UNIVERSITETET I OSLO

Det matematisk-naturvitenskapelige fakultet

Eksamen i:	STK2100 — Maskinlæring og statistiske metoder for prediksjon og klassifikasjon
Eksamensdag:	Torsdag 14. juni 2018.
Tid for eksamen:	14.30 – 18.30.
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

Vi skal i denne oppgaven se på et datasett om overlevelse etter Titanic katastrofen.
Variablene som inngår er

Survival 0=Nei, 1=Ja, faktor

Age Alder i måneder, en numerisk variabel

Pclass Billett klasse, 1=1st, 2=2nd, 3=3rd, faktor

Sex Kjønn (menn/kvinner), faktor

Sibsp Antall søsken/ektefeller ombord, numerisk.

Parch Antall foreldre/barn ombord, numerisk

Fare Billettpris, numerisk

Cabin Kabin nummer, faktor som i utgangspunktet har 148 ulike verdier, men som er redusert/gruppert til 9; N (no cabin), A, B, C, D, E, F, G, T.

Embarked Havn for ombordstigning, C=Cherbourg, Q=Queenstown, S=Southhampton, faktor

Vi vil se på et delsett av det totale sett bestående av 712 individer.
Vi vil starte med en enkel logistisk regresjonsmodell. Tilpasning ga følgende regresjonstabell:

(Fortsettes på side 2.)

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	3.8723	0.6692	5.79	0.0000
Pclass2	-0.6793	0.5053	-1.34	0.1788
Pclass3	-1.8027	0.5182	-3.48	0.0005
Sexmale	-2.6900	0.2279	-11.80	0.0000
Age	-0.0439	0.0085	-5.15	0.0000
SibSp	-0.3553	0.1306	-2.72	0.0065
Parch	-0.0691	0.1251	-0.55	0.5805
Fare	0.0029	0.0030	0.97	0.3298
CabinA	1.1274	0.7877	1.43	0.1524
CabinB	0.5580	0.6381	0.87	0.3819
CabinC	-0.0680	0.5821	-0.12	0.9070
CabinD	0.9392	0.6146	1.53	0.1265
CabinE	1.5267	0.6049	2.52	0.0116
CabinF	1.2172	0.7936	1.53	0.1251
CabinG	-0.8919	1.0124	-0.88	0.3783
EmbarkedQ	-0.7989	0.6051	-1.32	0.1867
EmbarkedS	-0.4351	0.2838	-1.53	0.1252

Når en bruker denne modellen til å predikere de samme data (ved å predikere til den mest sannsynlige klassen) får vi en feilrate på 19.10%. Log-likelihood verdien for denne modellen er -308.8.

(a) Forklar hvorfor regresjonstabellen lister opp færre rader enn antall nivåer for faktor variablene.

Gitt at vi her har en "Treatment" begrensning (vi setter koeffisienten svarende til det første nivået lik null), hva slags tolkning har da regresjonskoeffisientene for de øvrige nivåer?

(b) Beregn AIC-verdien for denne modellen. Diskutér hvorfor det kan være rimelig å forenkle modellen noe.

(c) Nedenfor er utskrift fra en alternativ modell:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	4.3254	0.4507	9.60	0.0000
Pclass2	-1.4063	0.2848	-4.94	0.0000
Pclass3	-2.6450	0.2859	-9.25	0.0000
Sexmale	-2.6190	0.2150	-12.18	0.0000
Age	-0.0449	0.0082	-5.46	0.0000
SibSp	-0.3786	0.1214	-3.12	0.0018

