Finanční Matematika – 6. přednáška
Poznámka k regresi, aposteriorní odhady

29. března 2016

Regrese (zpětný postup)
$Y = f(\mathbf{X}, \boldsymbol{\beta})$

Regrese (zpětný postup)
$Y = f(\mathbf{X}, \boldsymbol{\beta})$
Y je závislá (měřená) veličina,

Regrese (zpětný postup)
$Y = f(\mathbf{X}, \beta)$
Y je závislá (měřená) veličina, X potenciální proměnné, na kterých X závisí,

Regrese (zpětný postup)

$Y = f(\mathbf{X}, \boldsymbol{\beta})$

Y je závislá (měřená) veličina, X potenciální proměnné, na kterých X závisí, β neznámé parametry.

Regrese (zpětný postup)
$Y = f(\mathbf{X}, \boldsymbol{\beta})$
Y je závislá (měřená) veličina, X potenciální proměnné, na kterých X závisí, β neznámé parametry.
f je předpokládaná funkčí závislost. (např Poissonova, smíšená Poissonova)

Regrese (zpětný postup)
$Y = f(\mathbf{X}, \boldsymbol{\beta})$
Y je závislá (měřená) veličina, X potenciální proměnné, na kterých X závisí, β neznámé parametry.
f je předpokládaná funkčí závislost. (např Poissonova, smíšená Poissonova)

Nejčastější je hledání lineární závislosti, v našem jednoduchém případě hledáme závislost v podobě Poissonovy distribuční funkce.

Rovnice pro extrém věrohodnostní funkce:

$$\frac{\partial L(\boldsymbol{\beta})}{\partial \beta_j} = 0 \equiv \sum_{i=0} x_{ij}(k_i - \lambda_i) = 0,\ j = 1, \ldots, p$$

Označme řešení $\hat{\boldsymbol{\beta}}$, dále $\hat{\lambda}_i = d_i e^{\sum_{j=0}^{p} \hat{\beta}_j x_{ij}}$, pak

$$\hat{\Sigma}_{\hat{\beta}} = \left(\sum_{j=1}^{n} \mathbf{x_i} (\mathbf{x_i})^T \hat{\lambda} \right)$$

Interval spolehlivosti pro β_j:

$$[\hat{\beta}_j - z_{\alpha/2}\hat{\sigma}_{\hat{\beta}_j}, \hat{\beta}_j + z_{\alpha/2}\hat{\sigma}_{\hat{\beta}_j}],$$

kde $\hat{\sigma}^2_{\hat{\beta}_j}$ je prvek (j, j) matice $\hat{\Sigma}_{\hat{\beta}}$ odhadující rozptyl $\hat{\beta}_j$.

Bayesův vzorec

Theorem

Pro pravděpodobnost jevů A a B platí

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)} \tag{1}$$

$$P(A|B) = \frac{P(A)P(B|A)}{P(A)P(B|A) + P(A^c)P(B|A^c)}. \tag{2}$$

Řekněme, že testy připravenosti a znalostí, na základě kterých jsou studenti přijímáni na univerzitu, mají následující spolehlivost v testování inteligence osob: 99% inteligentních osob má pozitivní výsledek testu, zatímco u neinteligentních uchazečů má 0,5% z nich pozitivní výsledek testu. Chceme zjistit, s jakou pravděpodobností je náhodně vybraný student na univerzitě inteligentní.

Řekněme, že testy připravenosti a znalostí, na základě kterých jsou studenti přijímáni na univerzitu, mají následující spolehlivost v testování inteligence osob: 99% inteligentních osob má pozitivní výsledek testu, zatímco u neinteligentních uchazečů má 0,5% z nich pozitivní výsledek testu. Chceme zjistit, s jakou pravděpodobností je náhodně vybraný student na univerzitě inteligentní.

Máme tedy jev A „náhodně zvolená osoba je inteligentní“ a jev B „osoba prošla testem s pozitivním výsledkem“. Dle Bayesova vzorce můžeme opět rovnou spočíst pravděpodobnost, že nastal jev A za předpokladu, že nastal jev B. Musíme jen dodat všeobecnou pravděpodobnost $p = p(A)$, že náhodně zvolený uchazeč o studium je inteligentní.

$$P(A|B) = \frac{p \cdot 0{,}99}{p \cdot 0{,}99 + (1-p) \cdot 0{,}005}.$$

V následující tabulce je spočten pro různé hodnoty p vyjádřené v jednotkách promile. V prvním sloupci tedy je výsledek za předpokladu, že je mezi uchazeči o studium každý druhý inteligentní atd.

p	500	100	50	10	1	0,1
$P(A\|B)$	0,99	0,96	0,91	0,67	0,17	0,02

V následující tabulce je spočten pro různé hodnoty p vyjádřené v jednotkách promile. V prvním sloupci tedy je výsledek za předpokladu, že je mezi uchazeči o studium každý druhý inteligentní atd.

p	500	100	50	10	1	0,1
$P(A\|B)$	0,99	0,96	0,91	0,67	0,17	0,02

Pokud tedy je každý druhý uchazeč inteligentní, máme na univerzitě používající náš test 99% inteligentních studentů. Pokud ale naší představě o inteligenci odpovídá jen 1% populace a uchazeči jsou dobrým náhodným vzorkem, pak už máme na univerzitě jen zhruba dvě třetiny inteligentních studentů ...

