Spolehlivost odhadů v regresních modelech

Radim Demut

26. května 2011

Školitel: Martin Holeňa

Máme soubor regresních modelů a chceme na jejich základě určit predikci pro konkrétní objekt.

- Jak zkombinovat výsledky jednotlivých modelů?
- Jak určit spolehlivost jednotlivých modelů?
- Je známo, že každý model má jinou spolehlivost pro jiný objekt. Jak tedy určit spolehlivost pro konkrétní objekt?

Odhad spolehlivosti:

- "Hold-out" estimate: Rozdělíme data na testovací a trénovací a náš odhad chyby bude chyba na testovací množině
- Cross validace
- Konfidenční intervaly
- Heuristiky
- Konformní predikce

Rozdíl induktivní (off-line) a transduktivní (on-line) přístup:

- V případě induktivního přístupu vytvoříme nejdříve obecné pravidlo (indukce), pomocí něhož pak rozhodujeme (dedukce)
- V případě transduktivního přístupu se snažíme jít přímo od starých příkladů k predikci o novém objektu
- Rozdíl především v praktickém použití
- Induktivní přístup: odhad parametrů ve statistice
- Transduktivní přístup: odhad pomocí nejbližších sousedů

Předpokládáme, že dostáváme páry $(x_1, y_1), (x_2, y_2), \ldots$, které nazýváme příklady

$$x_i \in \mathbf{X} \text{ - objekt} \qquad y_i \in \mathbf{Y} \text{ - závisle proměnná.}$$

Označme:
$\mathbf{Z} := \mathbf{X} \times \mathbf{Y}$ a $z_i := (x_i, y_i)$
$\mathbf{Z}^*$ ... množina všech n-tic ze $\mathbf{Z}$

Páry $(x_1, y_1), (x_2, y_2), \ldots$ jsou generované náhodně a nezávisle z nějakého rozdělení Q na Z, které neznáme.

Jednoduchý prediktor je měřitelná funkce

$$D : \mathbf{Z}^* \times \mathbf{X} \to \mathbf{Y}.$$

Pro jakoukoliv posloupnost vzorů $(x_1, y_1), \ldots, (x_{n-1}, y_{n-1})$ a nový objekt x_n nám dá predikci $y_n \in \mathbf{Y}$.

Nyní ale budeme chtít predikovat podmnožiny $\mathbf{Y}$ dost velké na to, abychom si mohli být celkem jistí, že y_n leží v této množině.

Konfidenční prediktor je měřitelná funkce

$$\Gamma : \mathbf{Z}^* \times \mathbf{X} \times (0,1) \to 2^{\mathbf{Y}},$$

která pro každý konfidenční parametr $1 - \varepsilon$ dá podmnožinu prostoru $\mathbf{Y}$

$$\Gamma^{\varepsilon}(x_1, y_1, \ldots, x_{n-1}, y_{n-1}, x_n),$$

která se zmenšuje s rostoucím ε, tzn.

$$\Gamma^{\varepsilon_1}(x_1, y_1, \ldots, x_{n-1}, y_{n-1}, x_n) \subseteq \Gamma^{\varepsilon_2}(x_1, y_1, \ldots, x_{n-1}, y_{n-1}, x_n)$$

kdykoliv $\varepsilon_1 \geq \varepsilon_2$.

Míra nekonformity je měřitelné zobrazení

$$A : \mathbf{Z}^* \times \mathbf{Z} \to \overline{\mathbb{R}}.$$

Někdy je vhodné uvažovat míru nekonformity pro množinu příkladů velikosti n jako

$$A_n : \mathbf{Z}^{n-1} \times \mathbf{Z} \to \overline{\mathbb{R}},$$

tedy jako zúžení A na $\mathbf{Z}^{n-1} \times \mathbf{Z}$.

