ANALÝZA A KLASIFIKACE DAT

prof. Ing. Jiří Holčík, CSc.

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

V.
KLASIFIKACE PODLE MINIMÁLNÍ VZDÁLENOSTI
pokračování

PRAVDĚPODOBNOSTNÍ METRIKY

- používají kompletní informaci o struktuře klasifikačních tříd danou pomocí podmíněných hustot pravděpodobnosti $p(x|\omega_1)$ a $p(x|\omega_2)$;
- metriky tohoto typu splňují následující podmínky:
  1. $J = 0$, pokud jsou si hustoty pravděpodobnosti rovny, tj.
     $$p(x|\omega_1) = p(x|\omega_2);$$
  2. $J \geq 0$
  3. J nabývá maxima, pokud jsou klasifikační třídy disjunktní, tj. $p(x|\omega_1) = 0$ a $p(x|\omega_2) \neq 0$ a naopak.

PRAVDĚPODOBNOSTNÍ METRIKY

základní myšlenka:

- klasifikační chyba:

$$e = \frac{1}{2}\left\{1 - \int \left|P(\omega_1|\mathbf{x}) - P(\omega_2|\mathbf{x})\right|.p(\mathbf{x})d\mathbf{x}\right\}$$

integrál v tomto vztahu

$$J_K = \int \left|P(\omega_1|\mathbf{x}) - P(\omega_2|\mathbf{x})\right|.p(\mathbf{x})d\mathbf{x} =$$

$$= \int \left|p(\mathbf{x}|\omega_1).P(\omega_1) - p(\mathbf{x}|\omega_2).P(\omega_2)\right|d\mathbf{x}$$

se nazývá Kolmogorovova variační vzdálenost

Chyba bude maximální, když integrand bude nulový, tj. když obě váhované funkce hustoty pravděpodobnosti budou totožné. Naopak chyba bude nulová, pokud se obě hustoty nebudou překrývat $\Rightarrow$ čím větší vzdálenost mezi třídami, tím je menší chyba klasifikace a naopak.

PRAVDĚPODOBNOSTNÍ METRIKY

Podobně jsou definovány další pravděpodobnostní míry vzdálenosti obecně vztahem

$$J(\mathbf{x}) = \int f\left[p(\mathbf{x}|\omega_i), P(\omega_i), i = 1,2\right] d\mathbf{x}$$

Chceme, aby J(**x**) byla nezáporná funkce, pro kterou je J(**x**)=0, když jsou obě hustoty pravděpodobnosti totožné a je maximální, když se obě hustoty nepřekrývají.

PRAVDĚPODOBNOSTNÍ METRIKY

Chernoffova vzdálenost

$$J_C = -\ln \int p^s(\mathbf{x}|\omega_1).p^{1-s}(\mathbf{x}|\omega_2)d\mathbf{x}, \quad s \in \langle 0;1 \rangle$$

Bhattacharyyova vzdálenost

$$J_B = -\ln \int \sqrt{p(\mathbf{x}|\omega_1).p(\mathbf{x}|\omega_2)}.d\mathbf{x}$$

Divergence

$$J_D = \int \left[p(\mathbf{x}|\omega_1) - p(\mathbf{x}|\omega_2)\right].\ln\left(\frac{p(\mathbf{x}|\omega_1)}{p(\mathbf{x}|\omega_2)}\right).d\mathbf{x}$$

Patrickova-Fisherova vzdálenost

$$J_{PF} = \sqrt{\int \left[p(\mathbf{x}|\omega_1) - p(\mathbf{x}|\omega_2)\right]^2.d\mathbf{x}}$$

PRAVDĚPODOBNOSTNÍ METRIKY

resp. tzv. zprůměrněné verze, zahrnující i apriorní pravděpodobnost jednotlivých klasifikačních tříd

zprůměrněná Chernoffova vzdálenost

$$J_C = -\ln \int \left[p(\mathbf{x}|\omega_1)P(\omega_1)\right]^s \left[p(\mathbf{x}|\omega_2)P(\omega_2)\right]^{1-s} d\mathbf{x}, \quad s \in \langle 0;1 \rangle$$

