UNIVERSITETET I OSLO

Det matematisk-naturvitenskapelige fakultet

Eksamen i:	STK2100 — - FASIT
Eksamensdag:	Torsdag 15. juni 2017.
Tid for eksamen:	09.00 – 13.00.
Oppgavesettet er på 5 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) Et alternativ kunne være å velge treningssettet på en mer systematisk måte, f.eks ved å ta de første observ asjoner til dette. Det kan imidlertid være systematikk i hvordan data er samlet inn, og da vil ikke treningssettet være representativt for nye data.

P-verdien for AGE er ganske stor, noe som sier at det ikke er grunnlag for å påstå at et (lineær) bidrag av denne variabelen er viktig. Da vi ønsker enkle modeller, tilsier dette at vi bør fjerne denne. Det minker også sjansen for overtilpasning.

(b) Vi kan bruke AIC kriteriet for å velge mellom modeller. Dette gir

$$\text{AIC}_1 = -2 * (-742.34) + 2 * 20 = 1524.68$$
$$\text{AIC}_2 = -2 * (-742.90) + 2 * 19 = 1523.80$$

Da AIC verdien for den nye modellen er lavere, velger vi denne.

Alternativt kan en argumentere direkte fra den første regresjonstabellen at det ikke er grunn til å forkaste hypotesen om at regresjonskoeffisientent tilhørende AGE er signifikant forskjellig fra 0 og dermed velge den alternative modell.

Vi ser at noen av P-verdiene tilhørende den kategoriske variabel RAD er ganske store. En mulig forenkling av modellen kunne være å slå sammen noen av kategoriene.

(c) GAM er basert på glattingssplines som gir prediksjoner av typen $\hat{y}_i = \sum_{j=1}^{n} S_{ij} y_j$, dvs lineære kombinasjoner av observasjonene.

(Fortsettes på side 2.)

Frihetsgrader er definert gjennom $\sum_{i=1}^{n} S_{ii}$. Vi kan da beregne en AIC verdi for denne modellen:

$$\text{AIC}_{gam} = -2 * (-615.79) + 2 * 46.12 = 1323.82$$

som er adskillig lavere enn de tidligere modeller. Denne modellen er altså å foretrekke. Basert på plottene er dette også rimelig da flere av variablene viser ikke-lineære sammenhenger.

Vi får også en god reduksjon i estimert feilrate på testdata som er rimelig når vi får såpass stor forbedring i tilpasning.

(d) Når vi har kombinasjoner av oppsplittinger med ulike forklaringsvariable, får vi indirekte interaksjoner.

Dette er et produkt av Gaussiske sannsynlighetstettheter som da baserer seg på antagelser at hver observasjon er Gaussisk og at observasjonene er uavhengige av hverandre. Vi antar også at variasjonen er den samme for alle observasjoner.

For klassifikasjonstrær så antar vi at vi har samme modell innenfor et område R_m. Dermed blir $\mu_i = c_m$ for $\boldsymbol{x}_i \in R_m$.

For en modell med 9 endenoder har vi 9 c_m verdier og σ^2 som må estimeres. I tillegg har vi 8 oppsplittinger og hver av disse krever 2 parametre (hvilken variabel som skal brukes til oppsplitting, og hvilke terskelverdi). Dermed får vi $9 + 1 + 2 * 8 = 26$ parametre.

AIC for denne modellen blir da

$$\text{AIC}_{tree} = -2 * (-697 - 49) + 2 * 26 = 1446.98$$

Dette er noe bedre enn de lineære modeller, men noe dårligere enn GAM.

(e) I Bagging så bruker vi ulike bootstrap utvalg til å lage mange regresjonstrær og så tar vi gjennomsnitt av prediksjonene vi får fra hvert tre for en endelig prediksjon.

For Random Forrest gjør vi noe liknende, men for å gjøre trærne mer ukorrelerte, bruker vi kun et (tilfeldig) utvalg av forklaringsvariablene for hver oppsplitting.

For Boosting, så genereres trærne sekvensielt der vi tilpasser residualene fra tidligere modell med et nytt tre. For å hindre overtilpasning blir de tilpassede trær skalert ned med en læringsparameter λ.

Alle metodene har potensiale for å forbedre regresjonstrær da de kan minke variabiliteten (som kan være stor for trør) uten at forventningsskjevhet behøver å øke mye.

I dette tilfellet ser alle metodene ut til å gi forbedringer, uten at det er mye forskjell mellom de ulike metodene.

(Fortsettes på side 3.)

