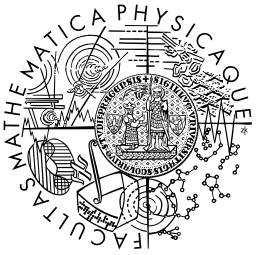


MATEMATICKO-FYZIKÁLNÍ FAKULTA
Univerzita Karlova

BAKALÁŘSKÁ PRÁCE

Michaela Krynická

Konformní predikce

Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: doc. RNDr. Matúš Maciak, Ph.D.
Studijní program: Finanční matematika

Praha 2023

Prohlašuji, že jsem tuto bakalářskou práci vypracoval(a) samostatně a výhradně s použitím citovaných pramenů, literatury a dalších odborných zdrojů. Tato práce nebyla využita k získání jiného nebo stejného titulu.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona v platném znění, zejména skutečnost, že Univerzita Karlova má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle §60 odst. 1 autorského zákona.

V ..... dne ....

Podpis autora

Děkuji panu doc. RNDr. Matúšovi Maciakovi, za jeho postřehy a podnětné rady. Dále děkuji svému Martinovi za to, že mi během psaní byl obrovskou oporou a za pomoc s programem R. Také bych ráda poděkovala Anše Vernerové, Erikovi a Radce za jejich bystré oči.

Název práce: Konformní predikce

Autor: Michaela Krynická

Katedra: Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: doc. RNDr. Matúš Maciak, Ph.D., Katedra pravděpodobnosti a matematické statistiky

Abstrakt: Hlavním cílem této práce je formalizovat koncept konformní predikce. Tato robustní, neparametrická metoda umožňuje konstrukci přesného predikčního intervalu na stanovené hladině, k čemuž stačí předpokládat, že vstupní data jsou nezávislá, stejně rozdělená. V kontextu náhodného výběru z jednorozměrného spojitého rozdělení vystavíme teoretické základy metody. Následně definujeme klíčový pojem míra nekonformity a prezentujeme algoritmické provedení, nejprve pro náhodný výběr, poté v kontextu regresní analýzy. V závěru práce porovnáváme na náhodně generovaných datech spolehlivost a efektivitu konformní predikce s konkrétní frekventistickou metodou.

Klíčová slova: konformní predikce, predikční interval, spolehlivost

Title: Conformal prediction

Author: Michaela Krynická

Department: Department of Probability and Mathematical Statistics

Supervisor: doc. RNDr. Matúš Maciak, Ph.D., Department of probability and mathematical statistics

Abstract: The main objective of this work is to formalize the concept of conformal prediction. This robust, nonparametric method allows the construction of an accurate prediction interval at a specified level, for which it is sufficient to assume that the input data are independent, equally distributed. In the context of random sampling from a one-dimensional continuous distribution, we expose the theoretical foundations of the method. Subsequently, we define the key concept of the degree of nonconformance and present the algorithmic design, first for random sampling and then in the context of regression analysis. At the end of the work, we compare the reliability and effectiveness of conformal prediction with a specific frequency method on randomly generated data.

Keywords: conformal prediction, prediction interval, confidence

Obsah

Úvod do konformní predikce

Predikce – předpověď budoucích výsledků na základě napozorovaných dat, je všudypřítomná součást našeho každodenního života. Od běžných záležitostí, jako je sledování předpovědi počasí, po komplexnější problémy, jako je odhadování účinnosti léků v rámci klinických studií.

Problematika predikce je v rámci statistiky dobře popsána primárně v kontextu *lineární regrese*. Lineární regrese zajišťuje platnost výsledků i při konečném počtu pozorování. Má ovšem příliš striktní předpoklady na to, aby bylo možné ji používat univerzálně. Metodou, která taktéž garantuje validní výsledky při konečném počtu pozorování je i *konformní predikce* (z anglického conformal prediction). Na rozdíl od lineární regrese ovšem postačuje předpokládat, že naše data jsou nezávislá, stejně rozdělená nebo dokonce pouze tzv. *zaměnitelná* (z angl. exchangeable). Konformní predikce je robustní, neparametrická metoda, jejíž teoretické základy popíšeme v této bakalářské práci.

Pojem konformní predikce byl poprvé použit v roce 1998 ve článku Gammerman, Vovk a Vapnik (1998) a v následujících letech postupně dále rozpracováván. Důležitým zdrojem je kniha Vovk, Gammerman a Shafer (2005), ze které vychází většina dalších publikací na toto téma. I přesto, že od první zmínky o konformní predikci již uplynulo téměř 25 let, tématem se stále zabývá pouze úzká skupina lidí. Inspirací pro tuto práci jsou zejména článek Shafer a Vovk (2008) a článek Fontana, Zeni a Vantini (2020). Oba tyto zdroje, stejně jako velká většina ostatních publikovaných článků, jsou populárně naučné a používají rozdílné značení. Hlavním cílem této práce je tedy zachytit klíčovou myšlenku metody konformní predikce, sjednotit používané značení a celý koncept matematicky formalizovat.

Vzhledem k minimálním předpokladům je možné konformní predikci použít pro téměř libovolný statistický model. Uplatňuje se v *regresních* i *klasifikačních* problémech, a to zejména v biomedicíně, zemědělství a ekologii. Dále také ve složitějších problémech *strojového učení*, například v rozpoznávání obličeje (z angl. face recognition). Tyto aplikace konformní predikce jsou mimo rozsah této práce, pro zájemce však doporučujeme knihu Balasubramanian, Ho a Vovk (2014).

Konformní predikce je způsob jak vytvořit predikční množinu pro následující pozorování, který funguje za minimálního předpokladu, že naše data jsou nezávislé, stejně rozdělené náhodné veličiny, respektive vektory. V kontextu této práce budeme zkoumat pouze specifický typ predikční množiny, a sice predikční interval. Budeme tedy pracovat s náhodnými veličinami, respektive vektory se spojitým rozdělením. Takto vzniklý predikční interval je validní na předem stanovené hladině a garantuje přesné pokrytí pro libovolné $n \in \mathbb{N}$, t.j. velikost vzorku napozorovaných dat.

V první kapitole práce vystavíme teoretické základy metody pro případ náhodného výběru z jednorozměrného spojitého rozdělení. Budeme vycházet primárně ze znalosti látky základního kurzu matematické statistiky. Ve druhé kapitole definujeme klíčové pojmy metody konformní predikce a prezentujeme algoritmické provedení. Popsat algoritmické provedení a demonstrovat fungování metody na jednoduchých příkladech. Ve čtvrté kapitole budeme prezentovat krátkou simulaci, kde

1. Teoretické základy

1.1 Formulace problému

V této kapitole budeme uvažovat pravděpodobnostní prostor $(\Omega, \mathcal{A}, \mathsf{P})$ a na něm reálnou náhodnou veličinu $X\colon (\Omega, \mathcal{A}) \to (\mathbb{R}, \mathcal{B})$, kde $\mathcal{B}$ značí borelovskou σ-algebru. Budeme pracovat s náhodným výběrem $\boldsymbol{X} = (X_1, \ldots, X_n)^\top$, kde konstantu $n \in \mathbb{N}$ nazýváme *rozsah náhodného výběru* a X_i je náhodná veličina s rozdělením $F_X \in \mathcal{F}$ pro předem stanovený statistický model $\mathcal{F}$.

Dále budeme uvažovat náhodnou veličinu X_{n+1}, která je nezávislá na náhodném výběru $\boldsymbol{X}$ a má stejné rozdělení F_X. Naším cílem bude na základě napozorovaných hodnot $X_1, \ldots, X_n$ sestavit interval, který obsahuje budoucí hodnotu náhodné veličiny X_{n+1} s určitou námi stanovenou pravděpodobností. Takový interval nazýváme *predikční interval*.

Definice 1. *(Predikční interval)*
Interval $B_n(\boldsymbol{X}) \subseteq \mathbb{R}$ nazveme predikčním intervalem pro náhodnou veličinu X_{n+1} v modelu $\mathcal{F}$ na hladině $1-\alpha$, pro $\alpha \in (0,1)$, právě když

$$\mathsf{P}[X_{n+1} \in B_n(\boldsymbol{X})] \geq 1 - \alpha.$$

Poznámka. Predikční interval na hladině $1-\alpha$ obsahuje budoucí hodnotu X_{n+1} s pravděpodobností $1-\alpha$. Na rozdíl od konfidenčního intervalu tedy nese informaci o hodnotě budoucího pozorování.

Poznámka. Dle tvaru intervalu rozlišujeme predikční intervaly oboustranné a jednostranné.

- Interval ve tvaru $(\eta_L(\boldsymbol{X}), \eta_R(\boldsymbol{X}))$, kde $\eta_L(\boldsymbol{X}) : \mathbb{R}^n \to \mathbb{R}$, $\eta_R(\boldsymbol{X}) : \mathbb{R}^n \to \mathbb{R}$ jsou měřitelná zobrazení splňující $\mathsf{P}[\eta_L(\boldsymbol{X}) < \eta_R(\boldsymbol{X})] = 1$ nazýváme *oboustranný predikční interval*. Obvykle jej konstruujeme tak, aby platilo

$$\mathsf{P}[X_{n+1} \leq \eta_L(\boldsymbol{X})] = \frac{\alpha}{2}, \quad \mathsf{P}[X_{n+1} \geq \eta_R(\boldsymbol{X})] = \frac{\alpha}{2}.$$

- Interval ve tvaru $(-\infty, \eta_R(\boldsymbol{X}))$ nazýváme *pravostranný (horní) predikční interval*. Platí $\mathsf{P}[X_{n+1} < \eta_R(\boldsymbol{X})] = 1 - \alpha$.

