Prøveeksamen STK2100 (fasit) - vår 2018

Geir Storvik

Vår 2018

Oppgave 1

(a) Vi har at

$$\begin{aligned}\boldsymbol{E} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}} =& \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{\beta} + \varepsilon) \\ =& [\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T]\boldsymbol{\varepsilon}\end{aligned}$$

Dette gir direkte at $E[\boldsymbol{E}] = \boldsymbol{0}$. Vi får at kovariansmatrisen til $\boldsymbol{E}$ blir

$$\begin{aligned}\text{Var}(\boldsymbol{E}) =& [\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T]\sigma^2\boldsymbol{I}[\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T]^T \\ =& [\boldsymbol{I} - 2\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T + \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T]\sigma^2 \\ =& [\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T]\sigma^2\end{aligned}$$

(b) Vi har at

$$\begin{aligned}E[\text{RSS}] =& E[\boldsymbol{E}^T\boldsymbol{E}] \\ =& E[\text{trase}(\boldsymbol{E}\boldsymbol{E}^T)] \\ =& \text{trase}(E[\boldsymbol{E}\boldsymbol{E}^T]) \\ =& \text{trase}(\text{Var}(\boldsymbol{E})) \\ =& \text{trase}([\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T]\sigma^2) \\ =& \sigma^2\text{trase}(\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T) \\ =& \sigma^2(\text{trase}(\boldsymbol{I}) - \text{trase}(\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T) \\ =& \sigma^2(n - \text{trase}(\boldsymbol{I}_p)) = \sigma^2(n-p)\end{aligned}$$

(c) Vi har at

$$\begin{aligned}\text{Cov}(\hat{\boldsymbol{y}}, \boldsymbol{E}) =& \text{Cov}(\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, [\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T]\boldsymbol{\varepsilon}) \\ =& \text{Cov}(\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\varepsilon}, [\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T]\boldsymbol{\varepsilon}) \\ =& \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\text{Cov}(\boldsymbol{\varepsilon}, \boldsymbol{\varepsilon})[\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T] \\ =& \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\sigma^2\boldsymbol{I}[\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T] \\ =& \sigma^2[\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T] \\ =& \sigma^2[\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T] = \boldsymbol{0}\end{aligned}$$

Dette betyr at feilen vi gjør er uavhengig av prediksjonen. Geometrisk betyr dette at $\hat{\boldsymbol{y}}$ er ortogonal på $\boldsymbol{E}$.

Oppgave 2

(a) Vi antar ε har forventning 0. Vi antar videre at Y er en ny observasjon som ikke er blitt brukt til å konstruere $\widehat{Y}$. Vi har først at

$$E[(Y-\hat{Y})^2] = E[E[(Y-\hat{Y})^2|\boldsymbol{x}]]$$

og det er nok å minimere $E[(Y-\hat{Y})^2]\boldsymbol{x}]$ for alle $\boldsymbol{x}$.

Videre er

$$\begin{aligned}
&E[(Y-\hat{Y})^2|\boldsymbol{x}] \\
&\quad = E[(Y-E[Y|\boldsymbol{x}]+E[Y|\boldsymbol{x}]-\widehat{Y})^2|\boldsymbol{x}] \\
&\quad = E[(Y-E[Y|\boldsymbol{x}])^2|\boldsymbol{x}] + E[(E[Y|\boldsymbol{x}]-\widehat{Y})^2|\boldsymbol{x}] + 2E[(Y-E[Y|\boldsymbol{x}])(E[Y|\boldsymbol{x}]-\widehat{Y})|\boldsymbol{x}] \\
&\quad = \mathrm{Var}(Y|\boldsymbol{x}) + (E[Y|\boldsymbol{x}]-\widehat{Y}(\boldsymbol{x}))^2
\end{aligned}$$

siden $E[Y|\boldsymbol{x}]$ og $\hat{Y}$ er konstanter gitt $\boldsymbol{x}$. Det første leddet avhenger ikke av $\widehat{Y}$ og det andre leddet blir lik 0 hvis $\widehat{Y}(\boldsymbol{x}) = E[Y|\boldsymbol{x}]$.

