FORMELSAMLING TIL STK2100

(Versjon Mai 2018)

1 Tapsfunksjoner

(*a*) For regresjon brukes vanligvis kvadratisk tap: $L(y, \hat{y}) = (y - \hat{y})^2$. Den optimale prediktor basert på input variable $\mathbf{x}$ er da $\widehat{Y} = E[Y|\mathbf{x}]$.

(*b*) For klassifikasjon brukes vanligvis 0-1 tap: $L(y, \hat{y}) = I(y = \hat{y})$ der $I(\cdot)$ er indikatorfunksjonen. Den optimale prediktor basert på input variable $\mathbf{x}$ er da $\widehat{Y} = \text{argmax}_k \Pr(Y = k|\mathbf{x})$.

2 Multippel lineær regresjon

(*a*) Modell:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i \,;\, i = 1, 2, \ldots, n \,;$$

der x_{ij}-ene er kjente tall og ϵ_i-ene er uavhengige og $N(0, \sigma^2)$-fordelte.

(*b*) Matriseform:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

der $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$ og $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)^T$ er henholdsvis n- og $(p+1)$-dimensjonale vektorer, og $\mathbf{X} = \{x_{ij}\}$ (med $x_{i0} = 1$) er en $n \times (p+1)$-dimensjonal matrise.

(*c*) Minste kvadraters estimator for $\boldsymbol{\beta}$ er $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$.

(*d*) La $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \ldots, \hat{\beta}_p)^T$. Da er $\hat{\beta}_j$-ene normalfordelte og forventningsrette, og

$$\text{Var}(\hat{\beta}_j) = \sigma^2 c_{jj} \quad \text{og} \quad \text{Cov}(\hat{\beta}_j, \hat{\beta}_l) = \sigma^2 c_{jl}$$

der c_{jl} er element (j, l) i $(p+1) \times (p+1)$ matrisen $\mathbf{C} = (\mathbf{X}^T\mathbf{X})^{-1}$.

(*e*) La $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ik}$, og sett $\text{SSE} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$. Da er $S^2 = \frac{\text{SSE}}{n-(p+1)}$ en forventningsrett estimator for σ^2, og $[n - (p+1)]S^2/\sigma^2 \sim \chi^2_{n-(p+1)}$. Videre er S^2 og $\hat{\boldsymbol{\beta}}$ uavhengige.

(*f*) La $\text{SE}(\hat{\beta}_j)^2$ være den variansestimatoren for $\hat{\beta}_j$ vi får ved å erstatte σ^2 med S^2 i formelen for $\text{Var}(\hat{\beta}_j)$ i punkt (*b*). Da er $(\hat{\beta}_j - \beta_j)/\text{SE}(\hat{\beta}_j) \sim t_{n-(p+1)}$.

(*g*) Vi kan teste hypotesen $H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$ ved å bruke testobservatoren

$$\mathrm{F} = \frac{(\mathrm{SST} - \mathrm{SSE})/p}{\mathrm{SST}/(n-p-1)}$$

der $\mathrm{SSE} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ og $\mathrm{SST} = \sum_{i=1}^{n}(y_i - \bar{y})^2$. Under H_0 er F F-fordelt med p og $n-p-1$ frihetsgrader.

(*h*) Vi kan teste hypotesen

$$H_0 : \beta_{i_1} = \beta_{i_2} = \cdots = \beta_{i_q} = 0$$

ved å bruke testobservatoren

$$\mathrm{F} = \frac{(\mathrm{SSE}_0 - \mathrm{SSE})/q}{\mathrm{SSE}/(n-p-1)} \overset{H_0}{\sim} F_{q,n-p-1}$$

der $\mathrm{SSE}_0 = \sum_{i=1}^{n}(y - \hat{y}_i)^2$ når $\hat{y}_i$ er beregnet *under* H_0 mens SSE er tilsvarende for full modell.

3 Maksimum likelihood metoden

Anta at $Y_1, Y_2, \ldots, Y_n$ har simultan punktsannsynlighet/sannsynlighetstetthet $f(y_1, y_2, \ldots, y_n \,|\, \boldsymbol{\theta})$, der $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)$ er en parametervektor (skalar hvis $d = 1$). Vi antar at $f(y_1, y_2, \ldots, y_n \,|\, \boldsymbol{\theta})$ tilfredsstiller visse deriverbarhetsbetingelser.

(*a*) Gitt observerte verdier $Y_i = y_i$; $i = 1, \ldots, n$; er likelihood-funksjonen $L(\boldsymbol{\theta}) = f(y_1, y_2, \ldots, y_n \,|\, \boldsymbol{\theta})$ og loglikelihood-funksjonen $l(\boldsymbol{\theta}) = \log L(\boldsymbol{\theta})$.