- Interval ve tvaru $(\eta_L(\boldsymbol{X}), \infty)$ nazýváme *levostranný (dolní) predikční interval*. Platí $\mathsf{P}[X_{n+1} > \eta_L(\boldsymbol{X})] = 1 - \alpha$.

Poznámka. Krajní body oboustranného predikčního intervalu se s rostoucím rozsahem náhodného výběru $\boldsymbol{X}$ blíží k hodnotám kvantilů rozdělení F_X. Například pro normální rozdělení $F_X = N(\mu, \sigma^2)$ a hladinu $(1-\alpha)$ se pro $n \to \infty$ blíží krajní body predikčního intervalu k hodnotám $\left(u(\frac{\alpha}{2}), u(1-\frac{\alpha}{2})\right)$, kde $u(\frac{\alpha}{2})$ je $\frac{\alpha}{2}$-kvantil rozdělení $N(\mu, \sigma^2)$.

Poznámka. Predikčnímu intervalu, jehož skutečná hladina je větší než požadované $1-\alpha$, se říká *konzervativní*. Takový predikční interval považujeme za validní, je ovšem obecně širší (tedy méně informativní), než by bylo nutné.

Máme náhodný výběr $(X_1, \ldots, X_n)^\top$ z rozdělení F_X a na něm nezávislou náhodnou veličinu X_{n+1} se stejným rozdělením. Hledáme $B_n(\boldsymbol{X})$, t.j. predikční interval náhodné veličiny X_{n+1}, takový, aby pro předem zvolené $\alpha \in (0,1)$ platilo $\mathsf{P}[X_{n+1} \in B_n(\boldsymbol{X})] \geq 1 - \alpha$. V konkrétním případě, kdy F_X je normální rozdělení, má úloha relativně přímočaré řešení. Konstrukce predikčního intervalu pak vychází z faktu, že existuje tzv. *pivotální statistika*, t.j. náhodná veličina, která má přesně Studentovo t-rozdělení s $n - 1$ stupni volnosti. Podrobnější odvození ukážeme v následující sekci. Pokud ovšem rozdělení F_X neznáme, zdá se na první pohled konstrukce predikčního intervalu na základě *konečného* počtu pozorování $X_1, \ldots, X_n$ jako komplikovaný úkol. V třetí a čtvrté sekci této kapitoly definujeme pojmy, s jejichž pomocí následně ukážeme, že pro náhodný výběr ze spojitého rozdělení se jedná o (až překvapivě) jednoduchou úlohu.

1.2 Normální rozdělení

V této sekci uvažujme náhodný výběr $\boldsymbol{X} = (X_1, \ldots, X_n)^\top$ z rozdělení $N(\mu, \sigma^2)$, kde $\mu \in \mathbb{R}$ a $\sigma^2 > 0$ jsou neznámé parametry. Ekvivalentně, náhodný výběr $\boldsymbol{X}$ pochází z rozdělení $F_X \in \mathcal{F}$, kde $\mathcal{F} = \{N(\mu,\sigma^2),\ \mu \in \mathbb{R},\ \sigma^2 > 0\}$. Dále uvažujme náhodnou veličinu X_{n+1} nezávislou na $\boldsymbol{X}$, takovou, že $X_{n+1} \sim N(\mu, \sigma^2)$. Lze dokázat, že platí (viz Anděl, 2007)

$$\frac{X_{n+1} - \overline{X}_n}{\sqrt{S_n^2(1 + \frac{1}{n})}} \sim t_{n-1},$$

kde

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

je *výběrový průměr* náhodného výběru $\boldsymbol{X}$,

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2$$

je *výběrový rozptyl* náhodného výběru $\boldsymbol{X}$ a t_{n-1} značí Studentovo t-rozdělení s $n-1$ stupni volnosti. Z toho plyne, že přesný predikční interval pro budoucí hodnotu X_{n+1} na hladině $1-\alpha$ má tvar

$$B_n(\boldsymbol{X}) = \left(\overline{X}_n \pm \sqrt{S_n^2\left(1 + \frac{1}{n}\right)}t_{n-1}\left(1 - \frac{\alpha}{2}\right)\right),$$

kde $t_{n-1}\left(1 - \frac{\alpha}{2}\right)$ je $\left(1 - \frac{\alpha}{2}\right)$-tý kvantil t-rozdělení s $n-1$ stupni volnosti.

Poznámka. Z předpisu intervalu $B_n(\boldsymbol{X})$ ihned vidíme, že je centrovaný okolo bodového odhadu střední hodnoty $\widehat{\mu} = \overline{X}_n$ a bude o to širší, čím větší je hodnota $S_n^2 = \widehat{\sigma}^2$ odhadu rozptylu σ^2. Stojí za povšimnutí, že konfidenční interval o stejné spolehlivosti $1-\alpha$ pro neznámou střední hodnotu $\mu \in \mathbb{R}$ má v tomto případě tvar

$$\left(\overline{X}_n \pm \sqrt{\frac{S_n^2}{n}}t_{n-1}\left(1 - \frac{\alpha}{2}\right)\right).$$

Od $B_n(\boldsymbol{X})$ se liší pouze přičtením odhadu rozptylu S_n^2 pod odmocninou.

1.3 Neznámé rozdělení – asymptotická metoda

Uvažujme nyní náhodný výběr $\boldsymbol{X} = (X_1, \ldots, X_n)^\top$ z neznámého rozdělení F_X, kde $F_X \in \mathcal{F} = \{$všechna rozdělení na $\mathbb{R}$ se spojitou distribuční funkcí$\}$. Dále uvažujme náhodnou veličinu X_{n+1} nezávislou na $\boldsymbol{X}$ se stejným rozdělením F_X. Pro předem zvolené $\alpha \in (0,1)$ hledáme predikční interval $B_n(\boldsymbol{X})$ pro náhodnou veličinu X_{n+1} na hladině $1-\alpha$. Vyjdeme z vlastností kvantilů rozdělení F_X.

Definice 2. *Pro $\alpha \in (0,1)$ definujeme kvantilovou funkci rozdělení F_X jako $F_X^{-1}(\alpha) = \inf\{x \in \mathbb{R} : F_X(x) \geq \alpha\}$. Kvantilem rozdělení F_X na hladině α (též α-kvantilem) rozumíme číslo $u_x(\alpha) = F_X^{-1}(\alpha)$.*

Z definice plyne, že pro $(1-\alpha)$-kvantil rozdělení F_X platí

$$\mathsf{P}[X_{n+1} \leq u_x(1-\alpha)] \geq 1-\alpha. \tag{1.1}$$

Kdybychom tedy znali hodnotu $u_x(1-\alpha)$, ihned dostaneme konzervativní (horní) predikční interval náhodné veličiny X_{n+1} ve tvaru $(-\infty, u_x(1-\alpha))$. V případě, že je distribuční funkce F_X ryze rostoucí, je kvantilová funkce jejím inverzem a zřejmě platí

$$\mathsf{P}\left[X_{n+1} \leq u_x(1-\alpha)\right] = F_X\left(u_x(1-\alpha)\right) = F_X\left(F_X^{-1}(1-\alpha)\right) = 1-\alpha.$$

Horní predikční interval $(-\infty, u_x(1-\alpha))$ náhodné veličiny X_{n+1} je tedy dokonce přesný. Distribuční funkci F_X ani kvantilovou funkci F_X^{-1} (tudíž ani hodnoty kvantilů) však neznáme. Budeme tedy namísto nich pracovat s jejich empirickými odhady.

Definice 3. *(Kulich (2017)) Funkci $\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^n \mathbb{I}\{X_i \leq x\}, \quad x \in \mathbb{R}$ nazýváme empirická distribuční funkce náhodného výběru $X_1, \ldots, X_n$.*

Poznámka. Funkce $\widehat{F}_n$ je neklesající, zprava spojitá a po částech konstantní se skoky v bodech X_i. Hodnota $\widehat{F}_n(x)$ pro pevné $x \in \mathbb{R}$ udává, jaká část náhodných veličin z $\boldsymbol{X}$ nepřekročí hodnotu $x \in \mathbb{R}$.

Definice 4. *Pro $\alpha \in (0,1)$ definujeme empirický (též výběrový) kvantil rozdělení F_X na hladině α jako $\widehat{u}_n(\alpha) = \inf\left\{x \in \mathbb{R} : \widehat{F}_n(x) \geq \alpha\right\}$.*

Když v nerovnosti (1.1) nahradíme kvantil $u_x(1-\alpha)$ jeho empirickým odhadem $\widehat{u}_n(1-\alpha)$, dostáváme přibližný výsledek $\mathsf{P}[X_{n+1} \leq \widehat{u}_n(1-\alpha)] \approx 1-\alpha$. Ukazuje se, že pro distribuční funkci F_X rostoucí na nějakém okolí bodu $u_x(1-\alpha)$ je výběrový kvantil $\widehat{u}_n(1-\alpha)$ konzistentním odhadem $u_x(1-\alpha)$. Asyptoticky pak platí, že $\mathsf{P}[X_{n+1} \leq \widehat{u}_n(1-\alpha)] \to 1-\alpha$ pro $n \to \infty$.