(b) La nå alle forventninger være betinget på $\boldsymbol{x}_0$. Anta videre at $\hat{f}(\boldsymbol{x}_0)$ er basert på treningsdata som ikke inkluderer $(\boldsymbol{x}_0, Y)$ slik at den nye ε_0 er uavhengig av $\hat{f}(\boldsymbol{x}_0)$. Da er

$$\begin{aligned}
&E[(Y-\widehat{f}(\boldsymbol{x}_0))^2] \\
&\quad = E[(f(\boldsymbol{x}_0)+\varepsilon_0 - E[\widehat{f}(\boldsymbol{x}_0)] + E[\widehat{f}(\boldsymbol{x}_0)] - \widehat{f}(\boldsymbol{x}_0))^2] \\
&\quad = E[(f(\boldsymbol{x}_0) - E[\widehat{f}(\boldsymbol{x}_0)])^2] + E[(E[\widehat{f}(\boldsymbol{x}_0)] - \widehat{f}(\boldsymbol{x}_0)])^2] + \mathrm{Var}(\varepsilon_0) + \\
&\qquad 2E[(f(\boldsymbol{x}_0) - E[\widehat{f}(\boldsymbol{x}_0)])(E[\widehat{f}(\boldsymbol{x}_0)] - \widehat{f}(\boldsymbol{x}_0))] + 2E[(f(\boldsymbol{x}_0) - E[\widehat{f}(\boldsymbol{x}_0))\varepsilon_0] + \\
&\qquad 2E[\varepsilon_0(E[\widehat{f}(\boldsymbol{x}_0)] - \widehat{f}(\boldsymbol{x}_0))] \\
&\quad = (f(\boldsymbol{x}_0) - E[\widehat{f}(\boldsymbol{x}_0)])^2 + \mathrm{Var}[\widehat{f}(\boldsymbol{x}_0)] + \mathrm{Var}(\varepsilon_0)
\end{aligned}$$

der vi i nest siste likhet har brukt at ε_0 er uavhengig av $\hat{f}$.

Det første leddet er forventningsskjevhet. Det andre leddet er varians til $\hat{f}$ mens det tredje leddet er varians til støyleddet.

(c) Det siste leddet kan vi ikke gjøre noe med. 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.

(*d*) En metode kunne være å bruke $\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$, dvs basere tilpasning på treningsdata. Dette vil imidlertid gi en underestimering av forventet feil, da tilpasningen er basert på å minimere dette uttrykket (referer oppgave 1 (*b*)).

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.

Oppgave 3

(*a*) Det er vanskelig å vurdere om dette er bra eller dårlig. Vil avhenge av hvor viktig det er med presisjon. Videre vil dette forvirringsmatrisen gi et for optimistisk svar.

Imidlertid gir regresjonstabellen inntrykk av at det er mange variable som ikke er så viktige å ha med slik at en bør utføre et modellvalg.

(*b*) Den maksimale likelihood verdien vil alltid være høyere for en modell med flere variable inkludert siden vi da kan maksimere over et større rom.

Kan bruke AIC kriteriet. Det gir

$$\begin{aligned}\text{AIC}_1 &= -2*(-97.83) + 2*(10) = 215.66\\ \text{AIC}_2 &= -2*(-98.71) + 2*(6) = 209.42\end{aligned}$$

Den andre modellen har nesten like god likelihood-verdi og langt færre parametre, noe som gjør at AIC verdien for den nye modellen er lavere og dermed å foretrekke.

Ser vi på forvirringsmatrisen for den nye modellen, så har den én mer feil, men gitt at dette er på treningsdata, så kan dette skyldes overtilpasning i den første modellen.

(*c*) P-verdiene gir P-verdier for testing av tilhørende hypotese $H_{0j} : \beta_j = 0$. Vi ser at mens den første modellen gir mange P-verdier som er store og dermed ikke gir grunnlag for å forkaste H_{0j} for flere j'er, så har den andre modellen alle små P-verdier.

En mulig årsak til at P-verdiene endrer seg i den andre modellen er at de tilhørende forklaringsvariable er korrelerte med noen av de variable vi har tatt bort fra modellen.

(*d*) 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 $\boldsymbol{x}_i \in R_m$.

