Oppgaver, STK4040, uke 44

30. oktober 2005

Oppgave 1: MSE

La f være en regresjon trent på et datasett X og $\boldsymbol{y}$, og la $\hat{y} = f(\boldsymbol{x})$ være predikert respons for en gitt ny observasjon $\boldsymbol{x}$. Vis at

$$\mathrm{MSE}(\hat{y}) = \mathrm{Bias}^2(\hat{y}) + \mathrm{Var}(\hat{y} - Y), \tag{1}$$

der Y er (den ukjente) responsverdien tilhørende $\boldsymbol{x}$. (Vi regner X og $\boldsymbol{x}$ som gitte.) Merk at dette gjelder generelt for regresjoner; vi har ikke antatt noe om formen til f (f.eks. at det er en lineær regresjon).

Oppgave 2: leverage

Anta modellen

$$\boldsymbol{y} = \mathrm{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{2}$$

hvor X er kjent og $\boldsymbol{\varepsilon} \sim N_n(0, \sigma^2\mathrm{I})$.

Gitt et datasett X og $\boldsymbol{y}$. La $e_i = \hat{y}_i - y_i$ være residualene fra en lineær regresjon av $\boldsymbol{y}$ på X. La h_i være diagonalelementene til hattmatrisen $\mathrm{H} = \mathrm{X}(\mathrm{X}^t\mathrm{X})^{-1}\mathrm{X}^t$. h_i kalles «leverage»n til $\boldsymbol{x}_i$, og er et mål på hvor mye $\boldsymbol{x}_i$ potensielt kan påvirke regresjonen. Merk at $h_i = \boldsymbol{x}_i^t(\mathrm{X}^t\mathrm{X})^{-1}\boldsymbol{x}_i$.

a)

La $\mathrm{X}_{(i)}$ og $\boldsymbol{y}_{(i)}$ være X og $\boldsymbol{y}$ uten den $i$te observasjonen, og la $\boldsymbol{b}_{(i)}$ være estimatet av $\boldsymbol{\beta}$ fra $\mathrm{X}_{(i)}$ og $\boldsymbol{y}_{(i)}$. Vis at

$$e_{(i)} = y_i - \boldsymbol{x}_i^t\boldsymbol{b}_{(i)} = e_i/(1 - h_i). \tag{3}$$

(Med andre ord: man kan beregne residualene fra en full kryssvalidering («leave one out cross-validation») av en lineærregresjon uten å måtte gjøre selve regresjonen om igjen for hver observasjon.)

Tips: for å spare litt regning, kan dere bruke at

$$(\mathrm{X}_{(i)}^t\mathrm{X}_{(i)})^{-1} = (\mathrm{X}^t\mathrm{X})^{-1} + \frac{(\mathrm{X}^t\mathrm{X})^{-1}\boldsymbol{x}_i\boldsymbol{x}_i^t(\mathrm{X}^t\mathrm{X})^{-1}}{1 - h_i}. \tag{4}$$

(Ekstraoppgave: Bruk at $\mathrm{X}_{(i)}^t\mathrm{X}_{(i)} = \mathrm{X}^t\mathrm{X} - \boldsymbol{x}_i\boldsymbol{x}_i^t$ til å vise (4).)

b)

Anta at $\mathrm{X} = (\mathbf{1}\ \mathrm{X}_1)$. La X_c være X_1 sentrert, og la $\boldsymbol{y} = \beta_0\mathbf{1} + \mathrm{X}_c\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon}$ være den ekvivalente modellen. Kall leverage'ne til denne regresjonen for g_i, og vis at

$$h_i = g_i + 1/n, \tag{5}$$

hvor n er antall observasjoner.