Projekt 1: Överanpassning och korsvalidering

4 maj 2017

Betrakta en mängd av parvisa observationer (x_i, y_i), $i = 1, \ldots, n$. I linjär regression vill man anpassa en linje $y = b_0 + b_1 x$ till data i modellen

$$y_i = b_0 + b_1 x_i + \epsilon_i$$

där ϵ_i:na är oberoende och normalfördelade med en okänd varians σ^2. ML-skattningarna $\hat{b}_0$ och $\hat{b}_1$ av b_0 och b_1 fås via minsta-kvadratmetoden, dvs man minimerar $\sum_{i=1}^{n}(y_i - (b_0 + b_1 x_i))^2$ över b_0 och b_1.

Om man istället använder modellen

$$y_i = b_0 + b_1 x + b_2 x^2 + \epsilon_i$$

ges ML-skattningarna av att minimera $\sum_{i=1}^{n}(y_i - (b_0 + b_1 x_i + b_2 x_i^2))^2$ och man kommer att få en skattad kurva som ligger närmare observerade data.

Generellt kan man anpassa data enligt en allmän polynommodell

$$y_i = \sum_{k=0}^{p} b_k x_i^k + \epsilon_i$$

och skattar b_k:na genom att minimera $\sum_{i=1}^{n}(y_i - \sum_{k=0}^{p} b_k x_i^k)^2$ och ju högre grad p, desto bättre kommer den skattade kurvan att anpassa sig till observerade data och om $p = n - 1$ får man en kurva som exakt kommer att kommer att passera genom alla datapunkter, dvs en *perfekt* anpassning.

Man kan nu fråga sig varför man inte helt enkelt låter $p = n-1$. Svaret ligger i att man då kommer att får en *överanpassning*. Ju högre p man väljer, desto ”svängigare” blir den skattade kurvan och man kan misstänka att om man fick nya observationer, skulle dessa passa mycket illa med den skattade kurvan.

För att utvärdera vilket p som är det optimala valet, använder man ofta så kallad *korsvalidering* och uppgiften är att utföra en sådan. Simulera datapunkter (x_i, y_i), $i = 1, \ldots, n$ med $n = 1000$ enligt den allmänna polynommodellen för de olika värdena 1, 2, 5 och 10 på p. Ni väljer själva x_i:na, b_k:na och σ^2. Välj x_i:na ganska samlade och σ^2 i förhållande till b_k:na och x_i:na så att det underliggande polynomet inte syns särskilt väl i en plot. Dessa observationer ska vi nu anpassa polynom av olika grad till. Dela slumpvis in observationerna i två olika mänger D_1 och D_2. Använd sedan observationerna i D_1 till att anpassa varsitt polynom av graderna $p = 1, 2, \ldots, 50$. Beräkna sedan $\sum_{i \in D_2}(y_i - \sum_{k=0}^{p} \hat{b}_k x_i)^2$ för vart och ett av p:na. Vilket p gör denna kvadratavvikelse minst? Eftersom vi själva simulerat data, vet vi ju vad det ”borde bli” och det är intressant att se om detta är det svar vi får. Vilket resultat man får beror på värdet på σ^2 i förhållande till b_k:na och x_i:na i simuleringen, så pröva gärna olika värden. Använd Matlab och redovisa med åskådliggörande plottade kurvor. Matlab har kommandot ”polyfit” som är mycket användbart.