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

Vi har 5 endenoder som gir 5 c_m parametre. I mange sitjasjoner bruker en dette som antall parametre.

I tillegg har vi imidlertid 4 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 $5 + 2 * 4 = 13$ parametre.

Dette gir en AIC verdi på

$$AIC = -2 * (-90.21) + 2 * 13 = 206.43$$

dvs noe bedre enn vi fikk med de tidligere modeller.

(*f*) Vi ser først at feilratene nå er større enn de var tidligere. Dette skyldes at vi ikke tok hensyn til overtilpasning tidligere.

Vi ser nå at faktisk logistisk regresjon gjør det bedre enn klassifikasjonstrær mhp feilrate. Vi vil dermed foretrekke logistisk regresjon i dette tilfellet.

En kunne muligens gjøre dette enda mer komplisert gjennom å si at hvilken node som blir splittet opp også er en parameter. Vi går imidlertid ikke inn på det her.

(*g*) Med Bagging så lager vi mange klassifikatorer basert på bootstrap utvalg av de opprinnelige data. For hver klassifikator får vi en klassifisering og vi kan da kombinere klassifiseringene ved å gjøre endelig klassifisering til den klasse som opptrer oftest.

Når vi lager bootstrap utvalg, vil noen observasjoner ikke være med i treningssettet. Disse kan da brukes til validering/testing. Når vi så ser over alle bootstrap utvalg, vil observasjonene være i et testsett flere ganger. Vi kan da gjøre en endelig klassifisering til den klasse som opptrer oftest.

Vi ser at for logistisk regresjon så er det ingen endring i forvirringsmatrisen mens for klassifikasjonstreet blir det en liten forbedring. Dette skyldes at logistisk regresjon i utgangspunktet har liten variabilitet og det hjelper da ikke så mye å ta gjennomsnitt over mange klassifikatorer.

Oppgave 4

(*a*) Vi har potensielt ikke-kontinuitet i punktet $x = c$. Kontinuitet medfører at

$$\beta_{0,1} + \beta_{1,1}c + \beta_{2,1}c^2 = \beta_{0,2} + \beta_{1,2}c + \beta_{2,2}c^2$$

mens kontinuerlige deriverte medfører at

$$\beta_{1,1} + 2\beta_{2,1}c = \beta_{1,2} + 2\beta_{2,2}c.$$

Vi har i utgangspunktet 6 parametre, men med 2 begrensninger, ender vi opp med 4 *frie* parametre.

(b) Innenfor intervallet $(-\infty, c)$ er

$$g(x) = \theta_0 + \theta_1 x + \theta_2 x^2$$

mens innenfor intervallet $[c, \infty)$ er

$$g(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 (x - c)^2$$

som begge er kvadratiske funksjon.

Videre er $(x - c)_+^2$ kontinuerlig i c som medfører at $g(x)$ også er kontinuerlig.

Vi har at for $x \neq c$ er

$$g'(x) = \theta_1 + 2\theta_2 x + 2\theta_3 (x - c)_+$$

der også $(x - c)_+$ er kontinuerlig i c slik at den deriverte også er kontinuerlig.

(c) I intervallet $(-\infty, c)$ må vi ha

$$\beta_{0,1} + \beta_{1,1} x + \beta_{2,1} x^2 = \theta_0 + \theta_1 x + \theta_2 x^2$$

som medfører at $\theta_0 = \beta_{0,1}, \theta_1 = \beta_{1,1}, \theta_2 = \beta_{2,1}$.

I intervallet $[c, \infty)$ må vi ha

$$\begin{aligned} \beta_{0,2} + \beta_{1,2} x + \beta_{2,2} x^2 =& \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 (x - c)^2 \\ =& \beta_{0,1} + c^2 \theta_3 + (\beta_{1,1} - 2c\theta_3) x + (\beta_{2,1} + \theta_3) x^2 \end{aligned}$$

som medfører 3 krav til θ_3:

$$\begin{aligned} \theta_3 =& c^{-2} (\beta_{0,2} - \beta_{0,1}) \\ \theta_3 =& \tfrac{1}{2c} (\beta_{1,1} - \beta_{1,2}) \\ \theta_3 =& \beta_{2,2} - \beta_{2,1} \end{aligned}$$

