Intermezzo: tre ligninger med fire ubekendte

PLS-regression gør det muligt at lave hurtigmetoder og finde sammenhænge i komplicerede data, men det er helt essentielt, at man sikrer sig mod, at modellen bruger tilfældige variationer. Et lille teoretisk eksempel viser nødvendigheden af validering

Af Rasmus Bro, Søren Balling Engelsen & Lars Nørgaard, Institut for Fødevarevidenskab, Det Biovidenskabelige Fakultet, Københavns Universitet

Hovedfordelen ved PLS-regression er, at man kan håndtere virkelig mange variable. Ofte er der flere variable end der er prøver, og det giver mulighed for "snyde" (over-fitte). Hvis man bruger for mange komponenter, så vil modellen kunne beskrive stort set hvad som helst, uden at det reelt giver mening. Heldigvis findes der en simpel måde at sikre sig mod dette: krydsvalidering. I denne klumme viser vi et lille eksempel på, hvorfor man må validere, når man har mange variable, specielt når disse variable er korrelerede.

Et lille eksempel

Lad os sige at vi har målt fire X-målinger på tre prøver. Desuden har vi en y-værdi for hver prøve, og den vil vi gerne prædiktere ud fra X-målingerne. De fire X-målinger er identiske og vores datamatrix, **X**, og reference-værdier, **y**, ser ud som følger

$$\mathbf{X} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 2 & 2 & 2 & 2 \\ 3 & 3 & 3 & 3 \end{bmatrix}, \mathbf{y} = \begin{bmatrix} 1 \\ 3 \\ 8 \end{bmatrix}$$

Som det kan ses, så er de fire variable ikke så informative, i og med at de er helt ens. Hvis vi anvender almindelig lineær regression[1], kan vi finde den bedste regressionsvektor ($\mathbf{b}_{LS}$ hvor LS står for Least Squares) som

$$\mathbf{b}_{LS} = \begin{bmatrix} .55 \\ .55 \\ .55 \\ .55 \end{bmatrix}$$

Hver X-variabel har samme vægt i modellen – samme regressionskoefficient, fordi de alle er ens. Vha. denne regressionsvektor kan vi nu prædiktere **y** ud fra **X**. For hver prøve tages første X-variabel gange .55, næste variabel gange .55 osv. Resultatet bliver

$$\hat{\mathbf{y}} = \mathbf{X}\mathbf{b}_{LS} = \begin{bmatrix} 2.2 \\ 4.4 \\ 6.6 \end{bmatrix}$$

Sammenlignes med den sande **y**-matrix, så går det ikke helt godt, men det er, hvad der er muligt med lineær regression. Så langt så godt! Der sker imidlertid noget interessant, hvis **X** ikke er så perfekt som ovenfor. Hvis **X** indeholder samme information som før, men nu med lidt støj, så bliver de fire variable ikke identiske

$$\mathbf{X} = \begin{bmatrix} 1.1 & 1.2 & .9 & .9 \\ 2.1 & 1.9 & 2.2 & 1.8 \\ 2.9 & 3.1 & 3.1 & 3.1 \end{bmatrix}, \mathbf{y} = \begin{bmatrix} 1 \\ 3 \\ 8 \end{bmatrix}$$

Som det kan ses, så er der tale om næsten samme problemstilling, blot med en smule variation pga. tilfældig støj. Resultatet af en lineær regression er dog markant anderledes.

$$\mathbf{b}_{LS} = \begin{bmatrix} -4.6 \\ -0.4 \\ 0.8 \\ 6.5 \end{bmatrix}$$

På trods af at de fire variable er næsten identiske, så er de fire regressionskoefficienter *helt* forskellige. Det er bestemt ikke et godt tegn, fordi vi forventer, at løsningen kun skal ændre sig lidt, når data ændres lidt (robusthed). Endnu mere bemærkelsesværdigt er det, at prædiktionerne nu pludselig bliver langt bedre end før

$$\hat{\mathbf{y}} = \mathbf{X}\mathbf{b}_{LS} = \begin{bmatrix} 1.0 \\ 3.0 \\ 8.0 \end{bmatrix}$$

