

Sekvenční rozpoznávání. Waldova analýza.

Petr Pošík

Czech Technical University in Prague
Faculty of Electrical Engineering
Dept. of Cybernetics

Motivace: Statistická
přejímka

Statistická přejímka
CUSUM diagramy

Sekvenční analýza

Závěr

Motivace: Statistická přejímka

Motivace: Statistická
přejímka

Statistická přejímka
CUSUM diagramy

Sekvenční analýza

Závěr

Situace:

- ✓ Náš podnik vyrábí stroje. Od subdodavatele nakupujeme hřídele o průměru 7.5 mm vždy v dávkách po 10 tis. kusech. Skutečné průměry dodaných součástí se od této jmenovité hodnoty samozřejmě trochu liší. Jak zjistit, zda má dávka akceptovatelnou kvalitu nebo zda ji odmítneme přijmout a vrátíme ji výrobcí?

Motivace: Statistická
přejímka

Statistická přejímka
CUSUM diagramy

Sekvenční analýza

Závěr

Situace:

- ✓ Náš podnik vyrábí stroje. Od subdodavatele nakupujeme hřídele o průměru 7.5 mm vždy v dávkách po 10 tis. kusech. Skutečné průměry dodaných součástí se od této jmenovité hodnoty samozřejmě trochu liší. Jak zjistit, zda má dávka akceptovatelnou kvalitu nebo zda ji odmítneme přijmout a vrátíme ji výrobcí?

Možnosti:

- ✓ 100% kontrola: neekonomické, při destruktivních zkouškách nemožné

Situace:

- ✓ Náš podnik vyrábí stroje. Od subdodavatele nakupujeme hřídele o průměru 7.5 mm vždy v dávkách po 10 tis. kusech. Skutečné průměry dodaných součástek se od této jmenovité hodnoty samozřejmě trochu liší. Jak zjistit, zda má dávka akceptovatelnou kvalitu nebo zda ji odmítneme přijmout a vrátíme ji výrobcí?

Možnosti:

- ✓ 100% kontrola: neekonomické, při destruktivních zkouškách nemožné
- ✓ Statistická přejímka:
 - ✗ otestujeme (změříme) jen malý počet kusů
 - ✗ na základě těchto měření usoudíme na kvalitu dávky jako celku
 - ✗ šetříme čas, práci i peníze
 - ✗ základní otázka: jak určit, kolik vzorků z dávky musíme změřit, abychom mohli dávku s dostatečnou jistotou přijmout/odmítnout?
 - ✗ možnost chyby:
 - ✓ zamítneme dobrou dávku (chyba 1. druhu, α)
 - ✓ přijmeme špatnou dávku (chyba 2. druhu, β)

Klasické testování hypotéz dává do souvislosti

- ✓ velikost efektu, tj. rozdíl mezi
 - ✗ nulovou hyp. H_0 (shoda se specifikacemi, např. $D = 7.5$ mm) a
 - ✗ alternativní hyp. H_1 (nepřijatelná odchylka, např. $D = 7.505$ mm),
- ✓ akceptovatelné pravděpodobnosti chyb 1. a 2. druhu (α a β) a
- ✓ velikost vzorku N .

Klasické testování hypotéz dává do souvislosti

- ✓ velikost efektu, tj. rozdíl mezi
 - ✗ nulovou hyp. H_0 (shoda se specifikacemi, např. $D = 7.5$ mm) a
 - ✗ alternativní hyp. H_1 (nepřijatelná odchylka, např. $D = 7.505$ mm),
- ✓ akceptovatelné pravděpodobnosti chyb 1. a 2. druhu (α a β) a
- ✓ velikost vzorku N .

Platí:

- ✓ Čím méně chyb požadujeme, tím větší vzorek budeme potřebovat (abychom si mohli být 100% jistí, museli bychom zkontrolovat celou dávku).
- ✓ Čím menší je “rozdíl” mezi hypotézami, tím větší vzorek budeme potřebovat.

Klasické testování hypotéz dává do souvislosti

- ✓ velikost efektu, tj. rozdíl mezi
 - ✗ nulovou hyp. H_0 (shoda se specifikacemi, např. $D = 7.5$ mm) a
 - ✗ alternativní hyp. H_1 (nepřijatelná odchylka, např. $D = 7.505$ mm),
- ✓ akceptovatelné pravděpodobnosti chyb 1. a 2. druhu (α a β) a
- ✓ velikost vzorku N .

