UNIVERSITETET I OSLO

Det matematisk-naturvitenskapelige fakultet

Eksamen i:	STK2100 — Løsningsforslag
Eksamensdag:	Torsdag 14. juni 2018.
Tid for eksamen:	14.30 – 18.30.
Oppgavesettet er på 6 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

(a) I modeller med faktorer, sier regresjonskoeffisientene noe om nivået til de ulike kategoriene. Imidlertid, når også et konstantledd er med, blir det for mange parametre og vi må begrense/redusere disse til en dimensjon lavere. Dette kan gjøres på ulike måter, en er å sette den første lik null, hvor de resterende koeffisientene måler avvik fra den første kategorien.

(b) Vi har

$$\text{AIC} = -2 * \text{log-lik} + 2 * p$$

der p er antall parametre i modellen. Her er $p = 17$ som gir AIC= $-2 * (-308.8) + 2 * 17 = 651.6$.

Siden flere av de estimerte koeffisientene har en tilhørende p-verdi som er ganske høy, tyder det på at vi bør ta bort noen variable.

(c) Når vi gjør begrensninger på modellen, vil vi ha et mindre rom å optimere likelihooden på, noe som medfører lavere verdi.

Her blir AIC=$-2 * (-318.0) + 2 * 6 = 648.0$. Da denne verdien er noe mindre enn hva vi fikk tidligere, er den nye modellen å foretrekke.

(d) For GAM har vi at $\hat{\boldsymbol{y}} = \boldsymbol{S}\boldsymbol{y}$ og frihetsgrader blir da beregnet ved trase($\boldsymbol{S}$). Vi får et høyere antall frihetsgrader her pga ikke-linearitet.

Her blir da AIC=$-2*(-312.2)+2*8.4 = 641.2$. Vi får da en forbedring i forhold til tidligere modeller.

Plottene viser ikke en veldig sterk ikke-linearitet, men gitt mengden data blir den likevel signifikant.

(Fortsettes på side 2.)

(e) Definisjonene av regionene vil være kombinasjoner av logiske operatorer basert på ulike forklaringsvariable. Dermed kommer interaksjoner inn.

Vi har at hver Y_i er binomisk fordelt med ett forsøk. Sannsynlighetene for å få 1 kan variere fra observasjon til observasjon. Dette er da markert ved å ha en indeks i på p_i. Ved å i tillegg anta uavhengighet mellom responsene, får vi da produktet av ledd av typen $p_i^{y_i}(1-p_i)^{1-y_i}$.

Siden vi for klassifikasjonstrær antar at sannsynlighetene er like innenfor hver region, blir da $p_i = c_m$ for $\boldsymbol{x}_i \in R_m$.

(f) Det er ikke helt opplagt hvordan en skal telle antall parametre i dette tilfellet.

Vi har 13 endenoder som gir 13 c_m parametre. I mange situasjoner bruker en dette som antall parametre.

I tillegg har vi imidlertid 12 oppsplittinger. Hver oppsplitting har to parametre, en som spesifiserer hvilken variabel som skal splittes opp og en som spesifiserer hvilken verdi oppsplittingen skal skje på. Totalt blir det dermed $13 + 2 * 12 = 37$ parametre.

(En kan imidlertid argumentere for at Sex ikke har noen ekstra spesifisering av hvor oppsplitting skal skje slik at en eventuelt også kunne bruke 36 parametre. Merk at for andre faktorer med mer enn 2 nivåer må en bestemme et "nivå" gjennom hvordan oppsplitting skjer.)

Dette gir en AIC verdi på

$$AIC = -2 * (-279.5) + 2 * 37 = 633.0$$

dvs noe bedre enn vi fikk med de tidligere modeller.

(g) Trær gir ofte overtilpasning. En mulighet er å stoppe oppsplitting tidligere, men da kan en miste interaksjoner som kommer senere. Det er derfor vanlig å først lage et stort tre og så beskjære dette for å minske variansen.