Som det kan ses, så er prædiktionerne faktisk *helt* perfekte. Det er jo dejligt, at modellen nu virker perfekt, men det er også næsten for godt til at være sandt. Sagen er at modellen "overfitter". Modellen bruger den tilfældige støj i data til at fitte/beskrive data perfekt, og det giver ikke rigtigt mening. At det ikke ▶

giver mening kan let ses, hvis man forsøger at prædiktere to nye prøver med støj. Begge prøver ligner den første prøve, hvor alle variable var ca. en, så vi forventer at svaret også er ca. en.

$$\mathbf{X} = \begin{bmatrix} 0.8 & 0.9 & 0.9 & 1.1 \\ 1.2 & 1.2 & 1 & 0.8 \end{bmatrix}, \hat{\mathbf{y}} = \mathbf{Xb} = \begin{bmatrix} 3.8 \\ 0.0 \end{bmatrix}$$

Som det ses, så kan små ændringer få prædiktionerne til at variere voldsomt – her fra 0.0 til 3.8. I praksis må man sige, at modellen er ubrugelig. Støjen har alt for stor indflydelse på regressionsmodellen. Problemet kan i bund og grund koges ned til, at vi har tre ligninger (prøver) og fire ubekendte (regressionskoefficienter), og det forhold gør, at man faktisk kan finde en hvilken som helst sammenhæng, man måtte ønske. I matematikken siger vi, at systemet er overbestemt, og der er en uendelighed af løsninger (regressionskoefficienter), som passer perfekt på de tre prøver.

Outro - løsningen er validering

En vigtig observation ovenfor er, at vi faktisk *ikke* kan bruge kalibreringsprøverne til at se, om modellen er god eller dårlig. Derimod kan vi ud fra nye prøver se, om modellen er god eller dårlig. Når man skal vurdere en kalibreringsmodel, bør man altså bruge et *testsæt,* sådan at modellen beregnes på kalibreringssættet og godheden eftervises på testsættet. Dette kaldes testsæt-validering, og i næste klumme vil vi vise, hvordan man kan lave noget tilsvarende, når ens samlede datasæt ikke er stort nok til at blive delt op i både et kalibrerings- og et testsæt. Dette kaldes krydsvalidering. Vi vil også vise, hvordan vi ikke blot kan bruge valideringen til at bestemme, hvor godt modellen prædikterer, men også til at bestemme hvor mange komponenter, vi skal bruge f.eks. i en PLS-model.

E-mail-adresser
Rasmus Bro: rb@life.ku.dk
Søren Balling Engelsen: se@life.ku.dk
Lars Nørgaard: lan@life.ku.dk

Fodnote
[1)] I virkeligheden anvendes en variant af lineær regression, som kan håndtere, at de fire variable er helt identiske.

Thorvalds madkort

Boxty Pancakes

1 lb/450 grams potatoes, peeled and grated
6 oz/170 grams plain flour
1/2 tspn baking powder
1/4 tspn salt
1 egg, beaten
4 fl. oz/125ml milk

Peel and grate the potatoes. Sift together the flour, salt and baking powder. Mix with the potatoes. Add the egg and enough milk to make a thick batter. Drop dessertspoonfuls of batter into a hot greased frying pan. Cook either side for 3-4 minutes or until brown.
Serve hot with butter.

Æg og bagepulver

I "Kemikeren i køkkenet" 2, 2005 (den kan hentes på www.danskkemi.dk) skrev jeg om, hvad det er bagepulver gør, og hvad æggenes rolle i grunden er. Af mange kageopskrifter fremgår det, at æggene medvirker til at få dejen til at hæve. Men det er bagepulverets rolle, det består nemlig af et carbonat og en syre som f.eks. natriumhydrogentartrat. Æggets proteiner koagulerer ved ca. 100°C og konsoliderer derved bagværkets struktur.

ThP, thorpe@tdcadsl.dk