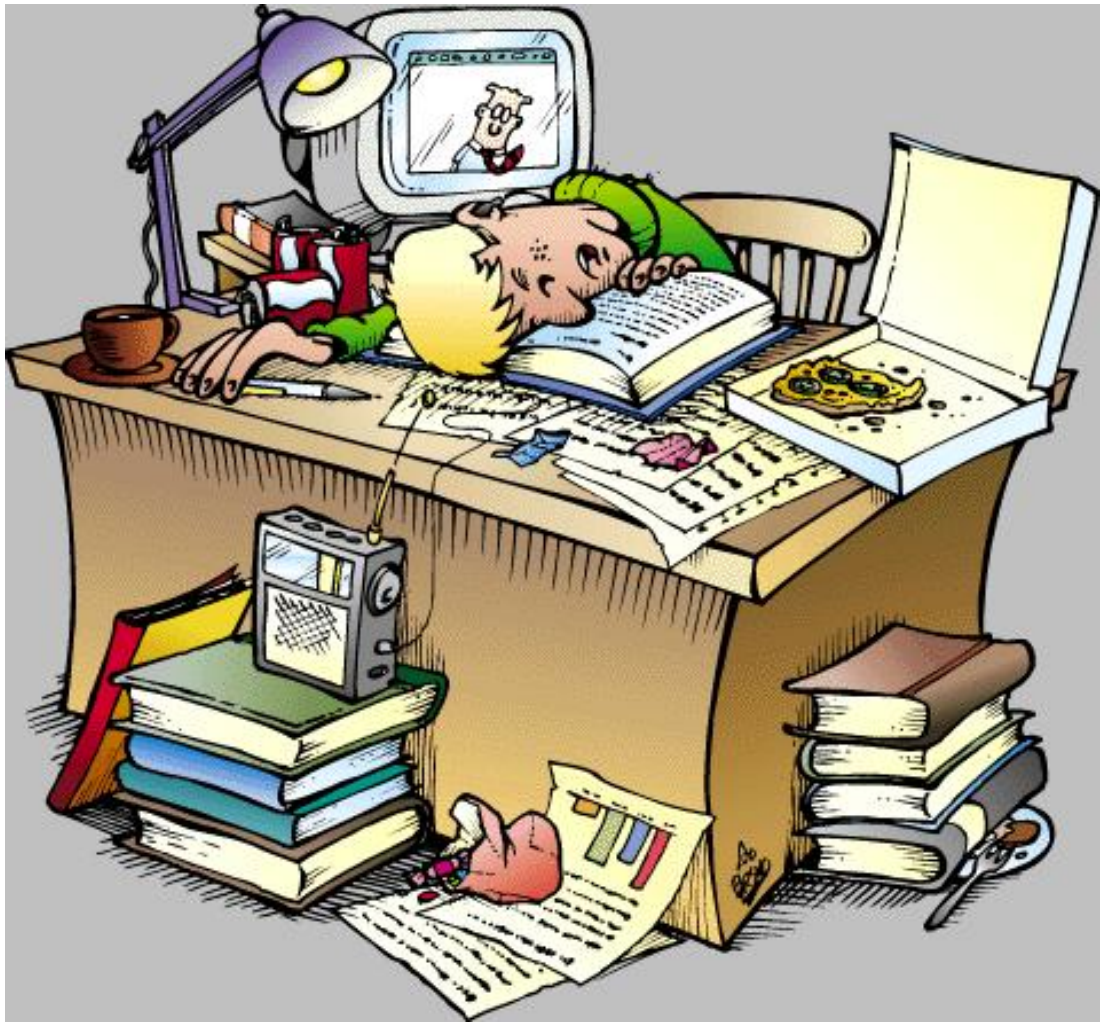
Vytěžování dat

Filip Železný

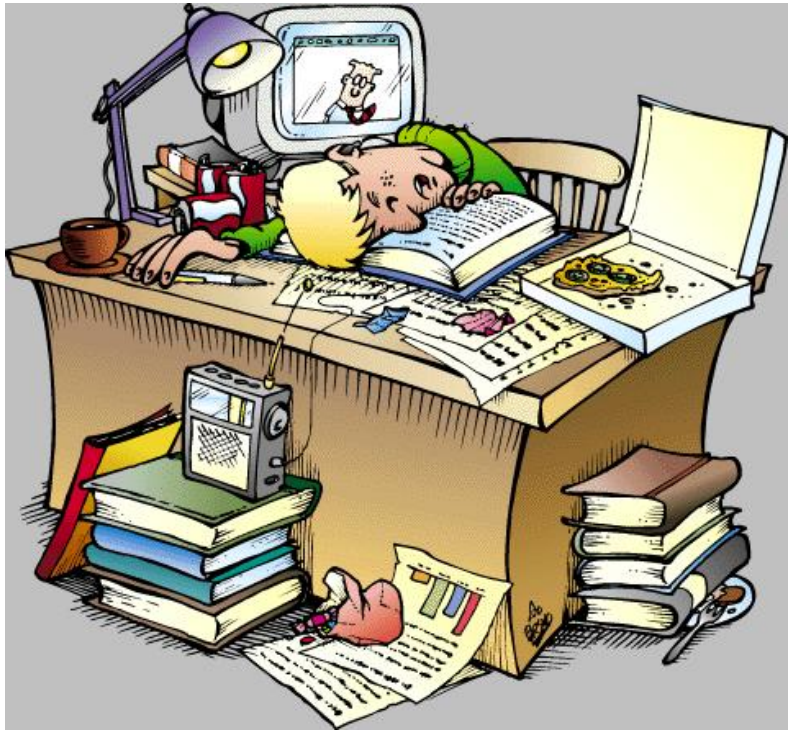
Katedra kybernetiky
skupina Inteligentní Datové Analýzy (IDA)