Tvrzení 1. *Nechť $\alpha \in (0,1)$. Nechť $\boldsymbol{X} = (X_1, \ldots, X_n)^\top$ je náhodný výběr z rozdělení, které má distribuční funkci spojitou a rostoucí na nějakém okolí bodu $u_x(\alpha)$. Potom*

$$\widehat{u}_n(\alpha) \xrightarrow[n\to\infty]{P} u_x(\alpha).$$

Důkaz. Viz Kulich (2017, Věta 2.5)

□

Z Tvrzení 1 a z nerovnosti (1.1) plyne

$$\lim_{n\to\infty} \mathsf{P}[X_{n+1} \leq \widehat{u}_n(1-\alpha)] = 1-\alpha.$$

1.4 Neznámé rozdělení – přesná metoda

Stejně jako v předchozí sekci mějme náhodný výběr $\boldsymbol{X} = (X_1, \ldots, X_n)^\top$ z neznámého spojitého rozdělení F_X. Dále mějme náhodnou veličinu X_{n+1} nezávislou na $\boldsymbol{X}$ se stejným rozdělením F_X. Hledáme predikční interval $B_n(\boldsymbol{X})$ náhodné veličiny X_{n+1} na hladině $1 - \alpha$ pro předem zvolené $\alpha \in (0{,}1)$.

V předchozí sekci (1.3) jsme ukázali, že pro neznámou distribuční funkci F_X, která je spojitá na určitém intervalu, umíme pomocí empirického kvantilu jednoduše sestrojit *asymptotický* predikční interval. Hledáme však metodu, která umožňuje konstrukci *přesného* predikčního intervalu $B_n(\boldsymbol{X})$ na základě náhodného výběru $\boldsymbol{X}$ s *konečným* rozsahem, a to pro libovolné spojité[1] rozdělení F_X. V následujícím textu definujeme pojmy *uspořádaný náhodný výběr* a *pořadí*. Ukážeme, že pro náhodný výběr ze spojitého rozdělení tyto pojmy na konstrukci hledaného přesného predikčního intervalu $B_n(\boldsymbol{X})$ přirozeně vedou. Uvedeme dva různé způsoby konstrukce. Popíšeme ideu prvního z nich a druhý budeme prezentovat formou věty. V závěru sekce předvedeme fungování prvního způsobu na vlastním příkladu.

Ještě než přejdeme k definici, povšimněme si, že jelikož mají náhodné veličiny $X_1, \ldots, X_n$ spojité rozdělení a jsou nezávislé, platí

$$\mathsf{P}(X_i = X_j) = 0 \text{ pro } i, j \in \{1, \ldots, n\}, i \neq j.$$

Definice 5. *(Uspořádaný náhodný výběr a pořadí) (Kulich (2017))*

1. *Seřadíme-li všechny náhodné veličiny $X_1, \ldots, X_n$ od nejmenší po největší, získáme uspořádaný náhodný výběr*

   $$X_{(1)} < X_{(2)} < \cdots < X_{(n)}.$$

   Symbol $X_{(k)}$ označuje k-tou nejmenší hodnotu mezi pozorováními $X_1, \ldots, X_n$ a nazýváme jej k-tá pořádková statistika.

2. *Pořadím náhodné veličiny X_i v uspořádaném náhodném výběru $X_{(1)}, \ldots, X_{(n)}$ rozumíme přirozené číslo $R_i \in \{1, \ldots, n\}$ takové, že $X_i = X_{(R_i)}$.*

Značení. Celý uspořádaný náhodný výběr budeme značit $\boldsymbol{X}_{(\cdot)}$, tedy

$$\boldsymbol{X}_{(\cdot)} = (X_{(1)}, \ldots, X_{(n)})^\top.$$

Poznámka. (Empirický kvantil, pořádková statistika) (Kulich (2017))

- Náhodné veličiny $X_1, \ldots, X_n$ jsou měřitelná zobrazení $X_i \colon (\Omega, \mathcal{A}) \to (\mathbb{R}, \mathcal{B})$. Pro každé $\omega \in \Omega$ tedy řadíme jejich realizace.
- Empirická distribuční funkce $\widehat{F}_n$ je po částech konstantní se skoky v bodech $X_{(1)}, \ldots, X_{(n)}$. Z toho plyne, že empirický kvantil je vlastně vhodně vybraná pořádková statistika. Jelikož platí

  $$\widehat{F}_n(X_{(k)}) = \frac{k}{n} \quad \text{a} \quad \widehat{F}_n(X_{(k)} - h) < \frac{k}{n} \text{ pro } \forall h > 0,$$

[1] Připomeňme, že metoda konformní predikce funguje pro libovolné neznámé rozdělení, v kontextu této práce však uvažujeme pouze rozdělení spojitá.

dostáváme

$$\widehat{u}_n(\alpha) = X_{(k_\alpha)}, \text{ kde } k_\alpha = \lceil n\alpha \rceil .$$

Poznámka. (Pořadí) (Kulich (2017))

- Pořadí R_i jsou náhodné veličiny.
- Platí $\mathsf{P}(R_i = k) = \frac{1}{n}$ pro všechna $i, k \in \{1, \ldots, n\}$, tedy R_i má diskrétní rovnoměrné rozdělení na množině $\{1, \ldots, n\}$.
- $R_i = \sum_{j=1}^{n} \mathbb{I}\{X_j < X_i\} + 1 \quad$ pro každé $i \in \{1, \ldots, n\}$.

Konstrukce predikčního intervalu

Nejprve popíšeme jednodušší verzi konstrukce predikčního intervalu. Následně zformulujeme větu, která vede na alternativní způsob konstrukce. Oba způsoby vychází z Definice 5 a pozorování uvedených v poznámkách.

Mějme náhodný výběr $\boldsymbol{X}$. Seřadíme-li napozorované hodnoty $X_1, \ldots, X_n$ od nejmenší po největší, dostáváme uspořádaný náhodný výběr $\boldsymbol{X}_{(\cdot)}$. Pořadí náhodné veličiny X_i má diskrétní rovnoměrné rozdělení na množině $\{1, \ldots, n\}$. Je tedy stejně pravděpodobné, že dané X_i bude první nejmenší mezi hodnotami $X_1, \ldots, X_n$, jako že bude druhé nejmenší atd. Protože náhodná veličina X_{n+1} je nezávislá na náhodném výběru $\boldsymbol{X}$ a má stejné rozdělení F_X, náhodný vektor $(\boldsymbol{X}^\top, X_{n+1})^\top$ je náhodným výběrem z rozdělení F_X o rozsahu $n+1$. Zdůrazněme, že zatímco hodnoty náhodných veličin $X_1, \ldots, X_n$ známe, hodnota náhodné veličiny X_{n+1} je neznámá a naším cílem je zkonstruovat pro ni predikční interval. Z teoretického hlediska platí, že uvažujeme-li pořadí veličin v rámci náhodného výběru $(\boldsymbol{X}^\top, X_{n+1})^\top$, R_{n+1}, t.j. pořadí náhodné veličiny X_{n+1} v uspořádaném náhodném výběru $X_{(1)}, \ldots, X_n, X_{(n+1)}$, má diskrétní rovnoměrné rozdělení na množině $\{1, \ldots, n, n+1\}$. Tento poznatek je *klíčovou myšlenkou metody konformní predikce.*

Vyznačíme-li napozorované hodnoty $X_1, \ldots X_n$ na reálnou osu, rozdělíme ji na $n+1$ intervalů ve tvaru $\left(-\infty, X_{(1)}\right), \left(X_{(1)}, X_{(2)}\right), \ldots, \left(X_{(n)}, \infty\right)$. V každém z nich bude neznámá náhodná veličina X_{n+1} ležet s pravděpodobností $\frac{1}{n+1}$. Nabízí se tedy přímočarý způsob, jak sestrojit $B_n(\boldsymbol{X})$, t.j. predikční interval pro X_{n+1} na hladině $1-\alpha$, jako sjednocení k sousedících intervalů, kde
$k = \inf\left\{i \in \{1, \ldots, n, n+1\} : \frac{i}{n+1} \geq 1-\alpha\right\}$. Tímto způsobem lze přímočaře konstruovat jednostranné i oboustranné predikční intervaly.

Poznámka. Na první pohled může konstrukce působit překvapivě. Pomáhá představit si, že v případě, že je skutečné rozdělení např. normální, bude k intervalů takto vybraných „na kraji“ dohromady „hodně širokých“, zatímco k intervalů vybraných uprostřed bude dohromady „užších“.

Poznámka. Z konstrukce predikčního intervalu popsané výše je zřejmé, že platí

$$\mathsf{P}[X_{n+1} \in B_n(\boldsymbol{X})] \in [1-\alpha, 1-\alpha+1/(n+1))$$

Lemma 2. *Nechť je Z náhodná veličina s diskrétním rovnoměrným rozdělením na množině $\{1,\dots,n\}$, nechť dále $\alpha \in (0,1)$. Potom platí*

$$\mathsf{P}\Big(Z \le \lceil n\alpha \rceil\Big) \ge \alpha.$$

Důkaz. Nechť nejprve $n\alpha \in \mathbb{N}$. Potom $\alpha = \frac{k}{n}$ pro nějaké $k \in \{1,\dots,n-1\}$. Dostáváme

$$\mathsf{P}\left(Z \le n \cdot \frac{k}{n}\right) = \mathsf{P}\left(Z \le k\right) = k \cdot \frac{1}{n} = \alpha.$$

Nyní nechť $n\alpha \notin \mathbb{N}$. Potom $\frac{j-1}{n} < \alpha < \frac{j}{n}$ pro nějaké $j \in \{1,\dots,n-1,n\}$. Dostáváme

$$\mathsf{P}\left(Z \le \lceil n\alpha \rceil\right) = \mathsf{P}\left(Z \le n \cdot \frac{j}{n}\right) = \frac{j}{n} > \alpha.$$

□

Věta 3. *Nechť je $\boldsymbol{X} = (X_1,\dots,X_n)^\top$ náhodný výběr z neznámého spojitého rozdělení F_X. Nechť je X_{n+1} náhodná veličina nezávislá na $\boldsymbol{X}$ a má stejné rozdělení F_X. Nechť $X_{(k)}, k \in \{1,\dots,n\}$ označuje k-tou pořádkovou statistiku v náhodném výběru $\boldsymbol{X}$ a nechť $\alpha \in (0,1)$. Pak platí*

$$P\Big(X_{n+1} \le \tilde{u}_n(1-\alpha)\Big) \ge 1-\alpha, \tag{1.2}$$

pro $\tilde{u}_n(1-\alpha) := X_{(\lceil (n+1)(1-\alpha) \rceil)}$, kde dodefinujeme $X_{(n+1)} := \infty$.

