(A) = 1$ hvis begivenheten A er tilfredsstilt og 0 ellers.

(d) Forklar hvordan man kan bruke *regresjonsmetoder* for å estimere $f_g(\boldsymbol{x})$ og dermed bruke regresjonsmetoder for å konstruere klassifikasjonsmetoder.

(e) Diskuter ulike metoder for å estimere forventet tap i denne klassifikasjonssettingen. Ta spesielt opp styrker og svakheter med ulike metoder.

(f) Anta nå $\hat{f}_g(\boldsymbol{x})$ er et estimat på $f_g(\boldsymbol{x})$. Vis at

$$\begin{aligned} &E[(f_g(\boldsymbol{x}_0) - \hat{f}_g(\boldsymbol{x}_0))^2|\boldsymbol{x}_0] \\ &=(f_g(\boldsymbol{x}_0) - E[\hat{f}_g(\boldsymbol{x}_0)|\boldsymbol{x}_0])^2 + E[(\hat{f}_g(\boldsymbol{x}_0) - E[\hat{f}_g(\boldsymbol{x}_0)|\boldsymbol{x}_0])^2|\boldsymbol{x}_0] \end{aligned}$$

Gi en fortolkning av de to hovedleddene på høyre side av likhetstegnet.

(Fortsettes på side 7.)

(g) La nå $\hat{f}_{g,1}(\boldsymbol{x})$ være et estimat på $f_b(\boldsymbol{x})$ basert på en ganske restriktiv metode/modell mens $\hat{f}_{g,2}(\boldsymbol{x})$ er basert på en mer fleksibel tilnærming. Diskuter de ulike leddene i likningen ovenfor i denne settingen.

Oppgave 3

I Ridge regresjon ønsker vi å minimere mhp $\boldsymbol{\beta}$

$$h(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

Vi vil anta forklaringsvariablene er sentrerte slik at $\sum_{i=1}^{n} x_{ij} = 0$ for alle j.

(a) Forklar hvorfor det kan være hensiktsmessig også å skalere x_{ij}-ene slik at $\frac{1}{n}\sum_{i=1}^{n} x_{ij}^2 = 1$ for alle j.

(b) Vis at

$$\begin{aligned}\hat{\beta}_0^{ridge} =& \bar{y} \\ \hat{\boldsymbol{\beta}}^{ridge} =& \hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X} + \lambda \boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{y}\end{aligned}$$

for passende spesifisering av $\boldsymbol{X}$. Her er $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)$.

(c) Anta nå at alle x-ene er ukorrelerte slik at $\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{I}$. Anta også at den sanne modell er $Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_j$ der $\varepsilon_1, ..., \varepsilon_p$ er uavhengige og alle har forventning 0 og varians σ^2.

Utled i dette tilfellet både forventningsvektoren og kovariansmatrisen til $\hat{\boldsymbol{\beta}}^{ridge}$.

Diskuter disse resultatene i forhold til de avveininger vi ofte må gjøre i regresjonssammenhenger.

Hint: Vis først at $E[\boldsymbol{Y}] = \beta_0\mathbf{1} + \boldsymbol{X}\boldsymbol{\beta}$ der $\mathbf{1}$ er en vektor bestående av 1-ere samt at $\boldsymbol{X}^T\mathbf{1} = \mathbf{0}$.