Představme si ale, že obdobné testování provedeme při plošném testování výskytu nějaké nemoci, třeba HIV. Dejme tomu, že máme stejně citlivý test jako výše a prověříme jím o přestávce mezi přednáškami všechny přítomné studenty. V tomto případě bychom měli předpokládat, že parametr p bude obdobný jako u celé populace, tj. řekněme jeden nakažený z 10000 obyvatel, což odpovídá poslednímu sloupci v tabulce. Pak ovšem je výsledek testu katastroficky nespolehlivý. Jen asi u 2 procent pozitivně otestovaných se jedná o skutečně nemocné studenty!

Představme si ale, že obdobné testování provedeme při plošném testování výskytu nějaké nemoci, třeba HIV. Dejme tomu, že máme stejně citlivý test jako výše a prověříme jím o přestávce mezi přednáškami všechny přítomné studenty. V tomto případě bychom měli předpokládat, že parametr p bude obdobný jako u celé populace, tj. řekněme jeden nakažený z 10000 obyvatel, což odpovídá poslednímu sloupci v tabulce. Pak ovšem je výsledek testu katastroficky nespolehlivý. Jen asi u 2 procent pozitivně otestovaných se jedná o skutečně nemocné studenty!
Všimněme si, že problém je zapříčiněn jakýmkoliv malým výskytem pozitivních výsledků u zdravých osob. I kdybychom zlepšili test tak, že bude na 100% účinný při testu pozitivní osoby, neovlivníme skoro vůbec výsledné pravděpodobnosti v tabulce.

Při lékařské diagnostice vzácných chorob je při pozitivním výsledku testu nutné provést další test. Přitom výsledek prvního testu $P(A|B)$ má roli apriorní pravděpodobnosti $P(A)$ při druhém testu. Tento postup umožňuje „kumulovat zkušenost“.
V obou případech tedy musíme při přípravě testu dbát na to, abychom si zajistili přiměřeně vysoké p. U procesu přijímání studentů na univerzitu to asi bude dobrý marketing univerzity, který zajistí, aby se neinteligentní osoby hlásily v daleko menší míře, než je jejich výskyt v populaci. U testování chorob nejspíš půjde o souběh dalších skutečností a činností (např. testování HIV pozitivitu pouze u rizikových skupin obyvatelstva a podobně).

Myslíme si, že mezi pojištěnci je 60% dobrých řidičů, zbytek špatných. Pravděpodobnost, že dobrý řidič havaruje k-krát se řídí rozdělením $Poi(\lambda_G)$, $\lambda_G = 0,05$, že havaruje špatný rozdělením $Poi(\lambda_B)$, $\lambda_b = 0,15$.
Očekávaný počet nehod za rok je tak

$$P[\text{dobrý}]\lambda_G + P[\text{špatný}]\lambda_B = 0,09$$

Pokud řidič nahlásí za rok k nehod, je pravděpodobnost, že je dobrý následující:

$$\begin{aligned} &\frac{P[k\text{nehod}|\text{dobrý}]P[\text{dobrý}]}{P[k\text{nehod}|\text{dobrý}]P[\text{dobrý}] + P[k\text{nehod}|\text{špatný}]P[\text{špatný}]} \\ =&\frac{e^{-\lambda_G}\lambda_G^k P[\text{dobrý}]}{e^{-\lambda_G}\lambda_G^k P[\text{dobrý}] + e^{-\lambda_B}\lambda_B^k P[\text{špatný}]} \end{aligned}$$

Očekávaný počet nehod řidiče v následujícím roce za předpokladu nahlášení k nehod je tak

$$P[\text{dobrý}|k \text{ nehod}]\lambda_G + P[\text{špatný}|k\text{nehod}]\lambda_B$$

V tabulce jsou uvedeny číselné hodnoty:

poč. nehod	0	1	2	3	4	5
$P[\text{dobrý}\|k]$	62,33	35,59	15,55	5,78	2,00	0,68
$P[\text{špatný}\|k]$	37,63	64,41	84,45	94,22	98,00	99,32
oček. poč. n.	0,0876	0,1144	0,1344	0,1442	0,1480	0,1493

V jedné vědomostní soutěži bylo hlavní výhrou Ferrari 599 GTB Fiorano. Soutěžící, který se dostal do posledního kola, byl přiveden před tři stejná vrata. Podmínkou získání výhry bylo správně uhodnout, za kterými vraty se automobil nachází. Soutěžící jedna vrata označil a poté asistent otevřel ta z neoznačených vrat, za nimiž byla koza. Poslední soutěžní otázkou bylo, zda soutěžící chce svůj tip měnit.