men hvis vi bruker begrensningene fra (a) har vi at

$$\begin{aligned} \tfrac{1}{2c} (\beta_{1,1} - \beta_{1,2}) =& \tfrac{1}{2c} (2\beta_{2,2} c - 2\beta_{2,1} c) = \beta_{2,2} - \beta_{2,1} \\ c^{-2} (\beta_{0,2} - \beta_{0,1}) =& c^{-2} [\beta_{1,1} c + \beta_{2,1} c^2 - (\beta_{1,2} c + \beta_{2,2} c^2)] \\ =& \beta_{2,1} - \beta_{2,2} + c^{-1} (\beta_{1,1} - \beta_{1,2}) \\ =& \beta_{2,1} - \beta_{2,2} + 2(\beta_{2,2} - \beta_{2,1}) = \beta_{2,2} - \beta_{2,1} \end{aligned}$$

som viser at det egentlig kun er et krav.

(*d*) Vi har potensielt ikke-kontinuitet i punktene $c_1, ..., c_{M-1}$. Kontinuitet medfører at

$$\beta_{0,m} + \beta_{1,m}c_m + \beta_{2,m}c_m^2 = \beta_{0,m+1} + \beta_{1,m+1}c_m + \beta_{2,m+1}c_m^2$$

mens kontinuerlige deriverte medfører at

$$\beta_{1,m} + 2\beta_{2,m}c_m = \beta_{1,m+1} + 2\beta_{2,m+1}c_m.$$

Vi har i utgangspunktet 3M parametre, men med 2(M-1) begrensninger, ender vi opp med $3M - 2(M-1) = M + 2$ frie parametre.

(*e*) Siden vi kan skrive modellen som en lineær kombinasjon av basisfunksjoner, kan de frie parametrene estimeres med vanlig minste kvadraters metode.

Oppgave 5

(*a*) La $H(\beta_0, \beta_1) = \sum_{i=1}^n K(x_i, x_0)(y_i - \beta_0 - \beta_1 x_i)^2$ (der vi for enkelthetsskyld her ikke har tatt med at β_0 og β_1 er avhengige av x_0). Da er

$$\frac{\partial}{\partial \beta_0} H(\beta_0, \beta_1) = -2 \sum_{i=1}^n K(x_i, x_0)(y_i - \beta_0 - \beta_1 x_i)$$
$$\frac{\partial}{\partial \beta_1} H(\beta_0, \beta_1) = -2 \sum_{i=1}^n K(x_i, x_0)(y_i - \beta_0 - \beta_1 x_i)x_i$$

som når vi setter lik 0 gir likningssystemet

$$\sum_{i=1}^n K(x_i, x_0)\hat{\beta}_0 + \sum_{i=1}^n K(x_i, x_0)x_i\hat{\beta}_1 = \sum_{i=1}^n K(x_i, x_0)y_i$$
$$\sum_{i=1}^n K(x_i, x_0)x_i\hat{\beta}_0 + \sum_{i=1}^n K(x_i, x_0)x_i^2\hat{\beta}_1 = \sum_{i=1}^n K(x_i, x_0)x_i y_i$$

Dette kan skrives på matriseform

$$\boldsymbol{M}\hat{\boldsymbol{\beta}} = \boldsymbol{K}\boldsymbol{y}$$

som gir en løsning $\hat{\boldsymbol{\beta}} = \boldsymbol{M}^{-1}\boldsymbol{K}\boldsymbol{y}$ som er en lineær kombinasjon av y_i-ene. Mer at begge matrisene $\boldsymbol{M}$ og $\boldsymbol{K}$ avhenger av x_0.

For $d = 2$ får vi en tilsvarende situasjon, men da er $\boldsymbol{M}$ en 3×3 matrise og $\boldsymbol{K}$ en $3 \times n$ matrise.

Da estimatet tilpasser (lokalt) en rett linje, vil estimatet kun være forventningsrett hvis den sanne f er (lokalt) lineær. I motsatt fall vil vi få en forventningsskjevhet.

(*b*) Vi har at $\sum_{i=1}^{n} S_{ii}$ angir frihetsgrader i slike modeller. Frihetsgrader er et mål på fleksibiliteten, og for å sammenlikne ulike metoder er det rimelig å sette frihetsgradene omtrent like.