(*b*) Maksimum likelihood *estimatet* er den verdien av $\boldsymbol{\theta}$ som maksimerer $L(\boldsymbol{\theta})$ eller ekvivalent maksimerer $l(\boldsymbol{\theta})$. Hvis vi erstatter de observerte y_i-ene med de stokastiske Y_i-ene, får vi maksimum likelihood *estimatoren*.

(*c*) Maksimum likelihood estimatet $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \ldots, \hat{\theta}_d)$ er en løsning av ligningene $s_j(\boldsymbol{\theta}) = 0$; $j = 1, \ldots, d$; der $s_j(\boldsymbol{\theta}) = (\partial/\partial\theta_j) l(\boldsymbol{\theta})$ er score-funksjonene. Vektoren av scorefunksjoner er $\mathbf{s}(\boldsymbol{\theta}) = (s_1(\boldsymbol{\theta}), \ldots, s_d(\boldsymbol{\theta}))^T$.

(*d*) Den observerte informasjonsmatrisen $\bar{\mathbf{J}}(\boldsymbol{\theta})$ er $d \times d$ matrisen med element (i, j) gitt ved $\bar{J}_{ij}(\boldsymbol{\theta}) = -\frac{\partial^2}{\partial\theta_i \partial\theta_j} l(\boldsymbol{\theta})$.

Den forventede informasjonsmatrisen (eller Fishers informasjonsmatrise) $\bar{\mathbf{I}}(\boldsymbol{\theta})$ er $d \times d$ matrisen med element (i, j) gitt ved $\bar{I}_{ij}(\boldsymbol{\theta}) = \mathrm{E}[\bar{J}_{ij}(\boldsymbol{\theta})]$.

For uavhengige og identisk fordelte observasjoner har vi at $\bar{\mathbf{I}}(\boldsymbol{\theta}) = n\mathbf{I}(\boldsymbol{\theta})$ der $\mathbf{I}(\boldsymbol{\theta})$ er forventet informasjon til én observasjon.

(*e*) Når ligningene i punkt (*c*) ikke har en eksplisitt løsning, kan vi finne maksimum likelihood estimatet ved å bruke Newton-Raphsons metode:

$$\boldsymbol{\theta}^{(s+1)} = \boldsymbol{\theta}^{(s)} + \bar{\mathbf{J}}^{-1}(\boldsymbol{\theta}^{(s)}) \mathbf{s}(\boldsymbol{\theta}^{(s)}),$$

ved å bruke Fishers scoringsalgoritme:

$$\boldsymbol{\theta}^{(s+1)} = \boldsymbol{\theta}^{(s)} + \bar{\mathbf{I}}^{-1}(\boldsymbol{\theta}^{(s)})\mathbf{s}(\boldsymbol{\theta}^{(s)}),$$

eller ved passende modifikasjoner av disse.

(f) Når vi har "tilstrekkelig mye" data, er $\hat{\theta}_i$ tilnærmet normalfordelt med forventning θ_i og med varians lik det i-te diagonalelementet til $\bar{\mathbf{I}}^{-1}(\boldsymbol{\theta})$. Kovariansen mellom $\hat{\theta}_i$ og $\hat{\theta}_j$ er tilnærmet lik element (i,j) i $\bar{\mathbf{I}}^{-1}(\boldsymbol{\theta})$. Vi kan estimere varianser/kovarianser ved å sette inn $\hat{\boldsymbol{\theta}}$ for $\boldsymbol{\theta}$ i $\bar{\mathbf{I}}^{-1}(\boldsymbol{\theta})$ eller i $\bar{\mathbf{J}}^{-1}(\boldsymbol{\theta})$.

4 Modell seleksjonskriterier

(a) Frihetsgrader: Antall frie parametre i en modell.

For modeller der $\hat{\mathbf{y}} = \mathbf{S}\mathbf{y}$, er antall frihetsgrader lik Trase$(\mathbf{S}) = \sum_i S_{ii}$.

(b) AIC er definert ved AIC $= -2l(\hat{\boldsymbol{\theta}}) + 2|\boldsymbol{\theta}|$ der $|\boldsymbol{\theta}|$ er antall *frie* parametre i modellen.

(c) BIC er definert ved BIC $= -2l(\hat{\boldsymbol{\theta}}) + \log(n)|\boldsymbol{\theta}|$.

5 Noen andre metoder for regresjon

(a) K-nærmeste nabo regresjon er definert ved

$$\hat{f}(\mathbf{x}_0) = \frac{1}{K}\sum_{\mathbf{x}_i \in \mathcal{N}_0} y_i$$

der $\mathcal{N}_0 \subset \{\mathbf{x}_1, ..., \mathbf{x}_n\}$ som inneholder de K nærmeste punkter til $\mathbf{x}_0$.