Platí:

- ✓ Čím méně chyb požadujeme, tím větší vzorek budeme potřebovat (abychom si mohli být 100% jistí, museli bychom zkontrolovat celou dávku).
- ✓ Čím menší je “rozdíl” mezi hypotézami, tím větší vzorek budeme potřebovat.

Pevná přejímka:

- ✓ Založena na klasickém testování hypotéz
- ✓ Ze zadané velikosti efektu a pstí α a β určíme potřebný počet testovaných kusů N
- ✓ Na základě N naměřených údajů se rozhodneme, zda dávku přijmeme či nikoli

Statistická přejímka (pokr.)

Motivace: Statistická přejímka

Statistická přejímka

CUSUM diagramy

Sekvenční analýza

Závěr

Klasické testování hypotéz dává do souvislosti

- ✓ velikost efektu, tj. rozdíl mezi
 - ✗ nulovou hyp. H_0 (shoda se specifikacemi, např. $D = 7.5$ mm) a
 - ✗ alternativní hyp. H_1 (nepřijatelná odchylka, např. $D = 7.505$ mm),
- ✓ akceptovatelné pravděpodobnosti chyb 1. a 2. druhu (α a β) a
- ✓ velikost vzorku N .

Platí:

- ✓ Čím méně chyb požadujeme, tím větší vzorek budeme potřebovat (abychom si mohli být 100% jistí, museli bychom zkontrolovat celou dávku).
- ✓ Čím menší je “rozdíl” mezi hypotézami, tím větší vzorek budeme potřebovat.

Pevná přejímka:

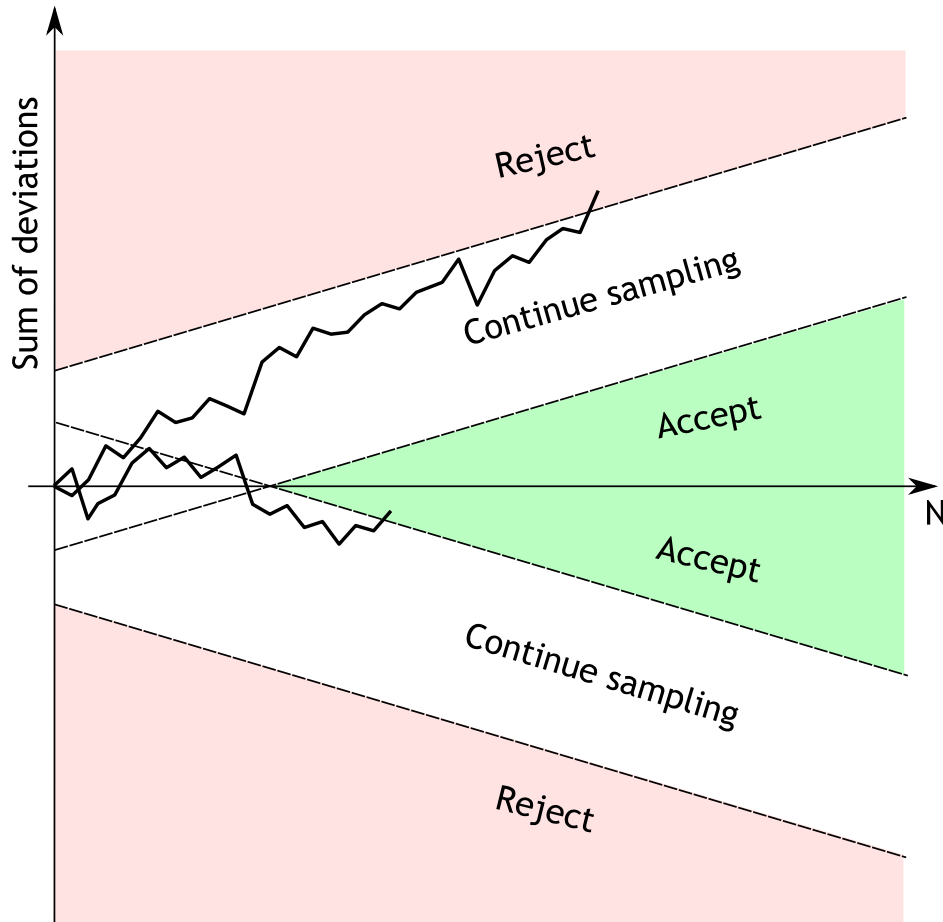
- ✓ Založena na klasickém testování hypotéz
- ✓ Ze zadané velikosti efektu a pstí α a β určíme potřebný počet testovaných kusů N
- ✓ Na základě N naměřených údajů se rozhodneme, zda dávku přijmeme či nikoli

Postupná (sekvenční) přejímka:

- ✓ Změříme 1 kus
- ✓ Přijmeme/zamítneme celou dávku, pokud nám dosud změřené údaje poskytují dostatek “důkazů”, jinak pokračujeme v měření dalšího kusu.