I prinsippet blir frihetsgrader her enda vanskeligere å beregne siden vi i prinsippet bør ta hensyn til hele prosessen for å generere det beskjærte treet. Hvis vi imidlertid kun forholder oss til størrelsen på det endelige tre, får vi 9+2*8=25 frihetsgrader. Da blir

$$AIC = -2 * (-287.349) + 2 * 25 = 624.698$$

som gir en ytterligere reduksjon i forhold til tidligere verdier. Kombinert med at vi å får et noe enklere tre å forholde oss til er derfor dette treet å foretrekke.

(h) For å få et realistisk mål på hvordan en metode fungerer, må det evalueres på data som ikke er blitt brukt til trening. En mulighet

(Fortsettes på side 3.)

er å dele opp i et treningssett og et testsett, men da vil vi få et mindre treningssett å estimere modellen med. Kryss-validering utnytter data bedre ved å "sirkulere" testsettet.

En ønsker ofte å måle metoder ved å se på hvordan det oppfører seg på nye datasett. Slike nye datasett er imidlertid ikke alltid tilgjengelig AIC (som kun bruker treningsdatasettet) vil ikke alltid gi et realistisk mål på hvor god en modell/metode er (baserer seg mye på modell-antagelser). Et bedre mål kan være prediksjonsfeil på nye data. Hvis vi imidlertid ikke har for mye data, vil vi tape endel estimeringsstyrke ved å ta bort en del av dataene til test. Kryss-validering har sin styrke i at det er en metode som både oppnår et stort testsett (faktisk hele datasettet) og samtidig får et treningssett som er ganske stort (en andel $(K-1)/K$ der K er antall grupper). En ekstra fordel med CV er at det kan parallelliseres slik at beregningstid ikke nødvendigvis blir alt for stor.

Bagging og Random Forest: Begge tar utgangspunkt i at trær kan ha stor variasjon (egentlig en hvilken som helst metode med stor varians) og robustifiserer dette ved å istedet kombinere mange prediktorer basert på ulike datasett. De ulike datasett blir konstruert ved bootstrapping. Bagging og Random Forest skiller seg ved at Bagging benytter alle forklaringsvariable ved hvert splitt mens Random Forest gjør begrensninger i settet av variable for å oppnå mindre korrelasjon mellom de ulike trær (prediktorer) som blir laget.

Nevrale nett er gitt ved

$$z_{im} = h(\boldsymbol{\alpha}_m^T \boldsymbol{x}_i), \qquad m = 1, ..., M \tag{1}$$

$$T_i = \beta_0 + \boldsymbol{\beta}^T \boldsymbol{z}_i \tag{2}$$

$$y_i = g(T_i) + \varepsilon_i \tag{3}$$

der både $h(\cdot)$ and $g(\cdot)$ er mulige *ikke-lineære* funktioner. Figur 1 illustrerer modellen. z-ene kan oppfattes som *latente* variable. Dype nett oppnås ved å ha flere lag med latente variable.

(i) Resultatene kan tyde på at interaksjoner likevel ikke er så viktige i denne situasjonen (alle de beste modellene er av GAM typen). Videre kan det se ut som ikke-lineæriteter *er* viktig, men at variabel-seleksjonen mhp GAM ikke fungerer så godt.

En mulig metode for å evaluere feilraten er å bruk den verdi man har fått på den valge metode. Merk imidlertid at selv om hver av feilratene kan være forventningsrette innenfor hver metode, så vil vi nå bruke minimum av 10 variable. Et slikt minimum vil ikke lenger være forventningsrett, og vil typisk være noe for optimistisk. Ideelt sett burde vi hatt et ekstra test-sett å vurdere den endelige modell på.

(Fortsettes på side 4.)