Hodíme mincí. Pokud padne líc, dáme do krabice bílou kulečníkovou kouli, pokud padne rub, dáme tam kouli černou. To opakujeme n-krát. Potom poslepu vybereme z krabice jednu kouli a nevrátíme ji zpět. Tato vybraná koule je bílá. Určete pravděpodobnost, že další poslepu vybraná koule je černá.

Označme B_i jev „v plné krabici je i bílých koulí“ (zřejmě $i \in \{0, 1, 2, \ldots, n\}$), A jev „první vytažená koule je bílá“ a C jev „druhá vytažená koule je černá“. Jev B_i je vlastně jevem, že v sérii n hodů mincí padl líc i-krát, tedy

$$P(B_i) = \binom{n}{i} p^i (1-p)^{n-i}.$$

Podmíněná pravděpodobnost vytažení bílé koule za podmínky, že v krabici je právě i bílých koulí, je rovna

$$P(A|B_i) = \frac{i}{n}.$$

Zajímá nás pravděpodobnost jevu C když víme, že nastal jev A, tedy $P(C|A)$. Poněvadž jevy B_i jsou neslučitelné, jsou neslučitelné i jevy $C \cap B_i$. Současně platí $C = \bigcup_{i=0}^{n}(C \cap B_i)$ a toto sjednocení je disjunktní. Proto můžeme psát

$$P(C|A) = P\left(\bigcup_{i=0}^{n}(C \cap B_i)|A\right) = \sum_{i=0}^{n} \frac{P\big((C \cap B_i) \cap A\big)}{P(A)} =$$

$$= \frac{1}{P(A)} \sum_{i=0}^{n} P\big(C \cap (A \cap B_i)\big) =$$

$$= \frac{1}{P(A)} \sum_{i=0}^{n} P(A \cap B_i) P(C|A \cap B_i) =$$

$$= \frac{1}{P(A)} \sum_{i=0}^{n} P(B_i) P(A|B_i) P(C|A \cap B_i).$$

Za pravděpodobnost $P(A)$ můžeme ještě dosadit ze vzorce pro celkovou pravděpodobnost a dostaneme

$$P(C|A) = \frac{\sum_{i=0}^{n} P(B_i)P(A|B_i)P(C|A \cap B_i)}{P(A)} = \\ = \frac{\sum_{i=0}^{n} P(B_i)P(A|B_i)P(C|A \cap B_i)}{\sum_{i=0}^{n} P(B_i)P(A|B_i)}. \tag{3}$$

Tato formulka bývá někdy nazývána *2. Bayesův vzorec*; obecně platí za předpokladu, že prostor Ω je disjunktním sjednocením jevů B_i.

Ještě si uvědomíme, že podle zadání úlohy jsme alespoň jednou hodili mincí a tedy $n \geq 1$. Nyní můžeme vypočítat

$$\begin{aligned}
\sum_{i=0}^{n} P(B_i)P(A|B_i) &= \sum_{i=0}^{n} \binom{n}{i} p^i (1-p)^{n-i} \cdot \frac{i}{n} = \\
&= \sum_{i=1}^{n} \frac{(n-1)!}{(i-1)!(n-i)!} p^i (1-p)^{n-i} = \\
&= \sum_{i=0}^{n-1} \frac{(n-1)!}{i!(n-i-1)!} p^{i+1} (1-p)^{n-i-1} = \\
&= p \sum_{i=0}^{n-1} \binom{n-1}{i} p^i (1-p)^{n-1-i} = \\
&= p\big(p + (1-p)\big)^{n-1} = p,
\end{aligned}$$

$$\sum_{i=0}^{n} P(B_i)P(A|B_i)P(C|A\cap B_i) =$$

$$= \sum_{i=0}^{n} \binom{n}{i} p^i (1-p)^{n-i} \cdot \frac{i}{n} \cdot \frac{n-i}{n-1} =$$

$$= \sum_{i=1}^{n-1} \frac{(n-2)!}{(i-1)!(n-i-1)!} p^i (1-p)^{n-i} =$$

$$= \sum_{i=0}^{n-2} \frac{(n-2)!}{i!(n-2-i)!} p^{i+1} (1-p)^{n-i-1} =$$

$$= p(1-p) \sum_{i=0}^{n-2} \binom{n-2}{i} p^i (1-p)^{n-2-i} =$$

$$= \begin{cases} p(1-p), & n > 1 \\ 0, & n = 1, \end{cases}$$

takže po dosazení do druhého Bayesova vzorce dostaneme hledanou pravděpodobnost

$$P(C|A) = \begin{cases} 0, & n = 1, \\ 1 - p, & n > 1. \end{cases}$$