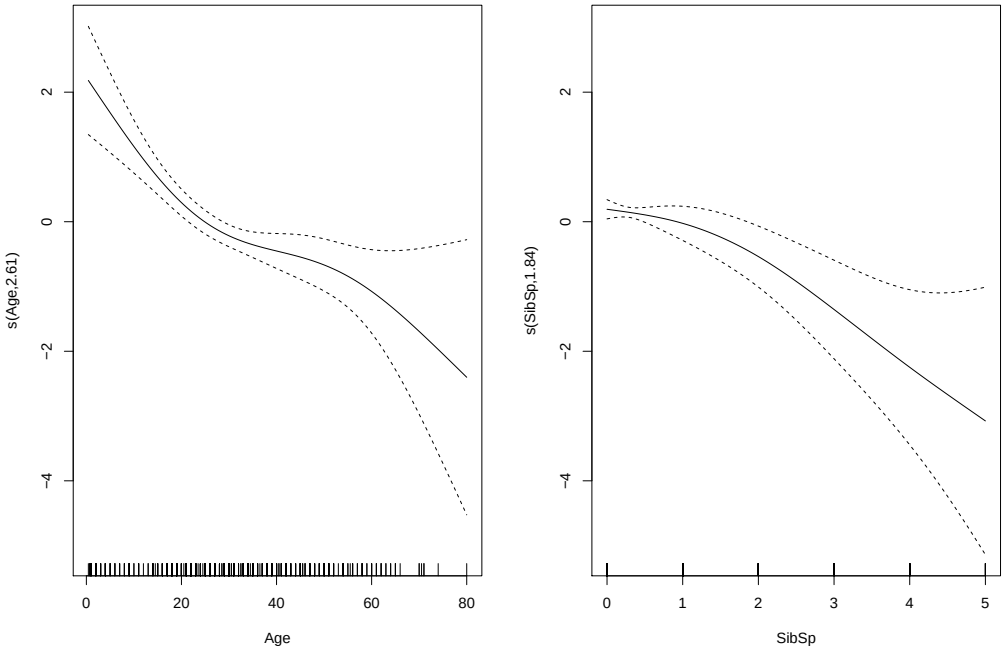
Når en bruker denne modellen til å predikere de samme data (ved å predikere til den mest sannsynlige klassen) får en en feilrate på 19.38%. Log-likelihood verdien for denne modellen er -318.0.

Forklar hvorfor log-likelihood verdien blir *mindre* i dette tilfellet.

Argumentér hvorfor denne modellen likevel er å foretrekke.

(Fortsettes på side 3.)

En alternativ modell er en generaliserte additive modell (GAM). Plottene nedenfor viser de ikke-lineære funksjoner som inngikk i en modellen der de samme forklaringsvariable som i oppgave (*c*) ble brukt. Log-likelihood verdien for denne modellen er -312.2 mens estimert antall frihetsgrader er 8.4.

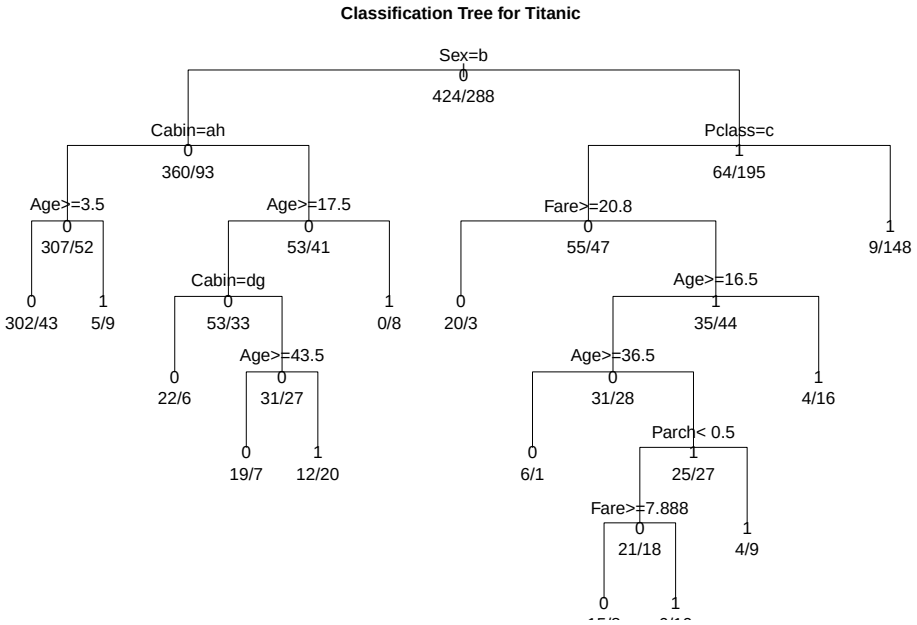


(d) Forklar hvordan antall frihetsgrader blir beregnet i dette tilfellet. Bruk dette til å sammenlikne denne modellen med de tidligere modeller.