Důkaz. Definujme funkci $R\colon \mathbb{R} \to \{1,\dots,n,n+1\}$,

$$R(x) = \sum_{i=1}^{n} \mathbb{I}\{X_i \le x\} + 1, \quad x \in \mathbb{R}.$$

Náhodná veličina $R(X_{n+1})$ udává pořadí náhodné veličiny X_{n+1} v náhodném výběru $(\boldsymbol{X}^\top, X_{n+1})^\top$. Jako taková má diskrétní rovnoměrné rozdělení na množině $\{1,\dots,n+1\}$. Z toho a z Lemmatu 2 plyne

$$\mathsf{P}\left(R(X_{n+1}) \le \lceil (n+1)(1-\alpha) \rceil\right) \ge 1-\alpha.$$

Což ekvivalentně znamená

$$\mathsf{P}\left(X_{n+1} \le X_{(\lceil (n+1)(1-\alpha) \rceil)}\right) \ge 1-\alpha.$$

Dle předpokladů věty je to už

$$\mathsf{P}\left(X_{n+1} \le \tilde{u}_n(1-\alpha)\right) \ge 1-\alpha.$$

□

Příklad. Uvažujme hodnoty v Tabulce 1.1. Jedná se o délku zobáku (z angl. bill length) a hloubku zobáků (z angl. bill depth) dvaceti tučňáků oslích (z angl. Gentoo) náhodně vybraných ze 124 tučňáků stejného druhu v datasetu „palmerpenguins“. Obě dvě veličiny uvádíme v mm. Dataset „palmerpenguins“ je volně dostupný ve statistickém programu R. V tomto příkladu budeme pracovat pouze s druhým sloupcem tabulky, t.j. s hloubkami zobáku tučňáků. První sloupec pak využijeme v navazujícím příkladu 3.2.

	délka zobáku	hloubka zobáku
1	46.2	14.9
2	50.0	15.2
3	49.5	16.2
4	51.1	16.5
5	46.3	15.8
6	49.0	16.1
7	46.1	13.2
8	50.2	14.3
9	45.3	13.8
10	49.8	16.8
11	47.6	14.5
12	44.5	14.7
13	47.5	14.0
14	44.4	17.3
15	49.1	14.8
16	46.8	15.4
17	48.7	15.1
18	49.6	16.0
19	47.8	15.0
20	49.3	—

Tabulka 1.1: Délka a hloubka zobáku v *mm* pro 20 náhodně vybraných tučňáků oslích z datasetu „palmerpenguins“.

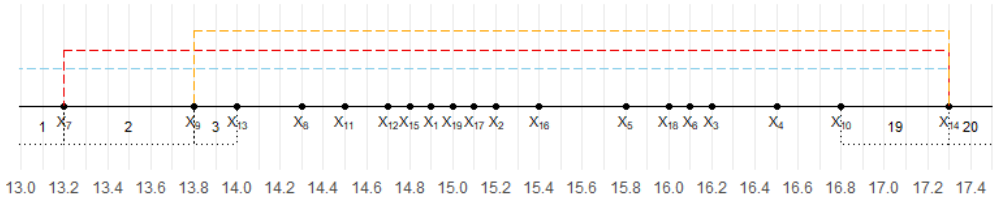
Předpokládáme, že hloubky zobáků vybraných tučňáků jsou nezávislé stejně rozdělené náhodné veličiny. Na základě 19 napozorovaných hodnot, označme je dle pořadí v tabulce jako $X_1, \ldots, X_{19}$, zkonstruujeme predikční interval pro neznámou hloubku zobáku dvacátého tučňáka, označme ji X_{20}.

Budeme postupovat způsobem popsaným v části „Konstrukce predikčního intervalu“, abychom demonstrovali jeho jednoduchost a intuitivní povahu. Vyznačíme napozorované hodnoty na reálnou osu a rozdělíme ji na 20 intervalů (viz naznačené očíslované intervaly na Obrázku 1.1). Pořadí náhodné veličiny X_{20} v náhodném výběru $X_1, \ldots, X_{20}$ má diskrétní rovnoměrné rozdělení na množině $\{1, \ldots, 20\}$. V každém z intervalů tedy neznámá náhodná veličina X_{20} leží s pravděpodobností $\frac{1}{20} = 0.05$. Postupně zkonstruujeme predikční interval (dále též PI) na hladinách $0.95, 0.90$ a 0.85.

- Pro $\alpha = 0.05$, konstruujeme predikční interval na hladině 0.95 jako sjednocení k sousedících intervalů, kde $k = \inf\left\{i \in \{1,\ldots,20\} : \frac{i}{20} \geq 0.95\right\} = 19$. Znamená to, že predikční interval bude kromě jediného vynechaného intervalu tvořit celá reálná osa. Vynechat můžeme buď interval nejvíce vpravo (označený číslem 20), nebo ten nejvíce vlevo (označený číslem 1). Je tedy zřejmé, že v tomto případě lze konstruovat pouze jednostranný PI. Když vynecháme interval nejvíce vpravo, dostaneme horní PI ve tvaru $(-\infty, X_{(19)}) = (-\infty, 17.3)$.

- V případě $\alpha = 0.10$ konstruujeme predikční interval pro hloubku zobáku dalšího náhodně vybraného tučňáka na hladině 0.90. Tentokrát bude PI sjednocením $k = \inf\left\{i \in \{1,\ldots,20\} : \frac{i}{20} \geq 0.90\right\} = 18$ sousedních intervalů. Vynecháváme tak dva z dvaceti vyznačených intervalů a je tedy zřejmé, že tentokrát lze zkonstruovat jednostranný i oboustranný predikční interval. Oboustranný PI je v tomto případě tvaru $(X_{(1)}, X_{(19)}) = (13.2, 17.3)$, horní predikční interval pak má tvar $(-\infty, X_{(18)}) = (-\infty, 16.8)$.

- Pro názornost prozkoumáme ještě případ $\alpha = 0.15$. PI na hladině 0.85 nyní konstruujeme jako sjednocení 17 sousedních intervalů. Konstrukce jednostranných intervalů je zřejmá. Případ oboustranného PI je však o něco zajímavější. V takovém případě máme dvě možnosti – zvolit interval tvaru $(X_{(1)}, X_{(18)}) = (13.2, 16.8)$, nebo interval tvaru $(X_{(2)}, X_{(19)}) = (13.8, 17.3)$. Délka prvního je 3.6, délka druhého 3.5. Zvolíme samozřejmě kratší, tedy více informativní interval. Dostáváme oboustranný predikční interval na hladině 0.85 ve tvaru $(X_{(2)}, X_{(19)}) = (13.8, 17.3)$.



Obrázek 1.1: Predikční intervaly pro hloubku zobáku (v mm) dalšího náhodně vybraného tučňáka oslího. Modře je vyznačen horní PI na hladině 0.95, červeně je vyznačen oboustranný PI na hladině 0.90 a oranžově je vyznačen oboustranný PI na hladině 0.85.

2. Algoritmické provedení

V sekci 1.4 jsme uvažovali náhodný výběr $\boldsymbol{X} = (X_1, \ldots, X_n)^\top$ z neznámého spojitého rozdělení F_X a dále náhodnou veličinu X_{n+1} nezávislou na $\boldsymbol{X}$ se stejným rozdělením F_X. Naším úkolem bylo na základě napozorovaného náhodného výběru $\boldsymbol{X}$ sestrojit predikční interval pro náhodnou veličinu X_{n+1} na dané hladině. To se nám přímočaře podařilo na základě uvědomění, že pořadí náhodné veličiny X_{n+1} v náhodném výběru $(\boldsymbol{X}^\top, X_{n+1})^\top = (X_1, \ldots, X_n, X_{n+1})^\top$ má diskrétní rovnoměrné rozdělení na množině $\{1, \ldots, n, n+1\}$. Aniž bychom si toho byli vědomi, použili jsme nejjednodušší verzi konformní predikce pro náhodný výběr z jednorozměrného spojitého rozdělení.

V případě, že však pracujeme se složitějším statistickým modelem, např. pozorujeme místo náhodných veličin náhodné vektory, nebo řešíme úlohu *klasifikace*[2], nelze napozorované hodnoty přirozeným způsobem seřadit. V tom případě není pojem *pořadí* dobře definovaný a postup ze sekce 1.4 nelze aplikovat. Pro složitější úlohy je tedy třeba metodu vhodně modifikovat. Principem takové modifikace je, že vhodným způsobem transformujeme napozorovaná data tak, aby je bylo možné seřadit a využít hlavní myšlenku konstrukce ze sekce 1.4.

V této kapitole definujeme klíčové pojmy konformní predikce a formulujeme algoritmus. Pro názornost budeme opět pracovat s náhodným výběrem z jednorozměrného spojitého rozdělení. Připomeňme, že pro takový případ již predikční interval sestrojit umíme (viz 1.4). Nyní se však na situaci podíváme z jiného úhlu pohledu. Používané značení je inspirováno článkem Fontana a kol. (2020), pro větší přehlednost však pro některé pojmy zavádíme vlastní značení.