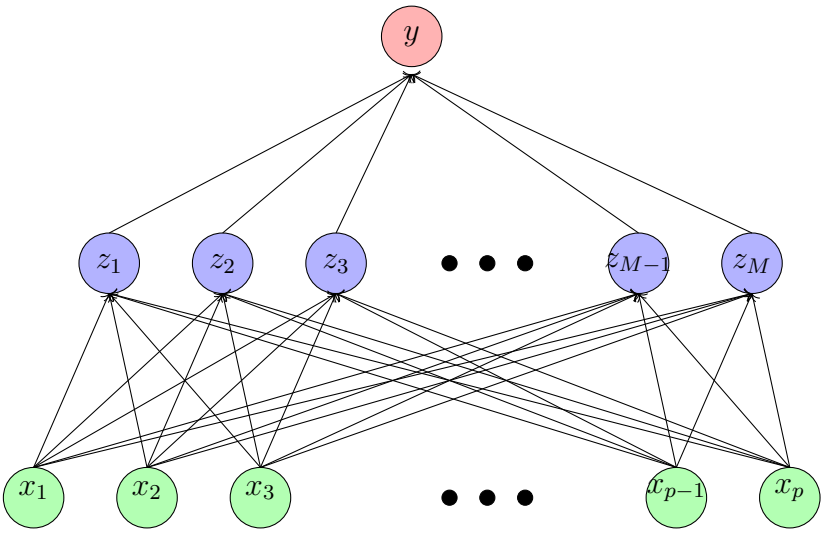

Figure 1: Visualisation of neural network with one hidden layer.

For det spesifikke problemet vil imidlertid prediksjon på nye data ikke være så aktuelt, man er mer interessert i å "lære" sammehenger. Sett fra dette perspektivet er det bra at en rimelig enkel model blir valgt, dog kanskje litt negativt at ikke noen av variablene blir valgt bort.

Oppgave 2

(a) Vi har at

$$\begin{aligned} Y_i =& \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i \\ =& \beta_0 + \beta_1 \bar{x}_1 + \beta_2 \bar{x}_2 + \beta_1(x_{i1} - \bar{x}_1) + \beta_2(x_{i2} - \bar{x}_2) + \varepsilon_i \\ =& \tilde{\beta}_0 + \beta_1 \tilde{x}_{i1} + \beta_2 \tilde{x}_{i2} + \varepsilon_i \end{aligned}$$

der

$$\begin{aligned} \tilde{\beta}_0 =& \beta_0 + \beta_1 \bar{x}_1 + \beta_2 \bar{x}_2 \\ \tilde{x}_{i1} =& x_{i1} - \bar{x}_1 \\ \tilde{x}_{i2} =& x_{i2} - \bar{x}_2 \end{aligned}$$

$\tilde{\beta}_0$ angir nå forventet nivå når begge forklaringsvariable har verdier lik gjennomsnittsverdiene av de observerte x-er.

(b) Hvis forklaringsvariablene har veldig ulike skalaer, kan det være hensiktsmessig å legge ulike straffeledd på disse. Et alternativ kunne være å skalere x-ene på forhånd. Ikke opplagt hva som er best.

Siden det er en en-til-en korrespondanse mellom $(\beta_0, \beta_1, \beta_2)$ og $(\tilde{\beta}_0, \beta_1, \beta_2)$ med $\tilde{\beta}_0 = \beta_0 + \beta_1 \bar{x}_1 + \beta_2 \bar{x}_2$ og vi har at

$$h(\beta_0, \beta_1, \beta_2) = \tilde{h}(\beta_0 + \beta_1 \bar{x}_1 + \beta_2 \bar{x}_2, \beta_1, \beta_2),$$

(Fortsettes på side 5.)

vil de to minimeringsproblemene være ekvivalente.

Vi har at

$$\begin{aligned}\frac{\partial}{\partial\tilde{\beta}_0}\tilde{h}(\tilde{\beta}_0,\beta_1,\beta_2) &= -2\sum_{i=1}^{n}(y_i-\tilde{\beta}_0-\beta_1\tilde{x}_{i1}-\beta_2\tilde{x}_{i2})\\ &= -2\sum_{i=1}^{n}(y_i-\tilde{\beta}_0)\end{aligned}$$

som hvis vi setter lik null gir optimal verdi $\hat{\tilde{\beta}}_0 = \bar{y}$.