Pokud máme danou míru nekonformity A_n a množinu příkladů $z_1, \ldots, z_n$, můžeme spočítat skóre

$$\alpha_i := A_n(\{z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_n\}, z_i),$$

které říká, jak moc z_i nezapadá mezi ostatní příklady.

Numerická hodnota α_i nám sama o sobě mnoho neřekne, proto se zavádí p-hodnota

$$\frac{|\{j = 1, \ldots, n : \alpha_j \geq \alpha_i\}|}{n},$$

která leží mezi $1/n$ a 1.

Konformní prediktor definovaný mírou nekonformity A je konfidenční prediktor Γ, kde

$$\Gamma^{\varepsilon}(x_1, y_1, \ldots, x_{n-1}, y_{n-1}, x_n),$$

je množina všech závisle proměnných $y \in \mathbf{Y}$ takových, že

$$\frac{|\{i = 1, \ldots, n : \alpha_i(y) \geq \alpha_n(y)\}|}{n} > \varepsilon,$$

kde

$$\begin{aligned} \alpha_i(y) &:= A(\{(x_1, y_1), \ldots, (x_{i-1}, y_{i-1}), (x_{i+1}, y_{i+1}), \ldots, (x_{n-1}, y_{n-1}), \\ &\quad (x_n, y)\}, (x_i, y_i)), \quad \forall i = 1, \ldots, n-1, \\ \alpha_n(y) &:= A(\{(x_1, y_1), \ldots, (x_{n-1}, y_{n-1})\}, (x_n, y)). \end{aligned}$$

Nechť D je jednoduchý prediktor

$$D : \mathbf{Z}^* \times \mathbf{X} \to \mathbf{Y}.$$

Pokud navíc máme nějakou míru vzdálenosti Δ

$$\Delta : \mathbf{Y} \times \mathbf{Y} \to \mathbb{R}$$

můžeme hodnoty $\alpha_i(y)$ a $\alpha_n(y)$ definovat jako

$$\alpha_i(y) := \Delta(y_i, D_{\{(x_1,y_1),\ldots,(x_{n-1},y_{n-1}),(x_n,y)\}}(x_i)).$$

$$\alpha_n(y) := \Delta(y, D_{\{(x_1,y_1),\ldots,(x_{n-1},y_{n-1}),(x_n,y)\}}(x_n))$$

nebo

$$\alpha_i(y) := \Delta(y_i, D_{\{(x_1,y_1),\ldots,(x_{i-1},y_{i-1}),(x_{i+1},y_{i+1}),\ldots,(x_{n-1},y_{n-1}),(x_n,y)\}}(x_i))$$

$$\alpha_n(y) := \Delta(y, D_{\{(x_1,y_1),\ldots,(x_{n-1},y_{n-1})\}}(x_n)).$$

Theorem

Pokud jsou příklady generovány náhodně a nezávisle z nějakého rozdělení P na $\mathbf{Z}$, potom pravděpodobnosti chyb jednotlivých predikcí jsou rovněž nezávislé a menší nebo rovné ϵ. V případě vyhlazeného konformního prediktoru jsou rovné ϵ.

Vyhlazený konformní prediktor je definovaný pomocí množiny $y \in \mathbf{Y}$

$$\frac{|\{i = 1, \ldots, n : \alpha_i(y) > \alpha_n(y)\}| + \tau_n |\{i = 1, \ldots, n : \alpha_i(y) = \alpha_n(y)\}|}{n} > \varepsilon,$$

kde τ_n jsou nezávislé náhodné veličiny rovnoměrně rozdělené na $[0, 1]$.

Pro klasifikaci chceme pro dané P a ε:

- Aby klasifikátor měl pro každou predikci pravděpodobnost chyby nejvýše ε
- Aby klasifikátor dával asymptoticky co nejméně vícenásobných predikcí pro dané P a ε
- Aby klasifikátor dával asymptoticky co největší množství prázdných predikcí pro dané P a ε, při splnění předchozího bodu

Takovýto klasifikátor se dá univerzálně zkonstruovat.