(b) Kjernemetoder (for eksempel lokal regresjon) er konseptuelt som K-nærmeste nabo regresjon, men påvirkning av en observasjon avhenger av avstanden fra interessepunktet,

$$w_i = \frac{1}{h} w\left(\frac{x_i - x_0}{h}\right)$$

der x_i er observasjonen og x_0 interessepunktet. Typiske kjerner $w(z)$ er

- Normal, $\frac{1}{\sqrt{2\pi}}\exp\{(-z^2/2)\}$, støtte $\mathcal{R}$;
- Rectangular, $1/2$, støtte $(-1, 1)$;
- Epanechnikov, $\frac{3}{4}(1 - z^2)$, støtte $(-1, 1)$;
- Biquadratic, $\frac{15}{16}(1 - z^2)^2$, støtte $(-1, 1)$;
- Tricubic, $\frac{70}{81}(1 - |z|^3)^3$, støtte $(-1, 1)$;

(*c*) Ridge regresjon: Minimer mhp $\boldsymbol{\beta}$

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

(*d*) Lasso regresjon: Minimer mhp $\boldsymbol{\beta}$

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

(*e*) Kubisk spline: Stykkevis polynomisk med basisfunksjoner

$$\begin{aligned} b_0(x) =& 1, \quad b_1(x) = x, \quad b_2(x) = x^2, \quad b_3(x) = x^3, \\ b_{3+k}(x) =& (x - c_k)_+^3, \quad k = 1, ..., K \end{aligned}$$

(*f*) Tre-baserte metoder: $f(\boldsymbol{x}) = \sum_{m=1}^{M} c_m I(\boldsymbol{x} \in R_m)$ der $\mathcal{R}^p = R_1 \cup R_2 \cup \cdots \cup R_M$ og regioner er definert gjennom sekvensiell oppsplitting basert på én variabel om gangen.

(*g*) Bagging og random forrest:

$$\hat{f}_{\text{avg}}(\mathbf{x}) = \frac{1}{B} \sum_{b=1}^{B} \hat{f}^b(\mathbf{x})$$

der $\hat{f}^1(\mathbf{x}), \hat{f}^2(\mathbf{x}), ..., \hat{f}^B(\mathbf{x})$ er B ulike prediktorer basert på ordinær bootstrapping (bagging) eller der oppsplitting kun vurderes blandt en delmengde av forklaringsvariablene (random forrest).

(*h*) Nevrale nett med ett latent lag: $f(\boldsymbol{x}) = \beta_0 + \sum_{m=1}^{M} \beta_k \sigma(\alpha_m^T \boldsymbol{x})$.

6 Noen metoder for klassifikasjon

(*a*) K-nærmeste nabo klassifikasjon er definert ved

$$\Pr(Y = j | \mathbf{X} = \mathbf{x}_0) = \frac{1}{K} \sum_{\mathbf{x}_i \in \mathcal{N}_0} I(y_i = j)$$

der $\mathcal{N}_0 \subset \{\mathbf{x}_1, ..., \mathbf{x}_n\}$ som inneholder de K nærmeste punkter til $\mathbf{x}_0$.

(*b*) Logistisk regresjon: $Y \in \{0, 1\}$ og

$$\Pr(Y = 1 | \mathbf{x}) = \frac{e^{\mathbf{x}^T \boldsymbol{\beta}}}{1 + e^{\mathbf{x}^T \boldsymbol{\beta}}} = 1 - \Pr(Y = 0 | \mathbf{x}).$$

(*c*) Bruk av Bayes teorem for klassifikasjon

$$\Pr(Y = k | X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}.$$

(*i*) LDA: $f_k(\mathbf{x}) = p(x|y = k) = N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$.

(*ii*) QDA: $f_k(\mathbf{x}) = p(x|y = k) = N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$.

7 Dimensjonsreduksjon

(*a*) Prinsipale komponenter: 1. prinsipale komponent definert gjennom $z_1 = \boldsymbol{\phi}_1^T \mathbf{x}$ der $\boldsymbol{\phi}_1$ er valgt slik at var(z_1) er størst mulig.

(*b*) Partial least squares: Bruker også responsen til å definere de transformerte variable.

8 Hierarkisk klusteranalysen

(*a*) Dekomponering av *total dissimilarity*-en,

$$\sum_{ij} d(i,i') = \sum_{k=1}^{K} \sum_{G(i)=k} \sum_{G(i')=k} d(i,i') + \sum_{k=1}^{K} \sum_{G(i)=k} \sum_{G(i')\neq k} d(i,i')$$

der $G(i)$ angir gruppen som i-te observagjon tilhører, K er totalt antall grupper od $d(i,i') = \sum_i d_j(x_{ij}, x_{i'j}))$ er *dissimilarity*-en mellom observasjoner i og i'. Her er x_{ij} j-ne delen av observasjonen i.

(*b*) *Dissimilarity measures* mellom grupper:

- single link, $\min_{i\in G, i'\in G'} d(i,i')$
- complete link, $\max_{i\in G, i'\in G'} d(i,i')$
- average link, $\frac{1}{n_G n_{G'}} \sum_{i\in G} \sum_{i'\in G'} d(i,i')$.