Interessant nok gir det også forbedringer i forhold til GAM. Dette kan skyldes at interaksjoner er viktig, noe GAM modellen ikke har med, men som trær indirekte gir muligheter for.

Oppgave 2

(a) K-nærmeste nabo metoden går ut på å finne de K nærmeste punkter $\boldsymbol{x}_i$ til $\boldsymbol{x}_0$ så estimere $P(Y = g|\boldsymbol{X} = \boldsymbol{x}_0)$ ved $\frac{1}{K}\sum_{\in\mathcal{N}_0} I(y_i = g)$, evt ved direkte klassifiksering å klassifisere til den klasse som opptrer oftest bland de nærmeste punktene.

Det er en enkel metode, fleksibel med liten forventningsskjevhet (hvis K er liten) men kan ha stor varians. K kan velges ved for eksempel kryss-validering.

(b) Siden små K gir veldig lokal tilpasning, mens stor K gir mer global struktur, vil $K = 1$ svare til det midterste plottet, $K = 10$ til det til venstre og $K = 100$ det til høyre.

$K = 100$ ser ut til å gi for restriktiv modell mens $K = 1$ gir for mange lokale tilpasninger. Jeg ville derfor foretrekke $K = 10$.

(c) Vi har at

$$E[L(Y, \widehat{Y})] = E[[L(Y, \widehat{Y})]|\boldsymbol{x}]$$

og minimering av forventet feil kan gjøres ved å minimere forventet feil for hver $\boldsymbol{x}$. Videre har vi at

$$\begin{aligned} E[[L(Y, \widehat{Y})]|\boldsymbol{x}] =& E[I(Y \neq \widehat{Y})|\boldsymbol{x}] \\ =& \sum_{g=1}^{G} I(\widehat{Y} = g)(1 - \Pr(Y = g|\boldsymbol{x})) \\ =& 1 - \sum_{g=1}^{G} I(\widehat{Y} = g)\Pr(Y = g|\boldsymbol{x}). \end{aligned}$$

Vi ser da at for å få dette minst mulig, må vi ha $\sum_{g=1}^{G} I(\widehat{Y} = g)\Pr(Y = g|\boldsymbol{x})$ størst mulig. Dette får vi til ved å velge $\widehat{Y} = \arg\max_g \Pr(Y = g|\boldsymbol{x})$.

For å kunne beregne $\widehat{Y}$, trenger vi da $\Pr(Y = g|\boldsymbol{x}) = E[I(Y = g)|\boldsymbol{x}] = f_g(\boldsymbol{x})$ for alle g.

(d) Vi kan innføre dummy variable $y_{ig} = I(y_i = g)$ og så tilpasse modeller $E[Y_{ig}] = f_g(\boldsymbol{x}_i)$ med vanlige regresjonsmetoder der nå Y_{ig} behandles som en numerisk respons.

Deretter kan en klassifisere til den klassen som estimerer høyest respons.

(Fortsettes på side 4.)

Merk at logistisk regresjon *ikke* er en regresjonsmetode, men en klassifikasjonsmetode!

(e) En mulighet er å bruke $\frac{1}{n}\sum_{i=1}^{n} I(\hat{y}_i = y_i)$. Dette vil imidlertid gi for optimistisk anslag da vi bruker samme data til trening og testing.

Alternative metoder er

- Dele data i to, bruke en del til trening og en til testing. Dette vil gi en forventningsrett estimator, men vil gi mindre data til trening.
- Kryss-validering: Her deler en data opp i K grupper. $K-1$ grupper blir brukt til trening mens den siste gruppen blir brukt til validering. Ved å utføre K slike tilpasninger og valideringer, får vi utnyttet all data til validering mens vi får brukt en andel $(K-1)/K$ av data til trening. En svakhet her er at vi fremdeles ikke bruker all data til trening og at vi validerer ulike modeller som ingen er lik den vi endelig ville bruke.
- (AIC: Dette er basert på å korrigere for underestimering av prediksjonsfeil gjennom å legge inn et straffeledd. Fordelen er at alle data kan utnyttes. Ulempen her er at den baserer seg mye mer på modell-antagelsene, som da må spesifiseres. Denne metoden har vi bare såvidt nevnt i denne sammenhengen)

(f) Alle forventninger nedenfor er betinget på $\boldsymbol{x}_0$.