Fra plottet kan det se ut som den tilpassede linjen består av to lineære segmenter. Det er derfor rimelig at $d = 1$ gir en god nok tilpasning.

(*c*) Naturlig splines er basert på et 3. grads polynom og kontinuitet i opptil 2. deriverte. Det er derfor rimelig at den estimerte funksjon er glattere.

Kun basert på prediksjonsfeil er lokal regresjon med $d = 1$ å foretrekke. Dog kan det i noen tilfeller være at en i utgangspunktet tror funksjonen skal være rimelig glatt (og kanskje også monoton). Slike kriterier kan gjøre at en vil godta litt større prediksjonsfeil.

(*d*) Dette er innenfor klassen av (generaliserte) additive modeller. De to estimatene ser veldig like ut. Dog er skalaen på y-aksen ganske forskjellig. Dette indikerer at når temperatur er med, så forklarer den noe av det samme som vind. Dette antyder at de to forklaringsvariable er korrelerte, noe som er rimelig.

(*e*) En kan starte med å sette $\hat{f}_2(x_2) = 0$ og så finne et estimat på $f_1(x_1)$. Deretter kan en beregne residualer $y - \hat{f}_1(x_1)$ og tilpasse $f_2(x_2)$ til denne. Man kan så skifte mellom disse til konvergens.

Oppgave 6

- Minste kvadraters metode er egnet for parallelisering da vi kan utføre de nødvendig beregninger oppdelt i ulike grupper av data. Slik gruppering vil også kunne være minne-besparende.

- Bagging kan enkelt parallelliseres da hvert bootstrap utvalg kan kjøres på hver sin prosessor. Dette vil være noe minne-besparende, men ikke mye siden ca 2/3 av dataene hele tiden vil være med.

- Kryss-validering kan enkelt parallelliseres da hvert kryss-validert sett kan valideres uavhengig av de øvrige kryss-validerte sett. Her bruker man nesten hele datasettet så det vil ikke være så veldig minne-besparende.

Oppgave 7

(*a*) Vi har at

$$\Pr(S|V) = \frac{\Pr(S)\Pr(V|S)}{\Pr(S)\Pr(V|S) + \Pr(R)\Pr(V|R)} = \frac{(1-r)q}{(1-r)q + rp}$$

og

$$\begin{aligned}
\Pr(S|V) &> 0.5 \\
&\Updownarrow \\
\frac{\Pr(S|V)}{\Pr(R|V)} &> 1 \\
&\Updownarrow \\
\frac{(1-r)q}{rp} &> 1 \\
&\Updownarrow \\
\frac{q}{p} &> \frac{r}{1-r}
\end{aligned}$$

dvs hvis vi observerer V og $\frac{q}{p} > \frac{r}{1-r}$ så klassifiserer vi til spam.

Tilsvarende:

$$\Pr(S|V^c) = \frac{\Pr(S)\Pr(V^c|S)}{\Pr(S)\Pr(V^c|S) + \Pr(R)\Pr(V^c|R)} = \frac{(1-r)(1-q)}{(1-r)(1-q) + r(1-p)}$$

og

$$\begin{aligned}
\Pr(S|V^c) &> 0.5 \\
&\Updownarrow \\
\frac{\Pr(S|V^c)}{\Pr(R|V^c)} &> 1 \\
&\Updownarrow \\
\frac{(1-r)(1-q)}{r(1-p)} &> 1 \\
&\Updownarrow \\
\frac{1-q}{1-p} &> \frac{r}{1-r}
\end{aligned}$$

dvs hvis vi observerer V^c og $\frac{1-q}{1-p} > \frac{r}{1-r}$ så klassifiserer vi til spam.

(*b*) Vi kan innføre c_R til å være et mål på tap for å klassifisere galt hvis mail er reell mens c_S er tilsvarende et mål på tap for å klassifisere galt hvis mail er spam. Typisk vil da $c_R > c_S$.