2.1 Základní značení, definice

Pokud není specifikováno jinak, uvažujeme v celé kapitole (stejně jako v sekci 1.4) náhodný výběr $\boldsymbol{X} = (X_1, \ldots, X_n)^\top$ z neznámého rozdělení F_X, kde $F_X \in \mathcal{F} = \{\text{všechna rozdělení na } \mathbb{R} \text{ se spojitou distribuční funkcí}\}$. Dále uvažujme náhodnou veličinu X_{n+1} nezávislou na $\boldsymbol{X}$ se stejným rozdělením F_X. Pro předem zvolené $\alpha \in (0{,}1)$ hledáme predikční interval $B_n(\boldsymbol{X})$ náhodné veličiny X_{n+1} na hladině $1-\alpha$. Predikční interval budeme hledat tak, že vybereme $x \in \mathbb{R}$ a budeme ověřovat, zda platí $X_{n+1} = x$. Vybrané $x \in \mathbb{R}$ budeme nazývat *ověřovaná* hodnota.

[2](Převzato z diplomové práce Semela (2016)) Úloha klasifikace je studována v kontextu mnohorozměrné statistiky a strojového učení. Touto úlohou rozumíme situaci, kdy máme k dispozici několik objektů (tzv. trénovací množinu), které náleží do právě jedné z předem specifikovaných disjunktních tříd. Zároveň na každém z objektů pozorujeme určitou sadu znaků. Cílem klasifikační analýzy je na základě těchto pozorování stanovit tzv. rozhodovací pravidlo (nazývané také jako klasifikátor), které umožní jednoznačně zařadit do jedné z uvažovaných tříd objekt s libovolnou sadou znaků.

Klíčovou roli v algoritmu konformní predikce hraje tzv. *míra nekonformity* (z angl. nonconformity measure). Jedná se o funkci, která vhodným způsobem přiřadí napozorovaným datům a hodnotě $x \in \mathbb{R}$ tzv. *skóre nekonformity* (z angl. nonconformity score). Skóre nekonformity pak lze přirozeným způsobem seřadit a porovnávat. V kontextu složitějších úloh o nich tedy můžeme přemýšlet jako o zástupných hodnotách pro napozorovaná data a ověřovanou hodnotu $x \in \mathbb{R}$, pro které lze přirozeně definovat pořadí.

Značení. Ve zbytku této kapitoly budeme používat následující značení

$$M := (X_1, \ldots, X_n, x), \quad x \in \mathbb{R}.$$

Definice 6. *(Míra nekonformity a skóre nekonformity)*
Uvažujme náhodný výběr $\boldsymbol{X} = (X_1, \ldots, X_n)^\top$, $x \in \mathbb{R}$ *a množinu* M, $M = (X_1, \ldots, X_n, x)$. *Míra nekonformity prvku* x *vzhledem k množině* M, *je libovolné zobrazení*

$$A(M, x) \colon \mathbb{R}^{n+1} \times \mathbb{R} \to \mathbb{R}.$$

Hodnotu

$$r_x := A(M,x)$$

nazýváme skóre nekonformity prvku $x \in \mathbb{R}$ *vzhledem k* M.

Míru nekonformity zavádíme primárně proto, abychom rozšířili pojem pořadí na obecnou úlohu. Zdůrazněme, že jelikož v této kapitole pracujeme s náhodným výběrem $\boldsymbol{X}$ z jednorozměrného spojitého rozdělení F_X, pořadí je za této situace přirozeně definované a míru nekonformity explicitně zavádět nepotřebujeme. V sekci 1.4 jsme nepřímo pracovali s mírou nekonformity $A(M,x) =...$

Zároveň pomocí míry nekonformity můžeme měřit, jak moc se dané $x \in \mathbb{R}$ liší od množiny napozorovaných hodnot, do které jsme zahrnuli x. Jinými slovy měříme, jak moc je x vzhledem k M *nekonformní*. Z tohoto pohledu by vhodná volba míry nekonformity mohla být také např.

$$A(M,x) = \left| \frac{1}{n+1} (\sum_{i=1}^{n} X_i + x) - x \right|.$$

Za předpokladu $X_{n+1} = x$, se jedná o Euklidovskou vzdálenost bodu x od $\frac{1}{n+1} \sum_{i=1}^{n+1} X_i = \overline{X}_{n+1}$, t.j. výběrového průměru náhodného výběru $(\boldsymbol{X}^\top, X_{n+1})^\top$ a zároveň bodového odhadu střední hodnoty náhodné veličiny X_{n+1}. Poznamenejme ještě, že pojem množina v Definici 6 používáme proto, abychom zdůraznili, že *nezáleží na pořadí* ve kterém hodnoty $X_1, \ldots, X_n, x$ vstupují do funkce A.

Poznámka. (Skóre nekonformity)

- Čím vyšší je hodnota r_x, tím více se x liší od množiny M.
- O r_x můžeme v jistém smyslu uvažovat jako o reziduu daného $x \in \mathbb{R}$.

2.2 Algoritmus

Uvažujeme náhodný výběr $\boldsymbol{X} = (X_1, \ldots, X_n)^\top$ z neznámého spojitého rozdělení F_X. Dále uvažujme náhodnou veličinu X_{n+1} nezávislou na $\boldsymbol{X}$ se stejným rozdělením F_X a míru nekonformity $A : \mathbb{R}^{n+1} \times \mathbb{R} \to \mathbb{R}$. Hledáme predikční interval $B_n(\boldsymbol{X})$ náhodné veličiny X_{n+1} na hladině $1 - \alpha$ pro předem stanovené $\alpha \in (0,1)$. Idea algoritmu konformní predikce spočívá v tom, že pro každé $x \in \mathbb{R}$ ověříme, zda patří do predikčního intervalu $B_n(\boldsymbol{X})$. Je zřejmé, že provést nespočetně mnoho takových pokusů nelze. V naší situaci (a obecně vždy, když konstruujeme predikční interval pro náhodnou veličinu se spojitým[3] rozdělením), proto samotný výpočet spoléhá na nějaký dodatečný princip. Příklad takového principu naznačíme v příkladu 3.2.

Značení. Pro názornost budeme značit

$$r_{n+1}^x := r_x = A(M, x), \quad x \in \mathbb{R},$$
$$r_i^x := r_{X_i} = A(M, X_i), \quad i \in \{1, \ldots, n\}.$$

Poznámka. Značení zavedené výše je intuitivnější, když si představíme, že pro vybrané $x \in \mathbb{R}$, pro které ověřujeme, zda patří do predikčního intervalu $B_n(\boldsymbol{X})$ „fiktivně dosadíme“ $X_{n+1} = x$. Pak pro každé $i \in \{1, \ldots, n, n+1\}$ píšeme $r_i^x = A(M, X_i)$.

Algoritmus 1: KP pro náhodný výběr

Vstup *Napozorovaný náhodný výběr* $\boldsymbol{X} = (X_1, \ldots, X_n)^\top$*, předem zvolená míra nekonformity* A*, hladina významnosti* $1 - \alpha$*, kde* $\alpha \in (0,1)$.

Výstup *Interval* $B_n(\boldsymbol{X})$*, t.ž.* $\mathsf{P}[X_{n+1} \in B_n(\boldsymbol{X})] \geq 1 - \alpha$.

Začátek

Krok 1: Vyber $x \in \mathbb{R}$.

Krok 2: Spočítej

$$r_{n+1}^x = A(M, x) = A((X_1, \ldots, X_n, x), x),$$
$$r_i^x = A(M, X_i) = A((X_1, \ldots, X_n, x), X_i), \quad i \in \{1, \ldots, n\}.$$

Krok 3: Spočítej

$$p(x) := \frac{\sum_{i=1}^{n+1} \mathbb{I}\{r_i^x \geq r_{n+1}^x\}}{n+1}.$$

Krok 4: Pokud $p(x) > \alpha$, zahrň x do $B_n(\boldsymbol{X})$.

Konec

---

[3] V případě klasifikace nabývá náhodná veličina, pro kterou kvantifikujeme nejistotu predikce, pouze malého počtu hodnot. V takovém případě konstruujeme predikční množinu. Označme tuto náhodnou veličinu jako Y a nechť má rozdělení F_Y. Nosič S_Y rozdělení F_Y je množina s malým počtem prvků, např. pro alternativní rozdělení – $F_Y = Alt(p), p \in (0,1)$, pouze dvouprvková, $S_Y = \{0,1\}$. V takové situaci je možné výpočet provést hrubou silou, tedy postupně po jednom pro každé $y \in S_Y$.

Poznámka. p je funkce $p : \mathbb{R} \to \{\frac{1}{n+1}, \frac{2}{n+1}, \ldots, 1\}$. Hodnota $p(x)$ pak bývá neformálně nazývána jako *p-hodnota* pro testování nulové hypotézy $H_0 : X_{n+1} = x$.

Jeden cyklus algoritmu tedy probíhá tak, že vybereme libovolné reálné x a ověřujeme, zda patří do $B_n(\boldsymbol{X})$, t.j. predikčního intervalu náhodné veličiny X_{n+1}. Pro zvolené $x \in \mathbb{R}$ spočítáme skóre nekonformity $r_1^x, r_2^x, \ldots, r_{n+1}^x$. Hodnota r_{n+1}^x udává, jak moc se vybrané $x \in \mathbb{R}$ liší od množiny M – množiny napozorovaných náhodných veličin $X_1, \ldots, X_n$, do které jsme navíc přidali tstovanou hodnotu $x \in \mathbb{R}$. Každá z hodnot r_i^x pak specifikuje, jak moc se od M liší napozorovaná hodnota X_i. Skóre nekonformity r_{n+1}^x, které je relativně velké vůči ostatním hodnotám r_i^x, indikuje, že vybrané $x \in \mathbb{R}$ má vzhledem k napozorovaným náhodným veličinám $X_1, \ldots, X_n$ neobvyklou hodnotu. Je intuitivní, že takové x do predikčního intervalu nezahrneme.