Kommentér om plottene viser signifikante ikke-lineariteter.

(e) Nok en alternativ modell kan fås ved å bruke klassifikasjonstrær. Nedenfor er et plot av et klassifikasjonstre basert på 11 endenoder.

(Fortsettes på side 4.)



Diskuteŕ hvorfor klassifikasjonstrær gir mulighet for å inkludere *interaksjoner* mellom forklaringsvariable.

Forklar hvorfor en likelihood funksjon for et klassifikasjonstre med en respons tilhørende to klasser kan skrives på formen

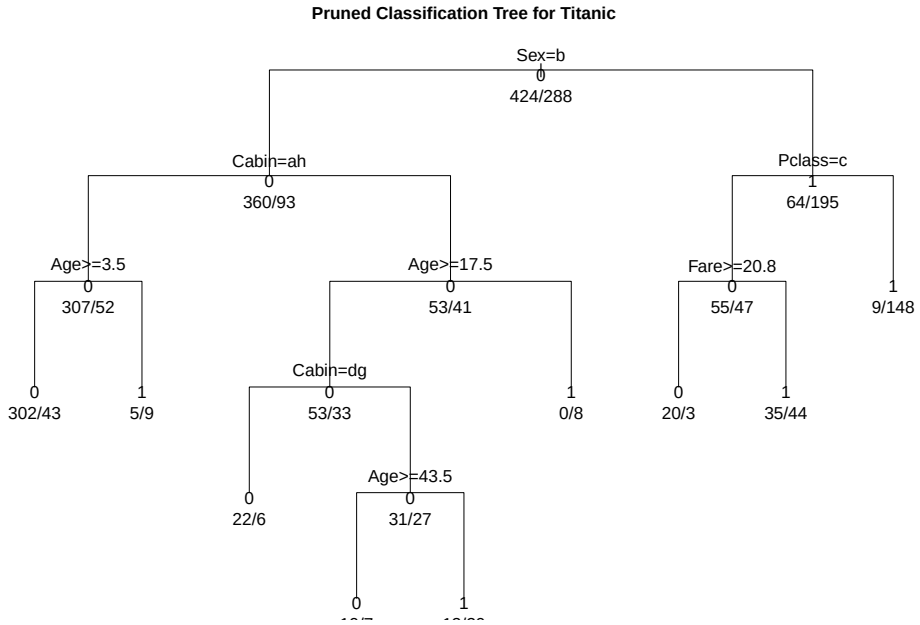
$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i}$$

der $p_i = c_m$ for $\boldsymbol{x}_i \in R_m$.

(f) For det spesifikke tre fikk vi en log-likelihood verdi på -279.452. Bruk dette til å vurdere denne modellen mot de tidligere modellene.

(g) Diskutér hvorfor det kan være hensiktsmessig å *beskjære* trær. Nedenfor er et tre beskåret ned til 9 endenoder. Log-likelihood verdien i dette tilfellet er -287.349. Vurder også denne modellen mot de tidligere modeller.

(Fortsettes på side 5.)



(h) Nedenfor er gitt en tabell over estimert feilrate basert på *kryss-validering* (delt opp i 8 grupper). Alternative metoder som Bagging, Random Forest og nevrale nett er også inkludert.

Metode	Feilrate (%)
Logistisk regresjon, alle variable	15.59
Logistisk regresjon, variabel seleksjon	17.84
GAM, alle variable	11.24
GAM, variabel seleksjon	16.85
Klassifikasjonstre, 11 noder	20.37
Klassifikasjonstre, 9 noder	19.94
Bagging	20.79
Random Forrest	19.38
Nevrale nett (150 latente noder)	20.37
Dype nett (3 latente lag med 50 noder i hver)	22.75

Diskutér fordelene med å bruke kryss-validering for å evaluere metoder.

