Um information

Helgi Tómasson
helgito@hi.is

Upplýsingar

Hvaða upplýsingar eru í gögnum? Upplýsingar eru mælikvarði á hversu nákvæmlega megi ályta um líkan (parametra í líkani).

Um AIC. Vandinn er að nálga óþekkta dreifingu g með f sem er úr þekktri fjölskyldu af dreifingum. *Kullback-Leibler* fjarlægð er skilgreind:

$$I(g, f) = \int g(x)\log(g(x))dx - \int g(x)\log(f(x))dx$$

Þetta er lágmarkað með því að gera:

$$-\int g(x)\log(f(x))dx = E_g(\log(f(X)), \tag{1}$$

sem minnst. Hér er dreifingin g óþekkt og því þarf að meta væntanlega gildið í (1). Eðlilegt er að meta það með úrtaksmeðaltali, þ.e. tek úrtak $X_1, \ldots, X_n$ úr g og reikna metilinn:

$$-\frac{1}{n}\sum_{i=1}^{n}\log(f(X_i)). \tag{2}$$

Oft þykir eðlilegt að velja f úr fjölskyldu sem stýrt er með stika (parameter), $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)'$. Hér flækjustigi líkans stýrt með fjölda stika. Því eru valið $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \ldots, \hat{\theta}_m)$ þannig að

$$-\frac{1}{n}\sum_{i=1}^{n}\log(f(X_i|\theta_1, \ldots, \theta_m)),$$

sé lágmarkað. Hér er fyrir hvert gildi á m fundið það gildi, $\hat{\boldsymbol{\theta}}$, sem lágmarkar (2). Síðan er fundið það gildi á m, sem lágmarkar:

$$-\frac{1}{n}\sum_{i=1}^{n}\log(f(X_i|\hat{\theta}_1, \ldots, \hat{\theta}_m)).$$

Hér er því um tvíþætta lágmörkun að ræða. Hún leiðir vanmats á (1). Með Taylorútvíkkun má sýna að vanmatið er um það bil $\frac{m}{n}$. Þannig rökstuddi Akaike (1973) að best væri að velja líkan með því að lágmarka:

$$-\frac{1}{n}\sum_{i=1}^{n}\log(f(X_i|\hat{\theta}_1, \ldots, \hat{\theta}_m)) + m/n.$$

Þetta er vel þekkt sem AIC (Akaike-Information-Criterion). Líkan er síðan fundið með því að lágmarka AIC. Algeng er að það sé skrifað sem:

$$AIC = -2\underbrace{\log(L(\hat{\boldsymbol{\theta}}_m)}_{A} + \underbrace{2m}_{B}.$$

Hér er A hefðbundið hámarksgildi sennileikafalls þar sem hámarkað hefur verið með tilliti til m-víðs stika $\hat{\boldsymbol{\theta}}$. Hluti B refsar fyrir aukið flækjustig líkans. AIC er þróað sem tíðnitölfræðihugtak. Schwarz (1978) leiðir út svipað með bayesískum rökum. Schwarz-criterion, sem stundum er nefnt BIC (Bayesian-information-criterion) er:

$$BIC = -2\log(L(\hat{\boldsymbol{\theta}}_m) + m\log(n).$$

Fyrri liður BIC er eins og AIC en seinni liðurinn refsar aðeins öðruvísi fyrir aukið flækjustig. Þó upphaflegi rökstuðningurinn með BIC sé bayesískur þá er BIC vel nothæft í hefðbundinni tölfræði því „a prior" dreifingin kemur ekki fyrir í formúlunni. Chatfield (1995) fjallar ítarlega um val á líkani. Ýmsar aðrar útgáfur af sömu hugmynd eru til, svo sem AICC, DIC, FPE.

Heimildir

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In Petroc, B. & Caski, F. (Eds.), *Second International Symposium in Information Theory*. Akademiai Kiado, Budapest.

Chatfield, C. (1995). Model uncertainty, data mining and statistical inference. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, *158*(3), 419–466.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464.