25. března 2009

Metafora pro tuto přednášku

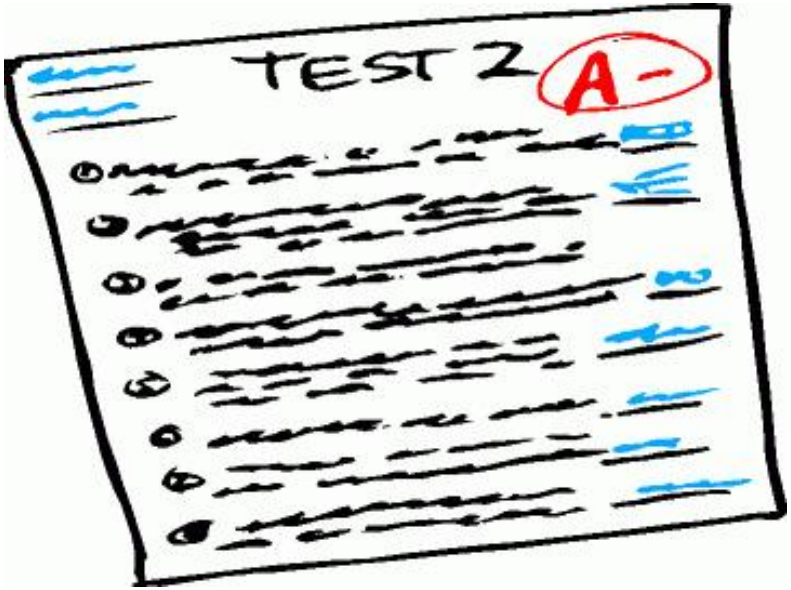


Metafora pro tuto přednášku



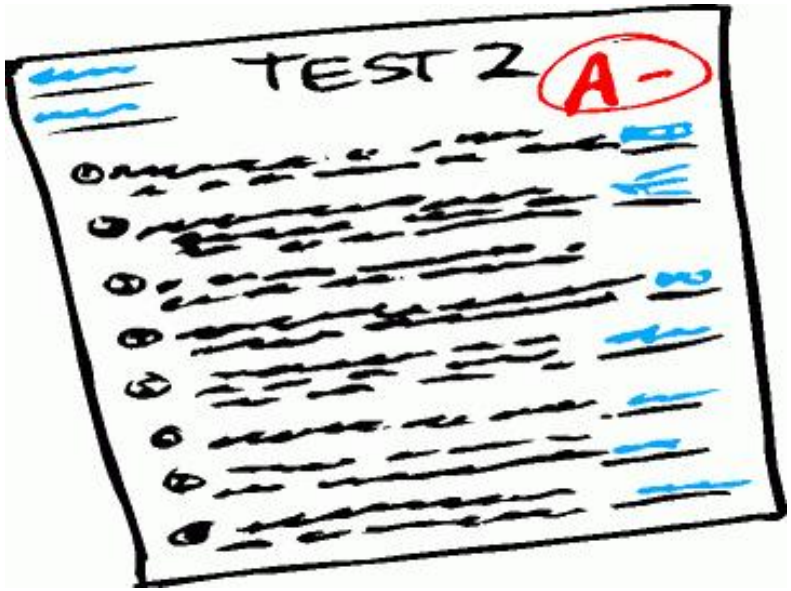
- Zajímá nás, jak student umí látku.
- Chybovost: *Err*