Hřebenová regrese

Uvažujme nyní lineární regresi, kde minimalizujeme součet čtverců

$$a||w||^2 + \sum_{i=1}^{n}(y_i - w^T x_i)^2.$$

Predikce pro objekty x_i potom je

$$\widehat{Y}_n := (\widehat{y}_1, \ldots, \widehat{y}_n)^T = \boldsymbol{X}_n(\boldsymbol{X}_n^T\boldsymbol{X}_n + a\boldsymbol{I}_d)^{-1}\boldsymbol{X}_n^T Y_n.$$

Definujme matici

$$\boldsymbol{H}_n := \boldsymbol{X}_n(\boldsymbol{X}_n^T\boldsymbol{X}_n + a\boldsymbol{I}_d)^{-1}\boldsymbol{X}_n^T$$

Míru nekonformity definujeme pomocí $\alpha_i := |e_i| = |y_i - \widehat{y}_i|$. Vektor $(\alpha_1, \ldots, \alpha_n)^T$ můžeme psát

$$(\alpha_1, \ldots, \alpha_n)^T = |Y_n - \boldsymbol{H}_n Y_n| = |(\boldsymbol{I}_n - \boldsymbol{H}_n) Y_n|.$$

Pro $Y := (y_1, \ldots, y_{n-1}, y)^T$ můžeme psát $(\alpha_1, \ldots, \alpha_n)^T$ jako

$$|A + By| := (\boldsymbol{I}_n - \boldsymbol{H}_n)(y_1, \ldots, y_{n-1}, 0)^T + (\boldsymbol{I}_n - \boldsymbol{H}_n)(0, \ldots, 0, 1)^T y$$

Odsud vidíme, že

- $\alpha_i = \alpha_i(y)$ se může měnit pouze po částech lineárně se změnou y.
- p-hodnota $p(y)$ se může měnit pouze v bodech, kde $\alpha_i(y) - \alpha_n(y)$ mění znaménko

Definujme pro $i = 1, \ldots, n$ množinu

$$S_i := \{y : \alpha_i(y) \geq \alpha_n(y)\} = \{y : |a_i + b_i y| \geq |a_n + b_n y|\},$$

kde a_i a b_i jsou složky A a B. Pro S_i mohou nastat tyto případy:

- pokud $b_i \neq b_n$, potom $\alpha_i(y) = \alpha_n(y)$ v bodech

$$-\frac{a_i - a_n}{b_i - b_n} \quad \text{a} \quad -\frac{a_i + a_n}{b_i + b_n},$$

- pokud $b_i = b_n \neq 0$ a $a_i \neq a_n$, potom $\alpha_i(y) = \alpha_n(y)$ v bodě

$$-\frac{a_i + a_n}{2b_i}$$

- pokud $b_i = b_n \neq 0$ a $a_i = a_n$, potom $S_i = \mathbf{R}$
- pokud $b_i = b_n = 0$, potom S_i je buď $\emptyset$ nebo $\mathbf{R}$

P-hodnotu tedy počítáme

$$p(y) = \frac{|\{i = 1, \ldots, n : y \in S_i\}|}{n}.$$

Můžeme to provést:

- seřadíme body, kde $\alpha_i(y) = \alpha_n(y)$, podle velikosti jako $y_{(1)}, \ldots, y_{(m)}$
- přidáme $y_{(0)} := -\infty$ a $y_{(m+1)} := \infty$
- spočteme $N(j)$, počet i, že $(y_{(j)}, y_{(j+1)}) \subseteq S_i$ pro $j = 0, \ldots, m$
- a $M(j)$, počet i, že $y_{(j)} \in S_i$ pro $j = 1, \ldots, m$
- $\Gamma_n^\varepsilon := \cup\{(y_{(j)}, y_{(j+1)}) : N(j)/n > \varepsilon\} \cup \{y_{(j)} : M(j)/n > \varepsilon\}$

