



Teorie informací a hodnocení statistických modelů

Bakalářská Práce

Tomáš Petit

505485@mail.muni.cz

Přírodovědecká fakulta, Masarykova Univerzita

27-06-2023

Motivace

Výběr modelů/pod-modelů v lineární regresi

- Výběr proměnných pro účel predikce v regresním modelu
- Snaha najít "spravedlivé" kritérium
- $y_i = \beta_0 + x_1\beta_1 + \dots + x_m\beta_m + \varepsilon_i \quad i = 1, \dots, n$

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

$$R_{adj}^2 = 1 - (1 - R^2) \frac{n-1}{n-p}$$

Cíle práce

- Představit v krátkosti základy teorie informace
- Poskytnout důkladné odvození Akaikeho informačního kritéria v plné obecnosti
- Předvést praktické využití pro širokou třídu statistických modelů

Obsah Práce

- Teorie Informace
 - Shannonova Entropie
 - Relativní Entropie
- Akaikeho informační kritérium
 - Metoda maximální věrohodnosti
 - Asymptotické vlastnosti věrohodnosti
 - Střední hodnota logaritmu věrohodnostní funkce
 - Vychýlení logaritmu věrohodnostní funkce
 - AIC
- Modelování pomocí AIC
 - Lineární a polynomiální regrese
 - Histogramy
 - Rovnost dvou diskrétních distribucí
 - Rovnost středních hodnot a rozptylů
 - Mallowsovo C_p
 - Analýza hlavních komponent

Kullback-Leiblerova divergence

Pravděpodobnostní distribuce P_X a Q_X náhodné veličiny X na pravděpodobnostním prostoru $(\Omega, \mathcal{F}, \mu)$

$$I(P_X; Q_X) = E_{P_X} \left[\ln \left(\frac{dP_X}{dQ_X} \right) \right] = \int_{\Omega} P_X \ln \left(\frac{dP_X}{dQ_X} \right) dP_X$$

- **Není metrikou**
- **Množství informace ztracené při nahrazení P_X za Q_X**
- **Problém praktického výpočtu \rightarrow potřeba odhadu**

Expected likelihood

$$E_{f(x)}[\ln(g(x))] = \int_{-\infty}^{\infty} f(x) \ln(g(x)) dx \quad (1)$$

$$= \sum_{\alpha=1}^n \hat{f}(x_{\alpha}) \ln(g(x_{\alpha})) \quad (2)$$

$$= \frac{1}{n} \sum_{\alpha=1}^n \ln(g(x_{\alpha})) \quad (3)$$

Tedy

$$n \int_{-\infty}^{\infty} f(x) \ln(g(x)) dx = \sum_{\alpha=1}^n \ln(g(x_{\alpha})) \quad (4)$$

Informační kritéria

Takeuchiho informační kritérium

$$\text{TIC} = -2 \sum_{\alpha=1}^n \ln f(X_{\alpha} | \hat{\theta}) + 2 \text{tr}\{J(\theta)I(\theta)^{-1}\} \quad (5)$$

Akaikeho informační kritérium

$$\text{AIC} = -2 \sum_{\alpha=1}^n \ln f(X_{\alpha} | \hat{\theta}) + 2p \quad (6)$$

Aplikace

- Hodnocení statistických modelů
 - Lineární regresní modely (polynomiální, ANOVA etc.)
 - PCA
 - Histogramy
 - Ekvivalence množin kategoriálních dat
 - Časové řady
 - Testování rovnosti středních hodnot