  - % chybných odpovědí na *všechny* možné otázky z dané látky

Statistický odhad



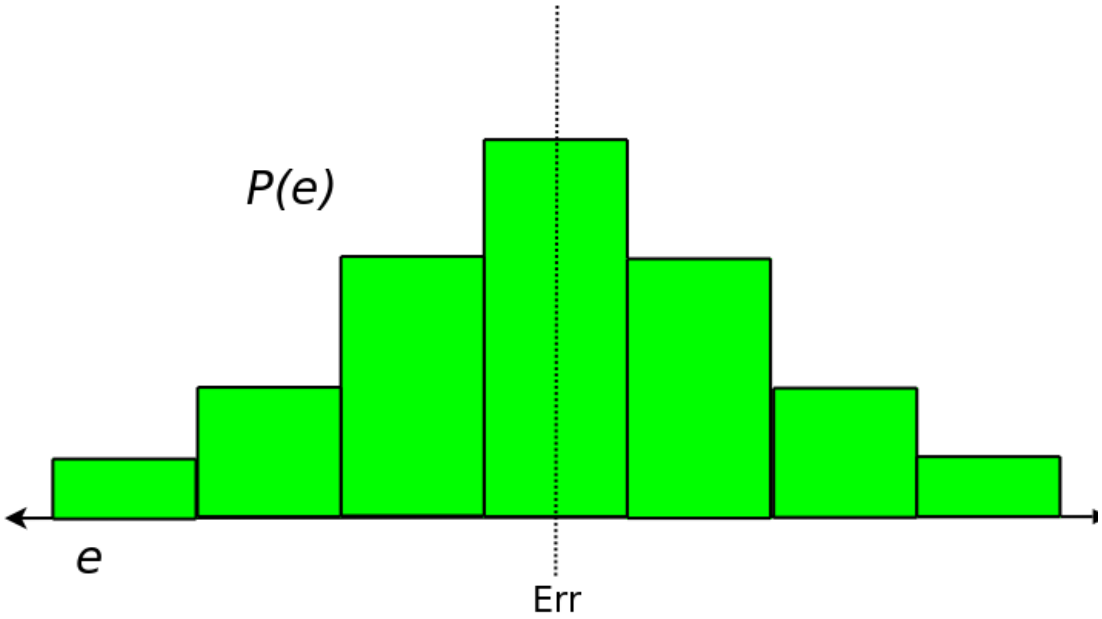
- Chybovost *Err* můžeme odhadnout *zkouškou*
  - *m* vybraných otázek ze všech otázek látky