Induktivní konformní prediktor (ICP)

Mějme ostře rostoucí posloupnost $m_1 < m_2 < \ldots$
Induktivní konformní prediktor Γ je definován:

- pokud $n \leq m_1$, potom $\Gamma^\varepsilon(x_1, y_1, \ldots, x_{n-1}, y_{n-1}, x_n)$ nalezneme pomocí běžného konformního prediktoru
- jinak najdeme k takové, že $m_k < n \leq m_{k+1}$ a

$$\Gamma^\varepsilon := \left\{ y \in \mathbf{Y} : \frac{|\{j = m_k + 1, \ldots, n : \alpha_j \geq \alpha_n\}|}{n - m_k} > \varepsilon \right\},$$

kde

$$\alpha_j(y) := A_{m_k+1}(\{(x_1, y_1), \ldots, (x_{m_k}, y_{m_k})\}, (x_j, y_j))$$

$$\alpha_n(y) := A_{m_k+1}(\{(x_1, y_1), \ldots, (x_{m_k}, y_{m_k})\}, (x_n, y))$$

Speciálně můžeme pro jednoduchý prediktor D a vzdálenost Δ psát

$$\alpha(y) := \Delta(y, D_{\{(x_1,y_1),\ldots,(x_{m_k},y_{m_k}),(x,y)\}}(x))$$

nebo

$$\alpha(y) := \Delta(y, D_{\{(x_1,y_1),\ldots,(x_{m_k},y_{m_k})\}}(x)).$$

Ve druhém případě nám tedy stačí počítat jednoduchý prediktor pouze při přechodu k dalšímu m_k v posloupnosti.

Příklady, jak definovat míru nekonformity:

- Míra nekonformity založená na 1-nearest neighbor

$$A(\{(x_1, y_1), \ldots, (x_l, y_l)\}, (x, y)) = \frac{\min_{i=1,\ldots,l:y_i=y} d(x, x_i)}{\min_{i=1,\ldots,l:y_i\neq y} d(x, x_i)}$$

- Míra nekonformity pro neuronové sítě, kde o_y je pravděpodobnost, že x patří do třídy y

$$A(x, y) = \frac{\sum_{y' \in \mathbf{Y}:y'\neq y} o_{y'}}{o_y + \gamma}$$

kde $\gamma \geq 0$ je vhodně zvolený parametr.

Pro případ Bayesova modelu, kterému úplně nevěříme, můžeme použít míru konformity v případě klasifikace

$$B(\{(x_1, y_1), \ldots, (x_l, y_l)\}, (x, y)) = p\{y\}$$

nebo v případě regrese

$$B(\{(x_1, y_1), \ldots, (x_l, y_l)\}, (x, y)) = \min(p((-\infty, y]), p([y, \infty))),$$

kde p je posteriorní pravděpodobnost za předpokladu x po tom, co jsme dostali $(x_1, y_1), \ldots, (x_l, y_l)$.

Vraťme se ke klasickému případu:

- Máme trénovací množinu $\{(x_1, y_1), \ldots, (x_n, y_n)\}$
- Máme nějaký model, který pomocí ní natrénujeme
- Chceme zjistit predikci pro x a odhadnou spolehlivost této predikce

Analýza citlivosti

- Spočteme predikci y objektu x
- Přiřadíme x závisle proměnnou $y + \varepsilon(l_{max} - l_{min})$, kde l_{min} a l_{max} jsou minimální a maximální hodnoty ze všech závisle proměnných z trénovací množiny a $\varepsilon \in E$
- Přidáme $(x, y + \varepsilon(l_{max} - l_{min}))$ do trénovací množiny, uděláme na této množině nový model a tímto modelem spočteme predikci y_ε objektu x