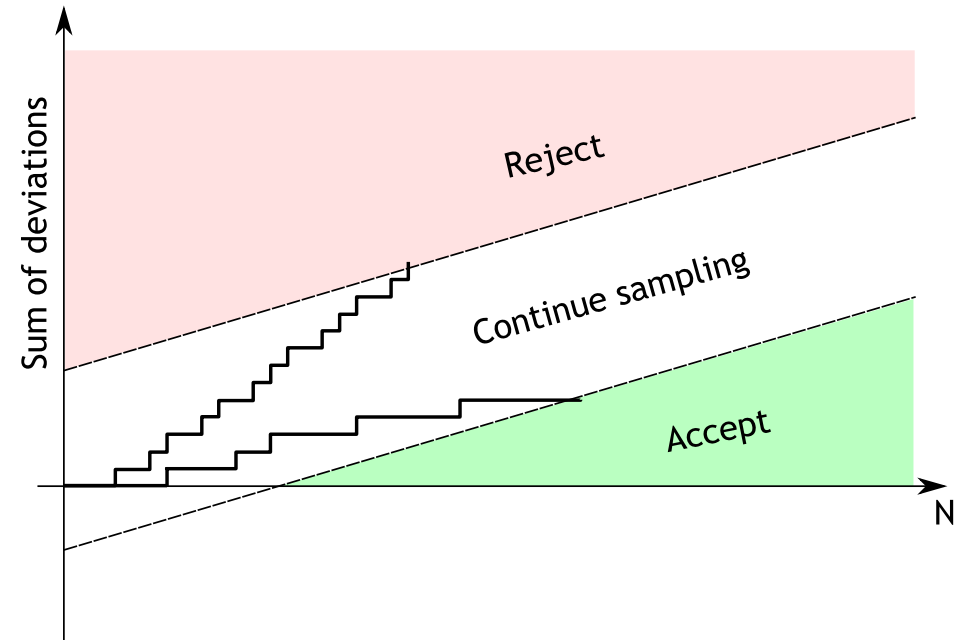
CUSUM diagramy

Oboustranný sekvenční test hypotézy o průměru:



- ✓ Součet odchylek může klesat i růst – každá odchylka může být “do plusu” i do “mínusu” od specifikace

Jednostranný sekvenční test hypotézy o četnosti:



- ✓ Součet odchylek může jenom růst (počet neshodných jednotek)

Motivace: Statistická
přejímka

Sekvenční analýza

Sekvenční analýza

Sekvenční rozhodovací
problém

SPRT (Sequential
Probability Ratio Test)

SPRT: Hranice A a B

SPRT: Závěrečná
doporučení

Závěr

Sekvenční analýza

- ✓ podoblast statistiky a strojového učení
- ✓ způsob provádění analýzy nějakým způsobem závisí na výsledcích předchozích kroků:
 1. výběr experimentu, měření nebo testu, který se provede v dalším kroku
 2. zastavení/pokračování analýzy

Výhody:

- ✓ Testování může být kratší než u klasické analýzy celého vzorku.
- ✓ Testy nemusí být stejného typu!
- ✓ Umožňuje měnit diagnostický postup v závislosti na výsledcích dřívějších kroků.
Porovnejte:
 - ✗ Chceme určit, zda pacient trpí rakovinou. Provedeme všechny následující testy: krevní obraz, rentgen, CT, magnetickou rezonanci, ultrazvuk, ... Na základě všech těchto měření rozhodneme, zda pacient rakovinou trpí nebo ne.
 - ✗ Chceme určit, zda pacient trpí rakovinou. Provedeme analýzu krevního obrazu. Nevykazuje-li žádné abnormality, prohlásíme jej za zdravého. Jinak, podle druhu abnormality jej pošleme buď na rentgen, CT, magnetickou rezonanci, nebo ultrazvuk, ...¹

¹Tento postup není medicínsky správně; je zde uveden jen jako možná ukázka.