$$\begin{aligned}
&E[(f_g(\boldsymbol{x}_0) - \widehat{f}_g(\boldsymbol{x}_0))^2] \\
&\quad = E[(f_g(\boldsymbol{x}_0) - E[\widehat{f}_g(\boldsymbol{x}_0)] + E[\widehat{f}_g(\boldsymbol{x}_0)] - \widehat{f}_g(\boldsymbol{x}_0))^2] \\
&\quad = E[(f_g(\boldsymbol{x}_0) - E[\widehat{f}_g(\boldsymbol{x}_0)])^2] + E[(E[\widehat{f}_g(\boldsymbol{x}_0)] - \widehat{f}_g(\boldsymbol{x}_0)])^2] + \\
&\qquad 2E[(f_g(\boldsymbol{x}_0) - E[\widehat{f}_g(\boldsymbol{x}_0)])(E[\widehat{f}_g(\boldsymbol{x}_0)] - \widehat{f}_g(\boldsymbol{x}_0))] \\
&\quad = (f_g(\boldsymbol{x}_0) - E[\widehat{f}_g(\boldsymbol{x}_0)])^2 + \mathrm{Var}[\widehat{f}_g(\boldsymbol{x}_0)]
\end{aligned}$$

Det første leddet er forventningsskjevhet. Det andre leddet er varians til $\hat{f}$. I en typisk situasjon må vi avveie disse leddene mot hverandre.

Alternativt

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

(g) For en restriktiv estimator vil vi kunne få en stor forventningsskjevhet men liten varians, mens det blir omvendt for en mer fleksibel estimator. For valg av de ulike estimatorer vil det da være en avveining mellom forventningsskjevhet og varians.

(Fortsettes på side 5.)

Oppgave 3

(a) Da vi har samme straff λ på alle β_j-ene er det hensiktsmessig at disse er på samme skala, og det kan vi få til ved å skalere variablene.

(b) Utledning:

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

$$\frac{\partial}{\partial \beta_l} h(\boldsymbol{\beta}_l) = -2 \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right) x_{il} + 2\hat{\beta}_l, l \geq 1$$

som gir normallikninger

$$\hat{\beta}_0 = \frac{1}{n} \sum_{i=1}^{n} y_i = \bar{y}$$

$$\sum_{j=1}^{p} \hat{\beta}_j \sum_{i=1}^{n} x_{ij} x_{il} + \beta_l = \sum_{i=1}^{n} (y_i - \bar{y}) x_{il} = \sum_{i=1}^{n} y_i x_{il}, \quad l = 1, ..., p$$

som på matriseform kan skrives ($\boldsymbol{X}$ har nå *ikke* 1 i første kollonne og $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)$)

$$(\boldsymbol{X}^T \boldsymbol{X} + \lambda \boldsymbol{I}) \hat{\boldsymbol{\beta}} = \boldsymbol{X}^T \boldsymbol{y}$$

Dette gir da $\hat{\boldsymbol{\beta}}^{ridge} = \hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T \boldsymbol{X} + \lambda \boldsymbol{I})^{-1} \boldsymbol{X}^T \boldsymbol{y}$.

(c) Vi får at

$$\begin{aligned}
E[\hat{\boldsymbol{\beta}}^{ridge}] =& (\boldsymbol{X}^T \boldsymbol{X} + \lambda \boldsymbol{I})^{-1} \boldsymbol{X}^T [\beta_0 \mathbf{1} + \boldsymbol{X} \boldsymbol{\beta}] \\
=& (\boldsymbol{I} + \lambda \boldsymbol{I})^{-1} \boldsymbol{X}^T \boldsymbol{X} \boldsymbol{\beta} \\
=& \frac{1}{1+\lambda} \boldsymbol{\beta} \\
V[\hat{\boldsymbol{\beta}}^{ridge}] =& V[(\boldsymbol{X}^T \boldsymbol{X} + \lambda \boldsymbol{I})^{-1} \boldsymbol{X}^T [\beta_0 \mathbf{1} + \boldsymbol{X} \boldsymbol{\beta} + \boldsymbol{\varepsilon}] \\
=& V[(\boldsymbol{X}^T \boldsymbol{X} + \lambda \boldsymbol{I})^{-1} \boldsymbol{X}^T \boldsymbol{\varepsilon} \\
=& (\boldsymbol{X}^T \boldsymbol{X} + \lambda \boldsymbol{I})^{-1} \boldsymbol{X}^T \sigma^2 \boldsymbol{I} \boldsymbol{X} (\boldsymbol{X}^T \boldsymbol{X} + \lambda \boldsymbol{I})^{-1} \\
=& \frac{1}{(1+\lambda)^2} \sigma^2 \boldsymbol{I}
\end{aligned}$$

Vi ser derfor at ved valg av $\lambda > 0$ får vi en forventningsskjevhet, men at vi får en reduksjon i varians.