Citlivost na rozptyl je potom

$$\mathrm{SEvar}(x) := \frac{\sum_{\varepsilon \in E}(y_\varepsilon - y_{-\varepsilon})}{|E|}$$

a na vychýlení

$$\mathrm{SEbias}(x) := \frac{\sum_{\varepsilon \in E}(y_\varepsilon - y) + (y_{-\varepsilon} - y)}{2|E|}.$$

Rozptyl "bagged" modelu

- Z trénovací množiny vezmeme "bootstrap" výběry $L^{(i)}$, $i = 1, \ldots, m$ a na každém z nich uděláme nový model
- Každý z modelů dá predikci $K_i, i = 1, \ldots, m$ pro objekt x
- Závisle proměnná k objektu x je predikována jako průměr jednotlivých predikcí

$$K := \frac{\sum_{i=1}^{m} K_i}{m}.$$

Odhad spolehlivosti je definován jako rozptyl predikcí

$$\mathrm{BAGV}(x) := \frac{1}{m} \sum_{i=1}^{m} (K_i - K)^2.$$

Lokální křížová validace

- Vezmeme množinu k nejbližších sousedů x
  $N = \{(x_1, C_1), \ldots, (x_k, C_k)\}$
- Pro každé $(x_i, C_i) \in N$ vygenerujeme model M_i na $N \setminus \{(x_i, C_i)\}$
- Pak spočteme predikci K_i pro objekt x_i pomocí modelu M_i a spočítáme chybu $E_i = |C_i - K_i|$. Odhad spolehlivosti je vážený průměr lokálních chyb

$$\mathrm{LCV}(x) := \frac{\sum_{(x_i,C_i)\in N} \frac{1}{d(x_i,x)} E_i}{\sum_{(x_i,C_i)\in N} \frac{1}{d(x_i,x)}},$$

kde d je nějaká vzdálenost na $\mathbf{X}$.

Lokální model chyby

- Spočteme predikci y objektu x
- Vezmeme množinu k nejbližších sousedů x
  $N = \{(x_1, C_1), \ldots, (x_k, C_k)\}$
- Chybu vezmeme jako rozdíl průměru ze závisle proměnných nejbližších sousedů a predikce