Sekvenční rozhodovací problém

Motivace: Statistická
přejímka

Sekvenční analýza

Sekvenční analýza

Sekvenční rozhodovací
problém

SPRT (Sequential
Probability Ratio Test)

SPRT: Hranice A a B

SPRT: Závěrečná
doporučení

Závěr

- ✓ Objekt x patří do jedné ze dvou tříd $\{-1, +1\}$
- ✓ Je dáno uspořádání měření $\{x_1, \dots, x_m\}$ na objektu x
- ✓ Sekvenční rozhodovací strategie je množina rozhodovacích funkcí $S = \{S_1, \dots, S_m\}$
- ✓ Každá rozhodovací funkce $S_i : \{x_1, \dots, x_i\} \rightarrow \{-1, +1, \#\}$

Sekvenční rozhodovací problém

Motivace: Statistická
přejímka

Sekvenční analýza

Sekvenční analýza

Sekvenční rozhodovací
problém

SPRT (Sequential
Probability Ratio Test)

SPRT: Hranice A a B
SPRT: Závěrečná
doporučení

Závěr

- ✓ Objekt x patří do jedné ze dvou tříd $\{-1, +1\}$
- ✓ Je dáno uspořádání měření $\{x_1, \dots, x_m\}$ na objektu x
- ✓ Sekvenční rozhodovací strategie je množina rozhodovacích funkcí $S = \{S_1, \dots, S_m\}$
- ✓ Každá rozhodovací funkce $S_i : \{x_1, \dots, x_i\} \rightarrow \{-1, +1, \#\}$

Sekvenční rozhodovací strategie S

- ✓ v kroku i použije k rozhodnutí funkci S_i , která na základě měření x_1, \dots, x_i zařadí x do jedné ze dvou tříd $\{-1, +1\}$ nebo vydá rozhodnutí $\#$ (“nedovedu rozhodnout”)
- ✓ pokud S_i vydá rozhodnutí $\#$, strategie provede měření x_{i+1} a použije funkci S_{i+1}
- ✓ je charakterizována chybami 1. a 2. druhu (α_S a β_S) a průměrnou dobou do rozhodnutí

$$\bar{T}_S = E(T_S(\mathbf{x})),$$

kde čas do rozhodnutí T_S je dán jako

$$T_S(\mathbf{x}) = \arg \min_i (S_i(\mathbf{x}) \neq \#).$$

Sekvenční rozhodovací problém

Motivace: Statistická
přejímka

Sekvenční analýza

Sekvenční analýza

Sekvenční rozhodovací
problém

SPRT (Sequential
Probability Ratio Test)

SPRT: Hranice A a B
SPRT: Závěrečná
doporučení

Závěr

- ✓ Objekt x patří do jedné ze dvou tříd $\{-1, +1\}$
- ✓ Je dáno uspořádání měření $\{x_1, \dots, x_m\}$ na objektu x
- ✓ Sekvenční rozhodovací strategie je množina rozhodovacích funkcí $S = \{S_1, \dots, S_m\}$
- ✓ Každá rozhodovací funkce $S_i : \{x_1, \dots, x_i\} \rightarrow \{-1, +1, \#\}$

Sekvenční rozhodovací strategie S

- ✓ v kroku i použije k rozhodnutí funkci S_i , která na základě měření x_1, \dots, x_i zařadí x do jedné ze dvou tříd $\{-1, +1\}$ nebo vydá rozhodnutí $\#$ (“nedovedu rozhodnout”)
- ✓ pokud S_i vydá rozhodnutí $\#$, strategie provede měření x_{i+1} a použije funkci S_{i+1}
- ✓ je charakterizována chybami 1. a 2. druhu (α_S a β_S) a průměrnou dobou do rozhodnutí

$$\bar{T}_S = E(T_S(\mathbf{x})),$$

kde čas do rozhodnutí T_S je dán jako

$$T_S(\mathbf{x}) = \arg \min_i (S_i(\mathbf{x}) \neq \#).$$

Optimální sekvenční rozhodovací strategie:

$$S^* = \arg \min_S \bar{T}_S \quad (1) \quad \text{tak, aby} \quad \alpha_S \leq \alpha \quad \text{a} \quad \beta_S \leq \beta.$$

SPRT (Sequential Probability Ratio Test)

Motivace: Statistická
přejímka

Sekvenční analýza

Sekvenční analýza

Sekvenční rozhodovací
problém

SPRT (Sequential
Probability Ratio Test)