Idea třetího kroku algoritmu staví na principu využitém v odvození predikčního intervalu v sekci 1.4. Seřadíme si spočítaná skóre nekonformity od nejmenšího po největší a zjišťujeme pořadí hodnoty r_{n+1}^x mezi hodnotami $r_1^x, r_2^x, \ldots, r_{n+1}^x$. Pokud r_{n+1}^x patří mezi $1 - \alpha$ nejmenších hodnot, zahrneme x do predikčního intervalu $B_n(\boldsymbol{X})$. Ve formulaci algoritmu je jako rozhodovací pravidlo použita o něco méně intuitivní hodnota $p(x)$. Důvodem je neformální analogie ke klasické p-hodnotě v rámci testování hypotéz. Hodnotu $p(x)$ interpretujeme jako část těch skóre nekonformity $r_i^x, i \in \{1, \ldots, n, n+1\}$, která jsou alespoň tak velká jako r_{n+1}^x. Tzn., pokud je počet skóre nekonformity $r_i^x, i \in \{1, \ldots, n, n+1\}$, která jsou alespoň tak velká jako r_{n+1}^x striktně větší než $\alpha \cdot (n+1)$, x zahrneme do $B_n(\boldsymbol{X})$.

3. Rozšíření metody

V předchozí kapitole jsme definovali klíčové pojmy metody konformní predikce pro náhodný výběr z jednorozměrného spojitého rozdělení a následně formulovali algoritmus. V této kapitole rozšíříme metodu na složitější úlohu. Budeme se zabývat konformní predikcí v *regresní analýze*.

Poznámka. (Regresní analýza) (Semela (2016))
Pojem regresní analýza či regrese, používáme jako souhrnné označení pro zkoumání závislosti mezi tzv. *závisle proměnnou* náhodnou veličinou a jednou nebo více nezávislými náhodnými veličinami nazývanými jako *vysvětlující proměnné*.

Hlavním účelem kapitoly je představit metodu konformní predikce v úloze, která má (na rozdíl od ryze ilustrační úlohy ve druhé kapitole) uplatnění v aplikované statistice. V první sekci rozšíříme značení a definice zavedené v druhé kapitole do rámce regresní úlohy. Budeme se zabývat konstrukcí predikčního intervalu pro závisle proměnnou náhodnou veličinu se *spojitým* rozdělením. Ve druhé sekci formulujeme algoritmus a uvedeme některé další vlastnosti míry nekonformity. V závěru kapitoly demonstrujeme fungování algoritmu na příkladu, který navazuje na příklad 1.4.

3.1 Základní značení, definice

V celé kapitole uvažujeme nezávislé, stejně rozdělené náhodné vektory $(Y_i, \mathbf{X}_i^\top)^\top$, $i \in 1, \ldots, n$. Každé Y_i je náhodná veličina s rozdělením $F_{Y_i} \in \mathcal{F}$, $\mathcal{F}$ = všechna rozdělení na $\mathbb{R}$ se spojitou distribuční funkcí. Každé $\mathbf{X}_i$ je p-dimenzionální vektor $\mathbf{X}_i = (X_{i1}, \ldots X_{ip})^\top, p \in \mathbb{N}$.

Poznámka. (Terminologie)

- V souladu se všeobecně používanou terminologií, budeme náhodné veličiny Y_i nazývat *závislé proměnné* a elementům vektorů $\mathbf{X}_i$ (náhodným veličinám X_{ij}) budeme říkat *vysvětlující proměnné*.
- Článek Fontana a kol. (2020) a podobně další publikace popisují situaci ve větší obecnosti a nazývají náhodné veličiny Y_i jako tzv. *štítky* (z angl. label) a náhodné vektory $\mathbf{X}_i$ jako tzv. *objekty* (z angl. object).

Značení. Označme pro $i \in 1, \ldots, n$

$$\boldsymbol{Z}_i = \left(Y_i, \boldsymbol{X}_i^\top\right)^\top, \quad \boldsymbol{Z}_i \in \mathbb{R}^{p+1},$$

$$\boldsymbol{Z} = \left(\boldsymbol{Z}_1^\top, \ldots, \boldsymbol{Z}_n^\top\right)^\top, \quad \boldsymbol{Z} \in \mathbb{R}^{n \times (p+1)}.$$

Uvažujme nyní náhodný výběr $\boldsymbol{Z} = \left(\boldsymbol{Z}_1^\top, \ldots, \boldsymbol{Z}_n^\top\right)^\top$ z neznámého sdruženého rozdělení $F_{\boldsymbol{Z}}$. Dále uvažujme náhodný vektor $\boldsymbol{Z}_{n+1} = \left(Y_{n+1}, \boldsymbol{X}_{n+1}^\top\right)^\top$, nezávislý na $\boldsymbol{Z}$ se stejným rozdělením $F_{\boldsymbol{Z}}$. Naším úkolem bude na základě napozorovaných hodnot vektorů $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n$ a $\boldsymbol{X}_{n+1}$ sestavit predikční interval pro budoucí hodnotu náhodné veličiny Y_{n+1} na stanovené hladině $1 - \alpha$.

Značení. Pro přehlednost budeme značit

$$\mathcal{X} := (\boldsymbol{Z}^\top, \boldsymbol{X}_{n+1}^\top)^\top.$$

Definice 7. *(Predikční interval 2)*
Interval $B_n(\mathcal{X}) \subseteq \mathbb{R}$ *nazveme* predikčním intervalem *náhodné veličiny* Y_{n+1} *v modelu* $\mathcal{F}$ *na hladině* $1 - \alpha$, $\alpha \in (0{,}1)$, *právě když*

$$P[Y_{n+1} \in B_n(\mathcal{X})] \geq 1 - \alpha.$$

Poznámka. Všechny vlastnosti predikčního intervalu uvedené v poznámkách v první kapitole lze přirozeně rozšířit i do kontextu této kapitoly.

Značení. Ve zbytku této kapitoly budeme používat následující značení

$$\begin{aligned} \boldsymbol{z} &:= \left(\boldsymbol{X}_{n+1}^\top, y\right), \quad y \in \mathbb{R}, \\ M &:= (\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n, \boldsymbol{z}). \end{aligned}$$

Definice 8. *(Míra nekonformity a skóre nekonformity 2)*
Uvažujme náhodný výběr $\boldsymbol{Z} = \left(\boldsymbol{Z}_1^\top, \ldots, \boldsymbol{Z}_n^\top\right)^\top$, *náhodný vektor* $\boldsymbol{X}_{n+1}$, *vektor* $\boldsymbol{z} = \left(\boldsymbol{X}_{n+1}^\top, y\right) \in \mathbb{R}^{p+1}$, *kde* $y \in \mathbb{R}$, *a množinu* $M = (\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n, \boldsymbol{z})$. *Míra nekonformity pro* $\boldsymbol{z}$ *vzhledem k množině* M, *je libovolné zobrazení*

$$A(M, \boldsymbol{z}) \colon \mathbb{R}^{(n+1)\times(p+1)} \times \mathbb{R}^{p+1} \to \mathbb{R}.$$

Hodnotu

$$r_{\boldsymbol{z}} := A(M, \boldsymbol{z})$$

nazýváme skóre nekonformity vektoru $\boldsymbol{z} = \left(\boldsymbol{X}_{n+1}^\top, y\right)$ *vzhledem k* M.

Poznámka. (Míra nekonformity)
Obecně závisí vhodná volba míry nekonformity na povaze řešené úlohy. Zpravidla však platí $A(M, \boldsymbol{z}) = A\left(M, \left(\boldsymbol{X}_{n+1}^\top, y\right)\right) = |y - \widehat{\theta}(M)|$, kde $\widehat{\theta}(M)$ značí nějaký *bodový odhad* náhodné veličiny Y_{n+1} založený na množině M.

Značení. Označme

$$\begin{aligned} r_{n+1}^y &:= r_{\boldsymbol{z}} = A(M, \boldsymbol{z}), \\ r_i^y &:= r_{\boldsymbol{Z}_i} = A(M, \boldsymbol{Z}_i), \quad i \in 1, \ldots, n. \end{aligned}$$

3.2 Algoritmus

Uvažujme značení zavedené v předchozí sekci. Algoritmus konformní predikce v regresní úloze formulujeme následovně.

Algoritmus 2: KP v regresní úloze

Vstup *Napozorovaný náhodný výběr* $\boldsymbol{Z} = \left(\boldsymbol{Z}_1^\top, \ldots, \boldsymbol{Z}_n^\top\right)^\top$, *napozorovaný náhodný vektor* $\boldsymbol{X}_{n+1}$, *předem zvolená míra nekonformity* A, *hladina významnosti* $1-\alpha$, *kde* $\alpha \in (0{,}1)$.

Výstup *Interval* $B_n(\mathcal{X})$, *t.ž.* $\mathsf{P}[Y_{n+1} \in B_n(\mathcal{X})] \geq 1-\alpha$.

Začátek

Krok 1: Vyber $y \in \mathbb{R}$.

