# UNIVERSITETET I OSLO

## Det matematisk-naturvitenskapelige fakultet

| | |
|---|---|
| Eksamen i: | STK2100 — - FASIT |
| Eksamensdag: | Torsdag 14. juni 2018. |
| Tid for eksamen: | 14.30 – 18.30. |
| Oppgavesettet er på 4 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 11 endenoder som gir 11 $c_m$ 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 $11 + 2 * 12 = 35$ parametre.

Dette gir en AIC verdi på

$$AIC = -2 * (-279.5) + 2 * 35 = 629.0$$

dvs noe bedre enn vi fikk med de tidligere modeller.

(g) 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 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.

(h) Trær kan ofte gi overtilpasning. Et alternativ da er å bygge et mindre tre. Det kan imidlertid være interaksjoner som er viktig

# 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

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

*(Fortsettes på side 3.)*

$\tilde{\beta}_0$ 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

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

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

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

*(Fortsettes på side 4.)*

som gir løsningen

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

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

Siden `Hits` er mye mer signifikant enn `PutPuts`, er det rimelig at det legges mest vekt på denne, og 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 $\lambda_j$'er.
- Et statistisk problem ved at vi kan lett få overtilpasning når vi nå innfører menge nye tuningparametre i metoden.