Příklad 1

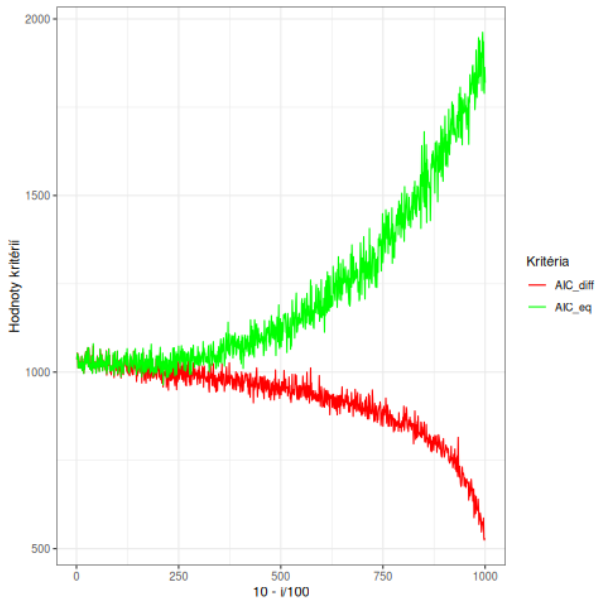
Volně dostupná data **Swiss**

- Model 1 - Fertility \sim Catholic
- Model 2 - Fertility \sim Catholic + Education + Infant.Mortality + Agriculture
- Model 3 - Full model

Model	AIC	R^2	R^2_{adj}
1	364.3479	0.215	0.1976
2	325.2408	0.6993	0.6707
3	326.0716	0.7067	0.671

Table: Srovnání AIC, R^2 a R^2_{adj}

Příklad 2



Shrnutí

- Základ v teorii maximální věrohodnosti
- Lze provést bodové odhady i intervaly spolehlivosti
- Lze použít jako základ pro statistiku

Děkuji za pozornost!

Otázka 1.1

Jaká je souvislost mezi p v (2.59) a k ve vzorci pro AIC_c ?

$$AIC_c = AIC + \frac{2p(p+1)}{n-p-1}$$

■ Překlep

Otázka 1.2

V odstavci před Příkladem 3.1.2 zmiňujete, že automatizace výběru modelu se nedoporučuje jako hlavní nástroj. Jaký postup byste Vy, resp. použité zdroje, doporučil?

- Partial least squares
- LASSO
 - $\min_{\beta_0, \beta} \{ \|y - \beta_0 - \mathbf{X}\beta\|_2^2 \}$ za podmínky $\|\beta\|_1 \leq t$
- Least Angle Regression
- Znalost dat a problematiky

Otázka 2.1

Ve třetí kapitole uvádíte dva přístupy výběru modelu pomocí AIC (forward-selection a backward-selection). Po použití obou těchto přístupů dostáváte stejný výsledný model. Funguje to tak vždy? Existuje ještě nějaký další přístup?

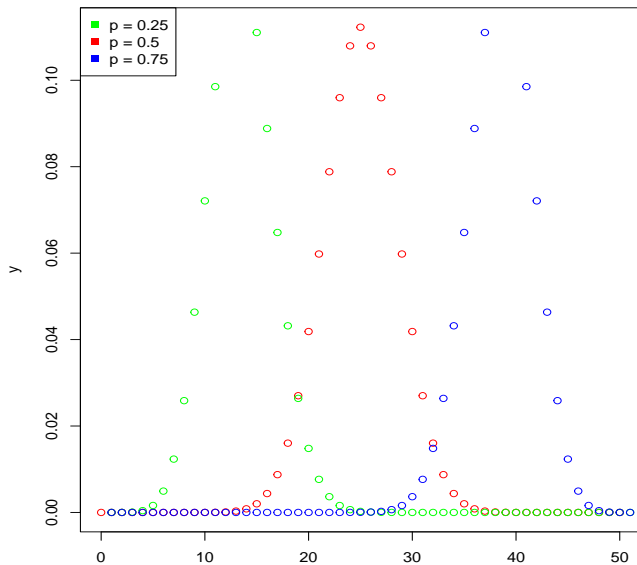
- Nemusí to vždy platit
- Bidirectional elimination

Otázka 2.2

Data z Příkladu 3.1.2 modelujete pomocí polynomiální regrese čtvrtého řádu. Je tento model pro uvedený datový soubor vhodný? Pokud ne, jaký jiný model byste použil?

Otázka 2.3

Na str. 28 aproximujete data z normálního rozdělení binomickým rozdělením s parametrem $p = 1/2$. Proč právě $p = 1/2$ dává nejlepší aproximaci?



Otázka 2.4

Co způsobuje oscilaci na Obrázcích 3.6 a 3.7?

Otázka 2.5

Jak bychom mohli zlepšit odhady rozptylů v Tabulce 3.6, kde je počet pozorování $n = 10$.

**MASARYK
UNIVERSITY**