(c) Vi har at

$$\begin{aligned}\frac{\partial}{\partial\beta_1}\tilde{h}(\tilde{\beta}_0,\beta_1,\beta_2) &= -2\sum_{i=1}^{n}(y_i-\tilde{\beta}_0-\beta_1\tilde{x}_{i1}-\beta_2\tilde{x}_{i2})\tilde{x}_{i1}\\ \frac{\partial}{\partial\beta_2}\tilde{h}(\tilde{\beta}_0,\beta_1,\beta_2) &= -2\sum_{i=1}^{n}(y_i-\tilde{\beta}_0-\beta_1\tilde{x}_{i1}-\beta_2\tilde{x}_{i2})\tilde{x}_{i2}\end{aligned}$$

som hvis vi setter lik null gir likningssystemet

$$\begin{aligned}\beta_1[\sum_{i=1}^{n}\tilde{x}_{i1}^2+\lambda_1]+\beta_2\sum_{i=1}^{n}\tilde{x}_{i2}\tilde{x}_{i1} &= \sum_{i=1}^{n}y_i\tilde{x}_{i1}\\ \beta_1\sum_{i=1}^{n}\tilde{x}_{i2}\tilde{x}_{i1}+\beta_2[\sum_{i=1}^{n}\tilde{x}_{i2}^2+\lambda_2] &= \sum_{i=1}^{n}y_i\tilde{x}_{i2}\end{aligned}$$

Hvis $\sum_i(x_{i1}-\bar{x}_1)(x_{i2}-\bar{x}_2) = \sum_i\tilde{x}_{i1}\tilde{x}_{i2} = 0$, forenkler likningssystemet seg til

$$\begin{aligned}\beta_1[\sum_{i=1}^{n}\tilde{x}_{i1}^2+\lambda_1] &= \sum_{i=1}^{n}(y_i-\bar{y})\tilde{x}_{i1}\\ \beta_2[\sum_{i=1}^{n}\tilde{x}_{i2}^2+\lambda_2] &= \sum_{i=1}^{n}(y_i-\bar{y})\tilde{x}_{i2}\end{aligned}$$

som gir løsningen

$$\begin{aligned}\hat{\beta}_1 &= \frac{\sum_{i=1}^{n}y_i\tilde{x}_{i1}}{\sum_{i=1}^{n}\tilde{x}_{i1}^2+\lambda_1}\\ \hat{\beta}_2 &= \frac{\sum_{i=1}^{n}y_i\tilde{x}_{i2}}{\sum_{i=1}^{n}\tilde{x}_{i2}^2+\lambda_2}\end{aligned}$$

Vi får da siden $\beta_0 = \tilde{\beta}_0 - \beta_1\bar{x}_1 - \beta_2\bar{x}_2$ at

$$\hat{\beta}_0 = \bar{y} - \frac{\sum_{i=1}^{n}y_i\tilde{x}_{i1}}{\sum_{i=1}^{n}\tilde{x}_{i1}^2+\lambda_1}\bar{x}_1 - \frac{\sum_{i=1}^{n}y_i\tilde{x}_{i2}}{\sum_{i=1}^{n}\tilde{x}_{i2}^2+\lambda_2}\bar{x}_2$$

(Fortsettes på side 6.)

(d) Den første metoden svarer til minste kvadraters metode. Den andre svarer til vanlig Ridge regresjon.

Det ser ut som det er viktigst å straffe β_1 (svarende til den minst signifikante variabel), og vanlig Ridge ser da ut til å legge mest vekt på denne variabelen. Valg av λ blir derfor mest påvirket av hvor mye vi trenger å straffe β_1, og g dermed at $\lambda_1 \approx \lambda$ (faktisk lik i dette tilfellet).

Hvis vi skulle bruke denne metoden på mange forklaringsvariable for vi iallefall to problemer:

- Et numerisk problem ved at vi må minimere med hensyn på mange λ_j'er.
- Et statistisk problem ved at vi kan lett få overtilpasning når vi nå innfører menge nye tuningparametre i metoden.