Statistický odhad



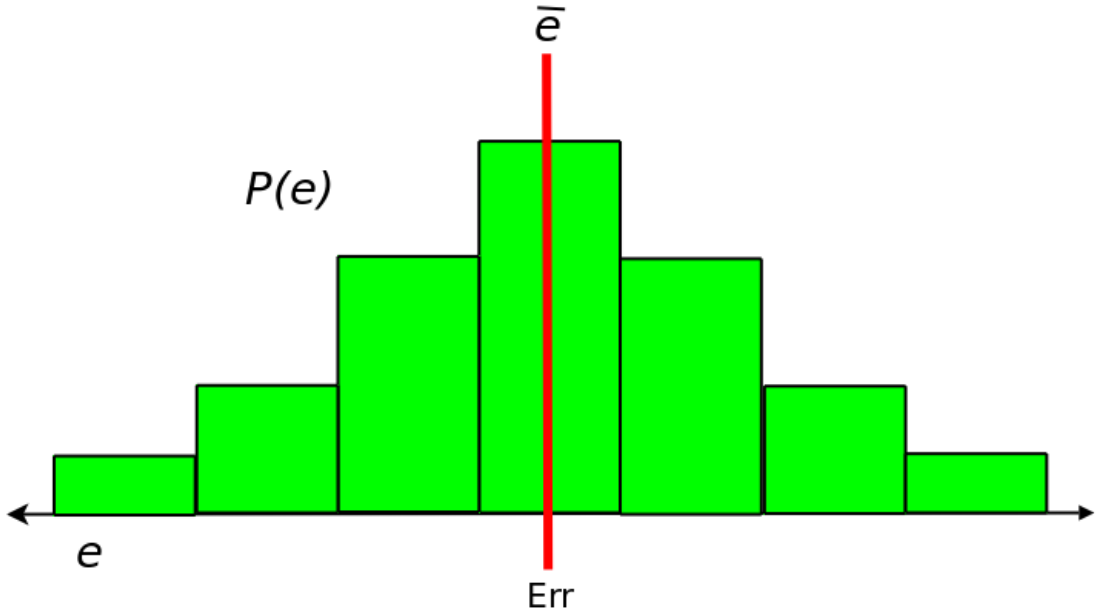
- Chybovost *Err* můžeme odhadnout *zkouškou*
  - *m* vybraných otázek ze všech otázek látky
- Předpokládejme, že otázky jsou vybrány náhodně
  - tzv. iid výběr: všechny z původního pr. rozdělení a na sobě nezávisle
- "Spravedlivá zkouška"

Nevychýlený odhad



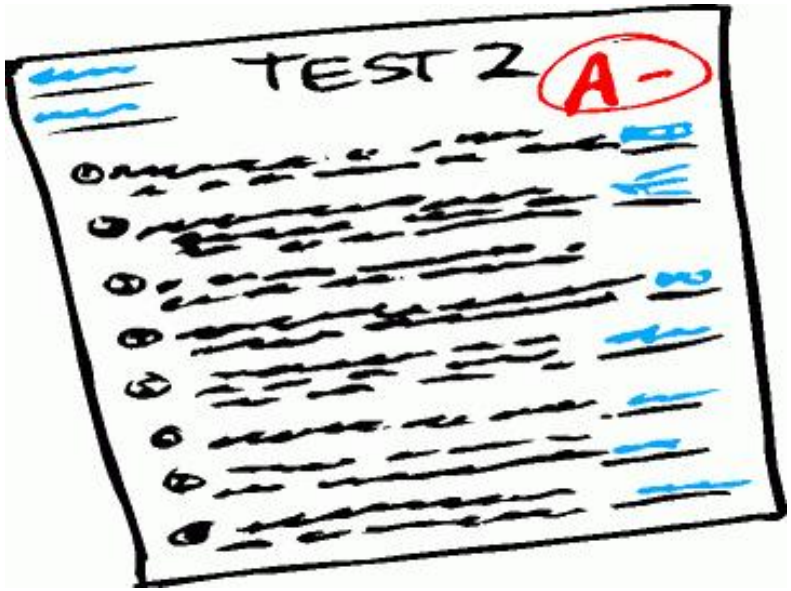
- Výsledek zkoušky e je náhodná veličina s rozdělením $P(e)$
- Střední hodnota $\bar{e} = \sum_e e \cdot P(e)$