Beskriv kort hvordan Bagging, Random Forest, Nevrale nett og dype nett fungerer.

(i) Diskutér mulige forklaringer på at de mest enkle metoder ser ut til å fungere best i dette tilfellet.

Anta du velger den metode med minst estimert feilrate, diskutér hvordan du kan si noe om hvor godt denne valgte metoden vil virke. Ta med både styrker og svakheter med ditt valg.

(Fortsettes på side 6.)

Oppgave 2

Anta en lineær regresjonsmodell

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i, \quad i = 1, ..., n$$

der $\varepsilon_i \sim N(0, \sigma^2)$ og alle støyledd er uavhengige.

(a) Vis at vi kan skrive om modellen til

$$Y_i = \tilde{\beta}_0 + \beta_1 \tilde{x}_{i1} + \beta_2 \tilde{x}_{i2} + \varepsilon_i, \quad i = 1, ..., n$$

der $\sum_i \tilde{x}_{i1} = \sum_i \tilde{x}_{i2} = 0$. Hva slags tolkning har $\tilde{\beta}_0$ i denne formuleringen av modellen?

(b) Anta vi ønsker å estimere $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)$ ved minimering av

$$h(\beta_0, \beta_1, \beta_2) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2})^2 + \lambda_1 \beta_1^2 + \lambda_2 \beta_2^2.$$

(Vi vil i det etterfølgende kalle de verdier som minimerer h for de *optimale* verdier).

Diskutér situasjoner hvor det kan være hensiktsmessig å bruke $\lambda_1 \neq \lambda_2$.

Vis at minimering av $h(\boldsymbol{\beta})$ kan oppnås ved minimering av

$$\tilde{h}(\tilde{\beta}_0, \beta_1, \beta_2) = \sum_{i=1}^{n} (y_i - \tilde{\beta}_0 - \beta_1 \tilde{x}_{i1} - \beta_2 \tilde{x}_{i2})^2 + \lambda_1 \beta_1^2 + \lambda_2 \beta_2^2.$$

Finn den optimale verdi for $\tilde{\beta}_0$.

(c) Sett opp et ligningssystem som de optimale verdier av (β_1, β_2) må tilfredsstille.

Under antagelsen om at $\sum_i (x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2) = 0$, utled eksplisitte uttrykk for de optimale verdier for (β_1, β_2). Hva blir den optimale verdi for β_0 da?

Vi vil nå se på `Hitters` datasettet der en ønsker å predikere `Salary` basert på mange ulike forklaringsvariable. Vi vil imidlertid kun se på to av disse her: `PutOuts` og `Hits`. En enkel lineær regresjon basert på de to forklaringsvariablene ga følgende resultater:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	535.9259	24.6013	21.78	0.0000
PutOuts	83.7694	25.8357	3.24	0.0013
Hits	172.7897	25.8357	6.69	0.0000

For å se på effekten av straffeledd, ble 3 ulike forsøk sammenliknet:

(Fortsettes på side 7.)

- $\lambda_1 = \lambda_2 = 0$.
- $\lambda_1 = \lambda_2 = \lambda$ der λ er bestemt ved minimering av kryss-validert estimat på kvadratisk feil.
- $\lambda_1 \neq \lambda_2$ der (λ_1, λ_2) er begge bestemt ved kryss-validert estimat på sum av kvadratisk feil

De kryss-validerte estimatene på sum av kvadratisk feil ble henholdsvis 163367, 163166 ($\lambda = 20.0$) og 163142 ($\lambda_1 = 20.0, \lambda_2 = 12.2$).

(d) Hva slags metoder svarer de 2 første forsøk til?

Basert på resultatene som er oppgitt, hvorfor kan det være rimelig at den optimale felles λ verdi i det andre forsøket svarer til λ_1 i det tredje forsøket?

Diskutér utfordringer i forbindelse med å innføre ulike straffeledd for de forskjellige forklaringsvariable når antall forklaringsvariable øker.