Krok 2: Spočítej

$$\begin{aligned} r_{n+1}^y &= A(M,\boldsymbol{z}), \\ r_i^y &= A(M, \boldsymbol{Z}_i), \quad i \in 1, \ldots, n. \end{aligned}$$

Krok 3: Spočítej

$$p(y) := \frac{\sum_{i=1}^{n+1} \mathbb{I}\{r_i^y \geq r_{n+1}^y\}}{n+1}.$$

Krok 4: Pokud $p(y) > \alpha$, zahrň y do $B_n(\mathcal{X})$.

Konec

Poznámka. Ať už jako míru nekonformity A zvolíme libovolnou funkci splňující Definici 8, algoritmus vždy vyprodukuje (konzervativní) predikční interval náhodné veličiny Y_{n+1} na hladině $1-\alpha$. Predikční interval však bude informativní (dostatečně malý) pouze v případě, že je míra nekonformity zvolená vhodně, tedy zejména, když je $\widehat{\theta}(M)$ rozumným bodovým odhadem náhodné veličiny Y_{n+1}. Situaci dobře ilustruje následující jednoduchý příklad.

Příklad. Uvažujme triviální míru nekonformity

$$A(M, \boldsymbol{z}) = 1, \quad M \subset \mathbb{R}^{n+1},\ \boldsymbol{z} \in \mathbb{R}^{p+1}.$$

Pak pro $\forall y \in \mathbb{R}$ platí

$$\begin{aligned} r_{n+1}^y &= 1, \\ r_i^y &= 1, \quad i \in 1, \ldots, n, \end{aligned}$$

a tedy

$$p(y) = 1.$$

Predikční interval je tedy celá reálná osa, t.j. $B_n(\mathcal{X}) = \mathbb{R}$. Zřejmě tedy splňuje $\mathsf{P}[Y_{n+1} \in B_n(\mathcal{X})] \geq 1-\alpha$, ale je zcela neužitečný.

Průběh algoritmu je analogický k průběhu popsanému v sekci 2.2. Jak již bylo zmíněno, zjevným problémem jeho implementace je, že nelze ověřovat pro každé $y \in \mathbb{R}$. Ukazuje se, že pro některé volby míry nekonformity lze provádění nespočetně mnoha cyklů algoritmu obejít. Například v případě, kdy ke konstrukci míry nekonformity použijeme bodové odhady sestrojené metodou *lineární regrese*. Průběh algoritmu pro takový případ ilustrujeme na následujícím příkladu. Předpokládáme základní znalost lineárního modelu.

Příklad. Navážeme na příklad 1.4. Uvažujeme délku zobáku a hloubku zobáku (oboje v mm) dvaceti tučňáků oslích náhodně vybraných ze 124 tučňáků stejného druhu, viz Tabulka 1.1. V první kapitole jsme pracovali pouze s hloubkou zobáku. Na základě 19 napozorovaných hodnot jsme zkonstruovali predikční interval pro neznámou hloubku zobáku dvacátého tučňáka na třech různých hladinách. Konstruovat predikční interval pro neznámou hloubku zobáku dvacátého tučňáka budeme i tentokrát, nyní ovšem vezmeme v úvahu i délky zobáků jednotlivých tučňáků.

Pro jednoduchost zkonstruujeme pouze predikční interval na hladině 0.95. Neznámou hloubku zobáku dvacátého tučňáka tentokrát označme jako Y_{20}. Volíme míru nekonformity založenou na bodových odhadech založených na metodě lineární regrese. Skóre nekonformity pak mají následující tvar

$$\begin{aligned} r_{20}^y &= |y - \widehat{Y}_{20}^y|, \\ r_i^y &= |Y_i - \widehat{Y}_i^y|, \quad i \in 1, \ldots, n. \end{aligned}$$

Predikční interval konstruujeme tak, že si zvolíme nějaké $y \in \mathbb{R}$, tedy potenciální hodnotu hloubky zobáku dvacátého tučňáka, a ověříme, zda patří do predikčního intervalu. Do intervalu zahrneme všechna taková $y \in \mathbb{R}$, pro která platí

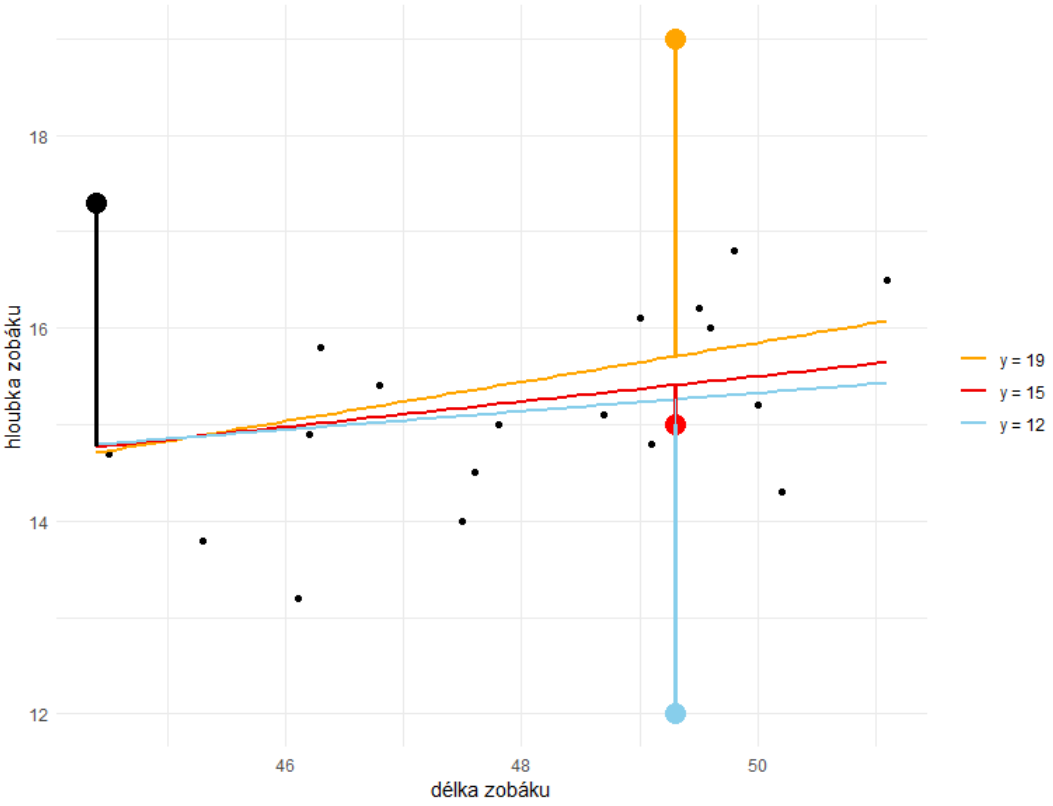
$$p(y) > 0.05 \quad \text{kde} \quad p(y) = \frac{\sum_{i=1}^{20} \mathbb{I}\{r_i^y \geq r_{20}^y\}}{20},$$

tedy všechna taková $y \in \mathbb{R}$, pro která platí $\sum_{i=1}^{20} \mathbb{I}\{r_i^y \geq r_{20}^y\} > 1$. Tedy ta $y \in \mathbb{R}$, pro která je skóre nekonformity r_{20}^y mezi $r_1^y, \ldots, r_{20}^y$ nejvýše devatenácté největší. Z obrázku 3.1 vidíme, že pro každou volbu $y \in \mathbb{R}$ konstruujeme novou regresní přímku a počítáme skóre nekonformity r_i^y jako vzdálenost hodnoty hloubky zobáku i-tého tučňáka od této přímky.

Podívejme se na hodnoty hloubek zobáků v Tabulce 1.1 a zvolme $y \in \mathbb{R}$ na potenciální dolní hranici predikčního intervalu. Zvolme nejprve $y = 12$. V prvním sloupci Tabulky 3.1 vidíme všechna skóre nekonformity pro volbu $y = 12$ seřazená od nejmenšího po největší. Hodnota r_{20}^{12} je největší, tedy $y = 12$ nezahrneme do predikčního intervalu, jedná se o příliš malou hodnotu. Podobná situace nastává, pro potenciální odhad horní hranice $y = 19$. Ve třetím sloupci Tabulky 3.1 se můžeme přesvědčit, že hodnota r_{20}^{19} je největší mezi ostatními skóre nekonformity a $y = 19$ do predikčního intervalu nezahrneme. Naopak při volbě $y = 15$ je r_{20}^{15} sedmé nejmenší skóre nekonformity, hodnota $y = 15$ tedy patří do predikčního intervalu. Postupným dosazováním hodnot za y v programu R dostaneme oboustranný predikční interval pro hloubku zobáku dalšího náhodně vybraného tučňáka na hladině 0.95 ve tvaru $(12.737, 18.231)$.