Nevychýlený odhad



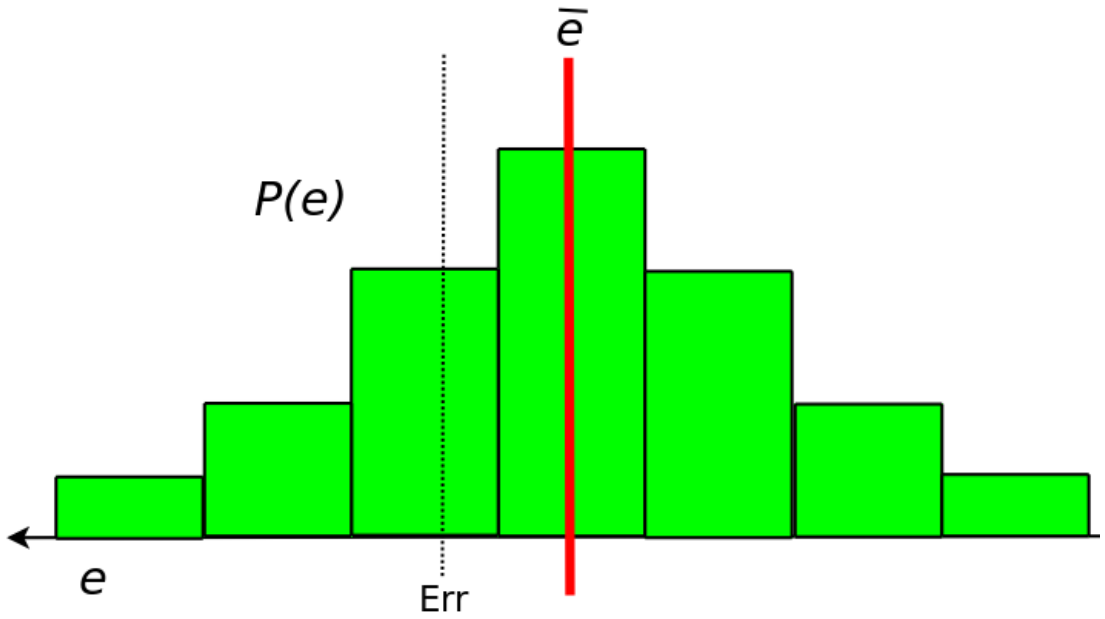
- Výsledek zkoušky e je náhodná veličina s rozdělením $P(e)$
- Střední hodnota $\bar{e} = \sum_e e \cdot P(e)$
- $\bar{e} = Err$: e je **nevychýleným** odhadem Err

Statistický odhad



- Předpokládejme, že schválně vybíráme těžké otázky.

Vychýlený odhad



- $\bar{e} \neq Err$: e je **vychýleným** odhadem Err
- "Nespravedlivá zkouška"

Výběr studenta

- Skupina studentů píše spravedlivou zkoušku

Výběr studenta

- Skupina studentů píše spravedlivou zkoušku
- Vybíráme **náhodného** studenta č. i s výsledkem zkoušky e_i
- Je e_i nevychýlený odhad Err_i?

Výběr studenta

- Skupina studentů píše spravedlivou zkoušku
- Vybíráme **náhodného** studenta č. i s výsledkem zkoušky e_i
- Je e_i nevychýlený odhad Err_i? **ANO**