zprůměrněná Bhattacharyyova vzdálenost

$$J_B = -\ln \int \sqrt{p(\mathbf{x}|\omega_1).P(\omega_1).p(\mathbf{x}|\omega_2)P(\omega_2)}.d\mathbf{x}$$

zprůměrněná divergence

$$J_D = \int \left[p(\mathbf{x}|\omega_1).P(\omega_1) - p(\mathbf{x}|\omega_2 P(\omega_2))\right] \ln\left(\frac{p(\mathbf{x}|\omega_1).P(\omega_1)}{p(\mathbf{x}|\omega_2).P(\omega_2)}\right).d\mathbf{x}$$

zprůměrněná Patrickova-Fisherova vzdálenost

$$J_{PF} = \sqrt{\int \left[p(\mathbf{x}|\omega_1).P(\omega_1) - p(\mathbf{x}|\omega_2).P(\omega_2)\right]^2.d\mathbf{x}}$$

PRAVDĚPODOBNOSTNÍ METRIKY

uvedené výrazy se liší zejména pracností výpočtu a vazbou k hodnotám chybné klasifikace. Tato vazba je vyjádřena hodnotami D(**x**) a H(**x**) – dolním a horním odhadem pravděpodobnosti chybného zařazení.

$$H_C(\mathbf{x}) = \min_{0 \le s \le 1} J_c(s)$$

$$H_B(\mathbf{x}) = J_B$$

V případě, že známe dichotomické pravděpodobnostní míry a je třeba řešit problém klasifikace do více tříd, lze definovat kritérium, např. podle vztahu

$$J(\mathbf{x}) = \sum_{r=1}^{R} \sum_{q=r+1}^{R} P(\omega_r).P(\omega_q).J_{rq}(\mathbf{x})$$

PRAVDĚPODOBNOSTNÍ METRIKY

- základní nevýhodou pravděpodobnostních metrik je požadavek na znalost hustot pravděpodobnosti a jejich integrace (numerické?)
- za určitých předpokladů o typu rozložení mohou být tyto vztahy integrovány analyticky

PRAVDĚPODOBNOSTNÍ METRIKY

za předpokladu normálního rozložení ($\boldsymbol{\mu}_i$ jsou střední hodnoty a $\boldsymbol{\Sigma}_i$ kovarianční matice)

Chernoffova vzdálenost

$$J_C = \frac{1}{2}s(1-s)(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}_s^{-1}(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)+\frac{1}{2}\ln\left(\frac{|\boldsymbol{\Sigma}_s|}{|\boldsymbol{\Sigma}_s|^{1-s}|\boldsymbol{\Sigma}_s|^{s}}\right), \quad s\in\langle 0;1\rangle$$

kde $\boldsymbol{\Sigma}_s = (1\text{-}s).\boldsymbol{\Sigma}_1 + s.\boldsymbol{\Sigma}_2$

Bhattacharyyova vzdálenost je pro s=0,5

Divergence

$$J_D = \frac{1}{2}(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)^T(\boldsymbol{\Sigma}_1^{-1}+\boldsymbol{\Sigma}_2^{-1})(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)+\mathrm{Tr}(\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\Sigma}_2+\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2^{-1}-2\mathbf{I})$$

Patrickova-Fisherova vzdálenost $J_{PF} = \frac{1}{\sqrt{(2\pi)}}[|2\boldsymbol{\Sigma}_1|^{-\frac{1}{2}}+|2\boldsymbol{\Sigma}_2|^{-\frac{1}{2}}-$

$$-2|\boldsymbol{\Sigma}_1+\boldsymbol{\Sigma}_2|^{-\frac{1}{2}}.\exp\left\{-\frac{1}{2}(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)^T(\boldsymbol{\Sigma}_1+\boldsymbol{\Sigma}_2)^{-1}(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)\right\}]$$

PRAVDĚPODOBNOSTNÍ METRIKY

pokud $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$, pak se vztahy pro Bhattacharyyovu a divergenční vzdálenost zjednoduší na

$$J_M = J_D = 8J_B = (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1),$$

což je výraz pro Mahalanobisovu vzdálenost.

Příprava nových učebních materiálů oboru Matematická biologie

je podporována projektem ESF

č. CZ.1.07/2.2.00/07.0318

„VÍCEOBOROVÁ INOVACE STUDIA MATEMATICKÉ BIOLOGIE“

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