SPRT: Hranice A a B

SPRT: Závěrečná
doporučení

Závěr

- ✓ Základní a nejdůležitější metoda sekvenční analýzy [Wal47]
- ✓ Objekt \mathbf{x} je charakterizován vnitřním stavem (třídou) $y \in \{-1, +1\}$; tento stav není znám a je třeba ho odhadnout na základě postupných měření x_1, \dots, x_m .
- ✓ Známe všechna sdružená podm. rozd. pravděpodobnosti $p(x_1, \dots, x_m | y = c)$

- ✓ Určeme si hypotézy a k nim příslušné chyby:

$$H_0 : y = +1 \quad \alpha = P(S(\mathbf{x}) = -1 | y = +1)$$

$$H_A : y = -1 \quad \beta = P(S(\mathbf{x}) = +1 | y = -1)$$

- ✓ Definujme poměr věrohodností R_m :

$$R_m(\mathbf{x}) = \frac{p(x_1, \dots, x_m | y = -1)}{p(x_1, \dots, x_m | y = +1)} \quad (2)$$

- ✓ SPRT je následující sekvenční strategie:

$$S_m^*(\mathbf{x}) = \begin{cases} -1, & \text{když } R_m(\mathbf{x}) \geq A, \\ +1, & \text{když } R_m(\mathbf{x}) \leq B, \\ \#, & \text{když } B < R_m(\mathbf{x}) < A. \end{cases} \quad (3)$$

- ✓ Hranice A a B jsou parametry testu a nastavují se podle požadovaných chyb α a β
- ✓ Lze ukázat, že
SPRT s optimálně nastavenými hranicemi A^* a B^* je optimálním sekvenčním testem ve smyslu kritéria (1).

SPRT: Hranice A a B

Motivace: Statistická
přejímka

Sekvenční analýza

Sekvenční analýza

Sekvenční rozhodovací
problém

SPRT (Sequential
Probability Ratio Test)

SPRT: Hranice A a B

SPRT: Závěrečná
doporučení

Závěr

Horní mez pro A^* :

Předpokládejme, že SPRT vydá rozhodnutí $S_m^*(\mathbf{x}) = -1$. Muselo tedy platit

$$R_m(\mathbf{x}) = \frac{p(x_1, \dots, x_m | y = -1)}{p(x_1, \dots, x_m | y = +1)} \geq A^*, \text{ neboli}$$

$$p(x_1, \dots, x_m | y = -1) \geq A^* \cdot p(x_1, \dots, x_m | y = +1).$$

Protože toto platí pro všechny sekvence \mathbf{x} takové, že $S_m^*(\mathbf{x}) = -1$, platí pro součet přes všechna taková \mathbf{x} :

$$\int_{\mathbf{x}: S_m^*(\mathbf{x}) = -1} p(x_1, \dots, x_m | y = -1) d\mathbf{x} \geq A^* \int_{\mathbf{x}: S_m^*(\mathbf{x}) = -1} p(x_1, \dots, x_m | y = +1) d\mathbf{x}$$
$$\underbrace{P(S_m^*(\mathbf{x}) = -1 | y = -1)}_{1-\beta} \geq A^* \cdot \underbrace{P(S_m^*(\mathbf{x}) = -1 | y = +1)}_{\alpha}$$

Horní mez A' pro A^* je tedy

$$\frac{1-\beta}{\alpha} = A' \geq A^* \tag{4}$$

SPRT: Hranice A a B

Motivace: Statistická
přejímka

Sekvenční analýza

Sekvenční analýza

Sekvenční rozhodovací
problém

SPRT (Sequential
Probability Ratio Test)

SPRT: Hranice A a B

SPRT: Závěrečná
doporučení

Závěr

Horní mez pro A^* :

$$\frac{1 - \beta}{\alpha} = A' \geq A^* \quad (4)$$

Dolní mez pro B^* :