$$\mathrm{CNK}(x) := \left| \frac{\sum_{i=1}^{k} C_i}{k} - y \right|.$$

Odhad spolehlivosti na základě hustoty

Odhad hustoty pro objekt x je

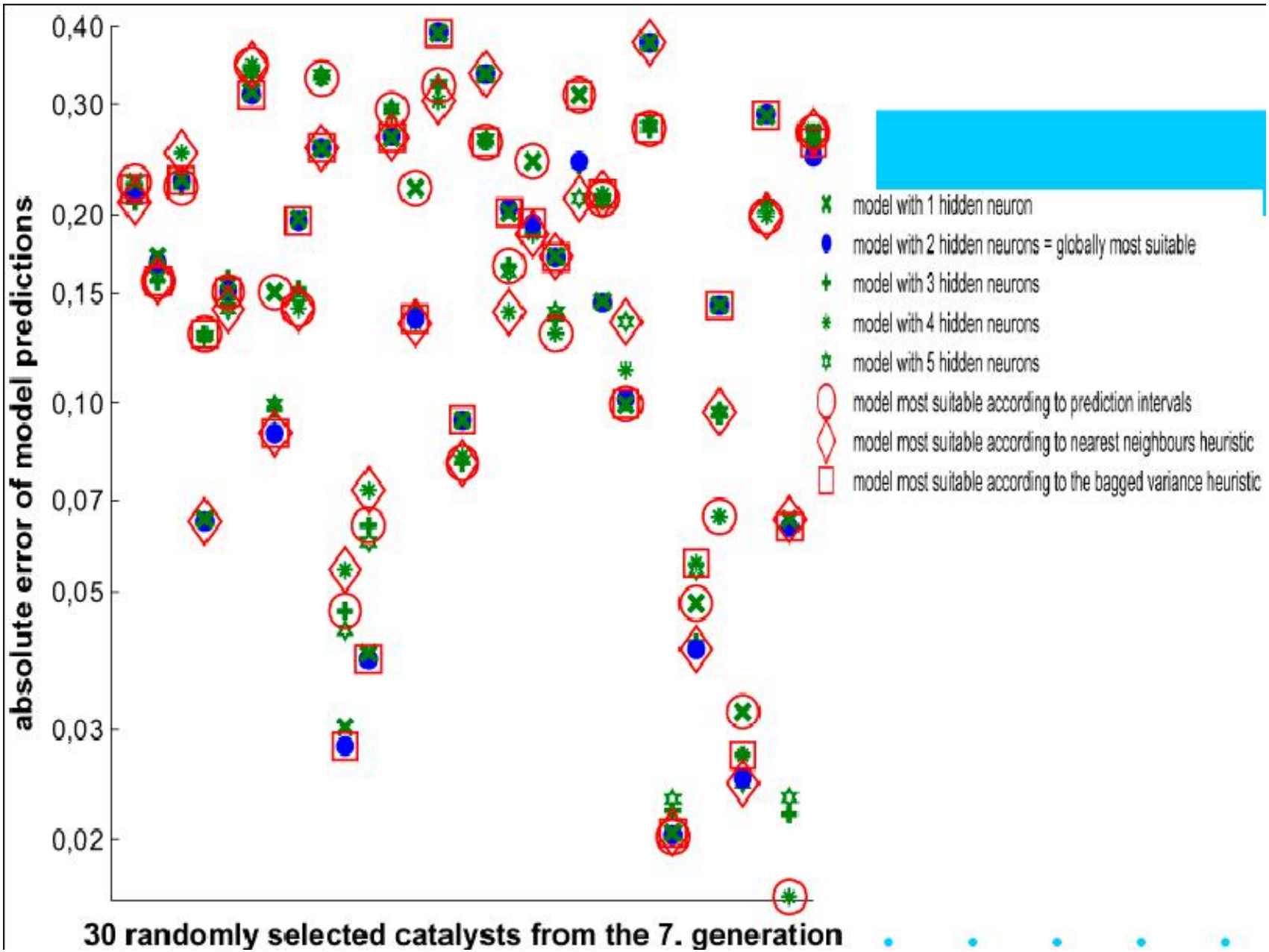
$$p(x) := \frac{\sum_{i=1}^{n} \kappa(d(x, x_i))}{n},$$

kde d značí nějakou vzdálenost na prostoru $\mathbf{X}$ a κ je jádrová funkce (můžeme volit například normální jádro).
Protože předpokládáme větší chybu při menší hustotě definujeme odhad spolehlivosti jako

$$\mathrm{DENS}(x) := \max_{i=1,\ldots,n} (p(x_i)) - p(x).$$

Data z chemického pokusu:

- Dvě diskrétní proměné a jedenáct spojitých proměnných
- Sedm generací
- V každé generaci 92 dat
- Prvních šest generací použito pro učení modelu, poslední pro testování
- Nejdříve rozclusterujeme podle diskrétních proměnných
- Potom na každém clusteru trénujeme RBF síť s různým počtem komponent


0,40
0,30
0,20
0,15
0,10
0,07
0,05
0,03
0,02
absolute error of model predictions
30 randomly selected catalysts from the 7. generation
model with 1 hidden neuron
model with 2 hidden neurons = globally most suitable
model with 3 hidden neurons
model with 4 hidden neurons
model with 5 hidden neurons
model most suitable according to prediction intervals
model most suitable according to nearest neighbours heuristic
model most suitable according to the bagged variance heuristic

V další práci se zaměříme na:

- Implementaci zde popsaných metod pro různé modely, speciálně ve FAKE-GAME.
- Pomocí simulací porovnáme různé metody pro kombinaci regresních modelů.
- Na základě simulací se pokusíme vylepšit stávající metody pro kombinaci regresních modelů.