Hvis vi definerer x til å våre observasjoner (V eller V^c), blir forventet tap da

$$E[L] = E[E[L|x]]$$

der

$$\begin{aligned}E[L|x] =& c_R \Pr(Y = R, \widehat{Y} = S|x) + c_S \Pr(Y = S, \widehat{Y} = R|x) \\ =& c_R[1 - \Pr(Y = S|x)]I(\widehat{Y} = S) + c_S 1 \Pr(Y = S|x)I(\widehat{Y} = R).\end{aligned}$$

Dette tilsier at for å minimere tap bør en sette $\widehat{Y} = S$ hvis $c_R[1 - \Pr(Y = s|x)] < c_S \Pr(Y = S|x)$ som er ekvivalent med at $\Pr(Y = S|x) > c_R/(c_R + c_S)$.

(c) Bygger på uavhengighet mellom forekomster av ord.

$$\begin{aligned}\Pr(S|\boldsymbol{V}) =& \frac{\Pr(S)\Pr(\boldsymbol{V}|S)}{\Pr(S)\Pr(\boldsymbol{V}|S) + \Pr(R)\Pr(\boldsymbol{V}|R)} \\ =& \frac{(1-r)\prod_{m=1}^{M} q_m^{V_m}(1-q_m)^{1-V_m}}{(1-r)\prod_{m=1}^{M} q_m^{V_m}(1-q_m)^{1-V_m} + r\prod_{m=1}^{M} p_m^{V_m}(1-p_m)^{1-V_m}}\end{aligned}$$

$$\begin{aligned}\frac{\Pr(S|V)}{\Pr(R|V)} >& 1 \\ \Updownarrow& \\ \frac{(1-r)\prod_{m=1}^{M} q_m^{V_m}(1-q_m)^{1-V_m}}{r\prod_{m=1}^{M} p_m^{V_m}(1-p_m)^{1-V_m}} >& 1 \\ \Updownarrow& \\ \frac{\prod_{m=1}^{M} q_m^{V_m}(1-q_m)^{1-V_m}}{\prod_{m=1}^{M} p_m^{V_m}(1-p_m)^{1-V_m}} >& \frac{r}{1-r}\end{aligned}$$

(d) Fordelen er at ved å se på par av ord, kan en lettere få med hele meningen med mailen. Ulempen er at det vil være flere parametre å estimere og at i korte mail så vil par forekomme svært sjeldent.

Oppgave 8

(a) Siden $\hat{\theta}$ er et gjennomsnitt, så er forventningen til $\hat{\theta}$ lik forventningen til enkeltobservasjoner. Vi har at $\widehat{F}(x)$ svarer til en diskret fordeling over punktene $\{x_1, ..., x_n\}$, hver med sannsynlighet $\frac{1}{n}$. Dermed er

$$\mathsf{E}^{\widehat{F}}[\hat{\theta}] = \mathsf{E}^{\widehat{F}}[X] = \sum_{i=1}^{n} \tfrac{1}{n} x_i = \bar{x}.$$

Av definisjonen til $\theta(F)$ blir også $\theta(\widehat{F})$ lik $\bar{x}$.

Bootstrap estimatet på skjevhet er definert ved $E^{\widehat{F}}[\hat{\theta}] - \theta(\widehat{F})$ og blir dermed lik 0.

(*b*) Vi har at

$$\begin{aligned}\mathrm{Var}^{\widehat{F}}[X] =& \mathsf{E}^{\widehat{F}}[(X - \mathsf{E}^{\widehat{F}}[X])^2] \\ =& \mathsf{E}^{\widehat{F}}[(X - \bar{x})^2] \\ =& \sum_{i=1}^{n} \frac{1}{n}[(x_i - \bar{x})^2] = s^2.\end{aligned}$$

Vi har videre at siden $\hat{\theta}$ er et gjennomsnitt at

$$\mathsf{E}^{\widehat{F}}[(\hat{\theta} - \mathsf{E}^{\widehat{F}}[\hat{\theta}])^2] = \mathrm{Var}^{\widehat{F}}[\hat{\theta}] = \frac{1}{n}\mathrm{Var}^{\widehat{F}}[X] = \frac{1}{n}s^2$$

som viser at bootstrap estimated for varians av $\hat{\theta}$ svarer til bruk av den empiriske variansen av de opprinnelige data.