Předpokládejme, že SPRT vydá rozhodnutí $S_m^*(\mathbf{x}) = +1$. Muselo tedy platit

$$R_m(\mathbf{x}) = \frac{p(x_1, \dots, x_m | y = -1)}{p(x_1, \dots, x_m | y = +1)} \leq B^*, \text{ neboli}$$

$$p(x_1, \dots, x_m | y = -1) \leq B^* \cdot p(x_1, \dots, x_m | y = +1).$$

Protože toto platí pro všechny sekvence \mathbf{x} takové, že $S_m^*(\mathbf{x}) = +1$, platí pro součet přes všechna taková \mathbf{x} :

$$\int_{\mathbf{x}: S_m^*(\mathbf{x})=+1} p(x_1, \dots, x_m | y = -1) d\mathbf{x} \leq B^* \int_{\mathbf{x}: S_m^*(\mathbf{x})=+1} p(x_1, \dots, x_m | y = +1) d\mathbf{x}$$
$$\underbrace{P(S_m^*(\mathbf{x}) = +1 | y = -1)}_{\beta} \leq B^* \cdot \underbrace{P(S_m^*(\mathbf{x}) = +1 | y = +1)}_{1-\alpha}$$

Dolní mez B' pro B^* je tedy

$$\frac{\beta}{1 - \alpha} = B' \leq B^* \quad (5)$$

SPRT: Závěrečná doporučení

Motivace: Statistická
přejímka

Sekvenční analýza

Sekvenční analýza

Sekvenční rozhodovací
problém

SPRT (Sequential
Probability Ratio Test)

SPRT: Hranice A a B

SPRT: Závěrečná
doporučení

Závěr

- ✓ Optimální hranice A^* a B^* není snadné určit. V praxi Wald doporučuje použít místo optimálních hodnot jejich meze:

$$A = A' = \frac{1 - \beta}{\alpha} \qquad B = B' = \frac{\beta}{1 - \alpha}$$

neboli použít přísnější hranice, než je nutné.

- ✓ Lze ukázat, že při použití těchto hranic se pravděpodobnosti chyb SPRT změní na α' a β' , pro něž platí $\alpha' + \beta' \leq \alpha + \beta$, takže zhoršit se může maximálně jedna chyba, zlepšit se musí alespoň jedna chyba.

Co Wald neřešil:

1. optimální pořadí jednotlivých měření
 - ✓ v jeho aplikacích byla všechna měření i.i.d. (nezávislá a stejně rozdělená), takže na jejich pořadí nezáleželo
2. odhad poměru věrohodností (2) z trénovacích dat
 - ✓ opět díky i.i.d. předpokladu platí

$$p(x_1, \dots, x_m | y = c) = \prod_{i=1}^m p(x_i | y = c), \qquad (6)$$

díky čemuž se R_m dá počítat inkrementálně z 1D rozdělení psti.

Motivace: Statistická
přejímka

Sekvenční analýza

Závěr

Shrnutí

Reference

Závěr

Sekvenční analýza

- ✓ často umožňuje rozhodnout o příslušnosti objektu ke třídě na základě menšího počtu příznaků než vyžadují obvyklé metody
- ✓ má mnoho aplikací, především tam, kde
 - ✗ je třeba minimalizovat čas do rozhodnutí
 - ✗ jednotlivá měření nejsou levná

Aplikace

- ✓ postupná statistická přejímka: měření jsou homogenní a nezáleží na pořadí
- ✓ klinická studie účinnosti léku: sekvenční testování po skupinách – pokud měření na skupině n lidí je dostatečně průkazné k potvrzení/vyvrácení účinnosti léku, test končí, jinak se pokračuje na další skupině n lidí ...
- ✓ detekce obličejů na fotkách a videosekvencích [Š09]
 - ✗ jednotlivá měření představují rozdíly jasů v různých místech obrázku
 - ✗ čas do rozhodnutí je kritický, klasifikátor má být použitelný v reálném čase!!!
 - ✗ měření nejsou nezávislá!!!
 - ✗ AdaBoost: pořadí jednotlivých měření (příznaků) a odhad poměru psí
 - ✗ rozhodnutí o tom, zda se jedná o pozadí nebo obličej (nebo #), dělá SPRT
 - ✗ výsledná kombinace: WaldBoost
- ✓ ...

Reference

Motivace: Statistická
přejímka

Sekvenční analýza

Závěr

Shrnutí

Reference

- [Fu68] K. S. Fu. *Sequential methods in pattern recognition and machine learning (Mathematics in science and engineering)*. Academic Press, 1st edition, 1968.
- [Š09] Jan Šochman. *Learning for Sequential Classification*. PhD thesis, Czech Technical University in Prague, Prague, Czech Republic, December 2009. Available online at http://cyber.felk.cvut.cz/phd/completed/Sochman-PhD12_2009.pdf.
- [Wal47] Abraham Wald. *Sequential Analysis*. Wiley, New York, 1947.