y = 12	y = 15	y = 19
$r_{17}^{12} = 0.038$	$r_{12}^{15} = 0.078$	$r_{12}^{19} = 0.028$
$r_{2}^{12} = 0.038$	$r_{1}^{15} = 0.101$	$r_{1}^{19} = 0.175$
$r_{1}^{12} = 0.046$	$r_{19}^{15} = 0.212$	$r_{16}^{19} = 0.203$
$r_{19}^{12} = 0.069$	$r_{17}^{15} = 0.230$	$r_{18}^{19} = 0.231$
$r_{12}^{12} = 0.115$	$r_{2}^{15} = 0.301$	$r_{19}^{19} = 0.401$
$r_{15}^{12} = 0.369$	$r_{16}^{15} = 0.320$	$r_{4}^{19} = 0.426$
$r_{16}^{12} = 0.408$	$r_{20}^{15} = 0.409$	$r_{3}^{19} = 0.452$
$r_{11}^{12} = 0.554$	$r_{18}^{15} = 0.552$	$r_{6}^{19} = 0.454$
$r_{18}^{12} = 0.792$	$r_{15}^{15} = 0.582$	$r_{17}^{19} = 0.485$
$r_{5}^{12} = 0.846$	$r_{11}^{15} = 0.685$	$r_{2}^{19} = 0.650$
$r_{6}^{12} = 0.938$	$r_{6}^{15} = 0.731$	$r_{5}^{19} = 0.704$
$r_{8}^{12} = 0.954$	$r_{3}^{15} = 0.765$	$r_{11}^{19} = 0.861$
$r_{3}^{12} = 1.000$	$r_{5}^{15} = 0.785$	$r_{15}^{19} = 0.867$
$r_{13}^{12} = 1.046$	$r_{4}^{15} = 0.855$	$r_{10}^{19} = 0.991$
$r_{9}^{12} = 1.077$	$r_{9}^{15} = 1.083$	$r_{9}^{19} = 1.092$
$r_{4}^{12} = 1.177$	$r_{13}^{15} = 1.172$	$r_{13}^{19} = 1.340$
$r_{10}^{12} = 1.577$	$r_{8}^{15} = 1.227$	$r_{8}^{19} = 1.591$
$r_{7}^{12} = 1.738$	$r_{10}^{15} = 1.326$	$r_{7}^{19} = 1.855$
$r_{14}^{12} = 2.492$	$r_{7}^{15} = 1.788$	$r_{14}^{19} = 2.592$
$r_{20}^{12} = 3.185$	$r_{14}^{15} = 2.535$	$r_{20}^{19} = 3.293$

Tabulka 3.1: Tabulka skóre nekonformity pro volby $y = 12$, $y = 15$ a $y = 19$.



Obrázek 3.1: Regresní přímky zkonstruované na základě volby $y = 12$, $y = 15$ a $y = 19$.

4. Simulační studie

Hlavní výhodou metody konformní predikce je, že za předpokladu, že uvažujeme nezávislé náhodné veličiny nebo vektory, produkuje predikční interval na stanovené hladině pro libovolný rozsah náhodného výběru $n \in \mathbb{N}$. V sekci 1.2 jsme uvedli frekventistickou metodu konstrukce predikčního intervalu, která platí za předpokladu, že uvažujeme náhodný výběr z normálního rozdělení. V této kapitole porovnáme výsledky získané použitím metody konformní predikce s výsledky obdrženými zmíněnou frekventistickou metodou ze sekce 1.2, která produkuje predikční interval ve tvaru

$$B_n(\boldsymbol{X}) = \left(\overline{X}_n \pm \sqrt{S_n^2 \left(1 + \frac{1}{n}\right)} \cdot t_{n-1}\left(1 - \frac{\alpha}{2}\right) \right).$$

4.1 Normální rozdělení

Budeme postupně pracovat se sadou náhodných výběrů o rozsahu $19, 39, 199$, z rozdělení $N(0, 1)$. Generování a výpočet opakujeme pro každý rozsah výběru 1000krát. Výpočet probíhá tak, že zkonstruujeme predikční interval na základě 19, respektive 39, respektive 199 pozorování, následně vygenerujeme další nezávislé pozorování a ověříme, zda leží ve vytvořeném predikčním intervalu. Při výpočtu předstíráme, že střední hodnotu ani rozptyl rozdělení neznáme a interval $B_n(\boldsymbol{X})$ konstruujeme pro frekventistickou metodu (dále také metodu F) přesně tak, jak bylo popsáno v sekci 1.2 a pro metodu konformní predikce (dále také metodu K) přesně tak, jak jsem popsali v sekci 1.4. Jelikož má metoda F v porovnání s metodou konformní predikce přísnější předpoklady, předpokládáme, že bude dávat lepší výsledky než univerzálnější metoda konformní predikce.

V Tabulce 4.1 vidíme, že pro všechny rozsahy náhodného výběru mají obě metody o něco menší pokrytí, než garantovaných 0.95, to může být ale pouze náhoda. Zdá se, že z hlediska pokrytí se metody neliší. Pro délky predikčních intervalů už je to jinak. Pro náhodný výběr o rozsahu 19 produkuje metoda K nekonečně dlouhý predikční interval. To souvisí s principem představeným v příkladu 1.4. Vybíráme totiž 19 intervalů z 20 a lze tak konstruovat pouze jednostranný predikční interval. Takový nedostatek metoda F nemá. Pro větší rozsahy výběrů jsou již délky predikčních intervalů podobnější. Metoda K má predikční intervaly o něco větší a také variabilnější. Pro všechny rozsahy náhodných výběrů tedy metoda F dává lepší výsledky, pro rozsah 19 pak značně lepší, jelikož výsledné predikční intervaly jsou výrazně kratší, tedy informativnější.

	n	metoda	pokrytí	prům. délka (sd)	min.	max.
N(0,1)	19	F	0.949	4.258 (0.710)	2.253	6.497
		K	0.948	∞ (—)	∞	∞
	39	F	0.936	4.053 (0.461)	2.707	5.837
		K	0.943	4.278 (0.657)	2.681	6.907
	199	F	0.949	3.952 (0.195)	3.182	4.594
		K	0.947	3.991 (0.272)	3.020	5.068
Exp(1)	19	F	0.914	4.100 (1.264)	1.269	10.530
		K	0.949	∞ (—)	∞	∞
	39	F	0.937	4.051 (0.871)	2.041	8.310
		K	0.947	4.279 (1.310)	1.715	11.778
	199	F	0.945	3.923 (0.376)	2.948	5.673
		K	0.950	3.754 (0.452)	2.653	5.513

Tabulka 4.1: V tabulce vidíme srovnání výsledků metod K a F při různých rozsazích náhodného výběru. Porovnáváme pokrytí, průměrnou délku predikčního intervalu, minimální délku predikčního intervalu a maximální délku predikčního intervalu.

4.2 Exponenciální rozdělení

Opět pracujeme se sadou náhodných výběrů o rozsahu $19, 39, 199$, tentokrát z rozdělení $Exp(1)$. Opět předstíráme, že střední hodnotu ani rozptyl rozdělení neznáme. Generování a výpočet opakujeme pro každý rozsah výběru 1000krát a stejně, jako bylo popsáno v předchozí sekci. Jelikož předpoklady metody F nejsou splněny, očekáváme tentokrát lepší výsledky pro metodu konformní predikce. Z Tabulky 4.1 vidíme, že zatímco konformní predikce splňuje garantované pokrytí, metoda F má pokrytí nižší než očekávaných 0.95. Pro náhodný výběru o rozsahu 19 je pokrytí výrazně nižší, pro narůstající rozsah se pak zlepšuje. Co se týče délek výsledných predikčních intervalů, metody si vedou podobně jako v případě normálního rozdělení.

Závěr

V první kapitole práce jsme definovali základní pojmy, na kterých staví metoda konformní predikce. Odvodili jsme nejjednodušší verzi metody pro náhodný výběr z jednorozměrného spojitého rozdělení a formulovali větu, která dokazuje její platnost. Fungování metody jsme následně demonstrovali na vlastním příkladu. Ve druhé kapitole jsme sjednotili značení převzaté z různých zdrojů a definovali klíčové pojmy pro formulaci algoritmického provedení metody. Následně jsme ve třetí kapitole definice a značení rozšířili do rámce regresní analýzy a opět předvedli názornou ukázku na příkladu. Ve čtvrté kapitole jsme provedli jednoduchou simulaci. V kontextu náhodného výběru z jednorozměrného spojitého rozdělení jsme porovnávali výsledky, získané konformní predikcí, s výsledky založenými na frekventistické metodě představené v sekci 1.2.

Zdůrazněme, že na vybrané téma, pokud je autorce známo, neexistuje žádná odborná publikace. Autorka tak za hlavní přínos této práce považuje výstavbu teoretických základů metody v první kapitole, formulaci Věty 3 a sjednocení značení a formulaci definic pro algoritmické provedení metody.

Tato bakalářská práce je pouze úvodem do problematiky konformní predikce. Metodu jsme uvažovali pouze v kontextu konstrukce predikčního intervalu, nabízí se tedy rozšířit její zkoumání na případ predikční množiny, t.j. zkoumat konformní predikci i v klasifikačních úlohách. Dále, mimo jiné, existuje celá řada modifikací metody formulovaných primárně za účelem zmírnění konzervativní povahy predikčního intervalu.

Seznam použité literatury

Anděl, J. (2007). *Statistické metody*. Karolinum. ISBN 978-80-7378-003-6.

Balasubramanian, V., Ho, S.-S. a Vovk, V. (2014). *Conformal prediction for reliable machine learning: theory, adaptations and applications*. Newnes, 1 edition. ISBN 9780123985378.

Fontana, M., Zeni, G. a Vantini, S. (2020). Conformal prediction: a unified review of theory and new challenges. *arXiv preprint arXiv:2005.07972*.

Gammerman, A., Vovk, V. a Vapnik, V. (1998). Learning by transduction. In *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*, UAI'98, page 148–155, San Francisco, CA, USA, 1998. Morgan Kaufmann Publishers Inc. ISBN 155860555X.

Kulich, M. (2017). Poznámky k přednášce matematická statistika 1. URL `https://www.karlin.mff.cuni.cz/~kulich/vyuka/ms1/doc/ms1_170112.pdf`. Přístup z [17.07.2023].

Semela, O. (2016). Robustní optimalizace v klasifikačních a regresních úlohách. Diplomová práce.

Shafer, G. a Vovk, V. (2008). A tutorial on conformal prediction. *Journal of Machine Learning Research*, **9**(3), 371–421.

Vovk, V., Gammerman, A. a Shafer, G. (2005). *Algorithmic learning in a random world*, volume 29. Springer, New York.