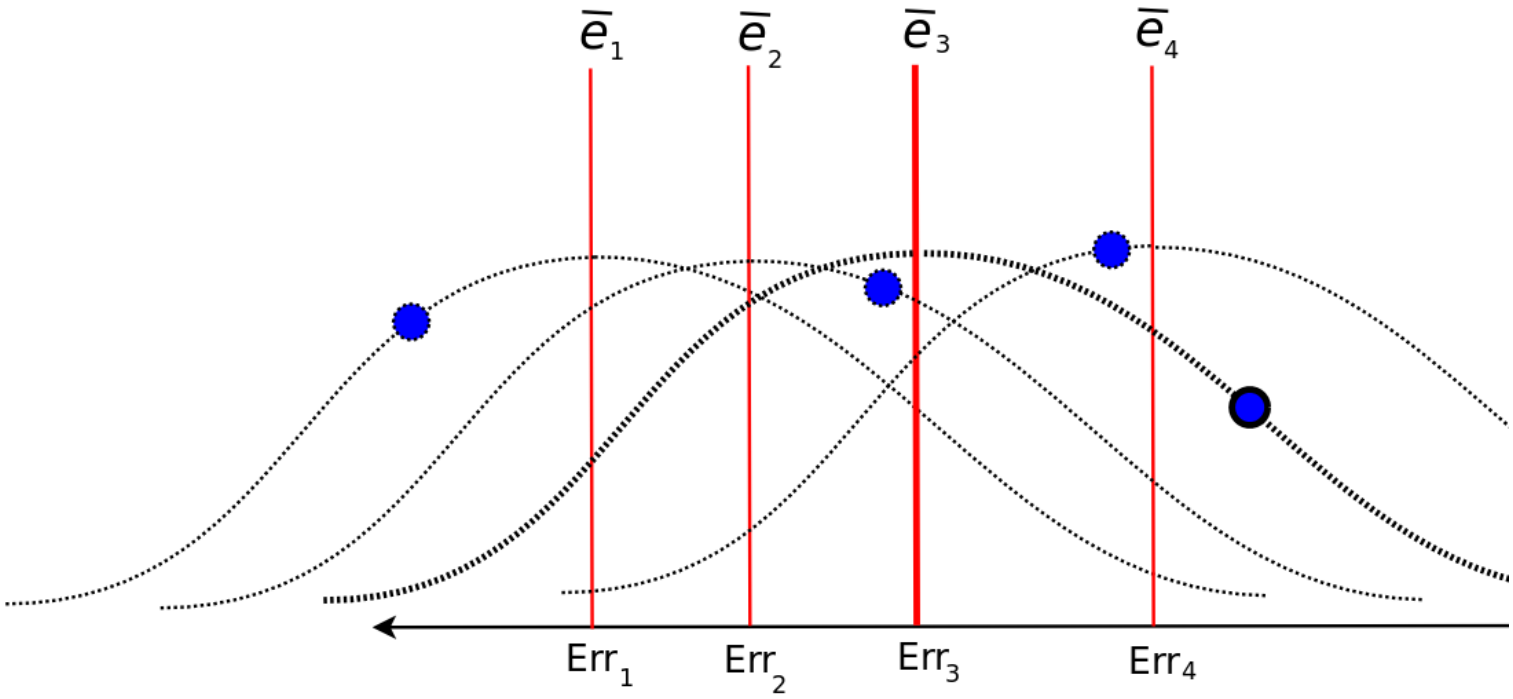
Výběr studenta

- Skupina studentů píše spravedlivou zkoušku
- Vybíráme studenta č. j s **nejlepším** výsledkem zkoušky e_j
- Je e_j nevychýlený odhad Err_j?

Výběr studenta

- Skupina studentů píše spravedlivou zkoušku
- Vybíráme studenta č. j s **nejlepším** výsledkem zkoušky e_j
- Je e_j nevychýlený odhad Err_j? **NE!**

Výchylka zavedená výběrem



- Výběrem nejlepšího e_j zavádíme výchylku!
- e_j již není nevychýleným odhadem Err_j
- přesto, že student j psal spravedlivou zkoušku

Výběr studenta

- Jak získat nevychýlený odhad pro již vybraného nejlepšího studenta?

Výběr studenta

- Jak získat nevychýlený odhad pro již vybraného nejlepšího studenta?
- Napíše novou zkoušku, otázky vybrány opět iid, nezávisle na předchozí zkoušce
- Její výsledek e'_j je nevychýleným odhadem E_j

Analogie

Student **Klasifikátor**

Analogie

Student	Klasifikátor
Chybovost studenta *Err*	**Skutečná chyba klasifikátoru** *Err*
Minimalizovaná výběrem studenta nejlépe ovládajícího látku. V praxi obvykle nevíme, který to je.	*Minimalizovaná výběrem klasifikátoru maximalizující aposteriorní pravděpodobnost. V praxi obvykle nevíme, který to je.*

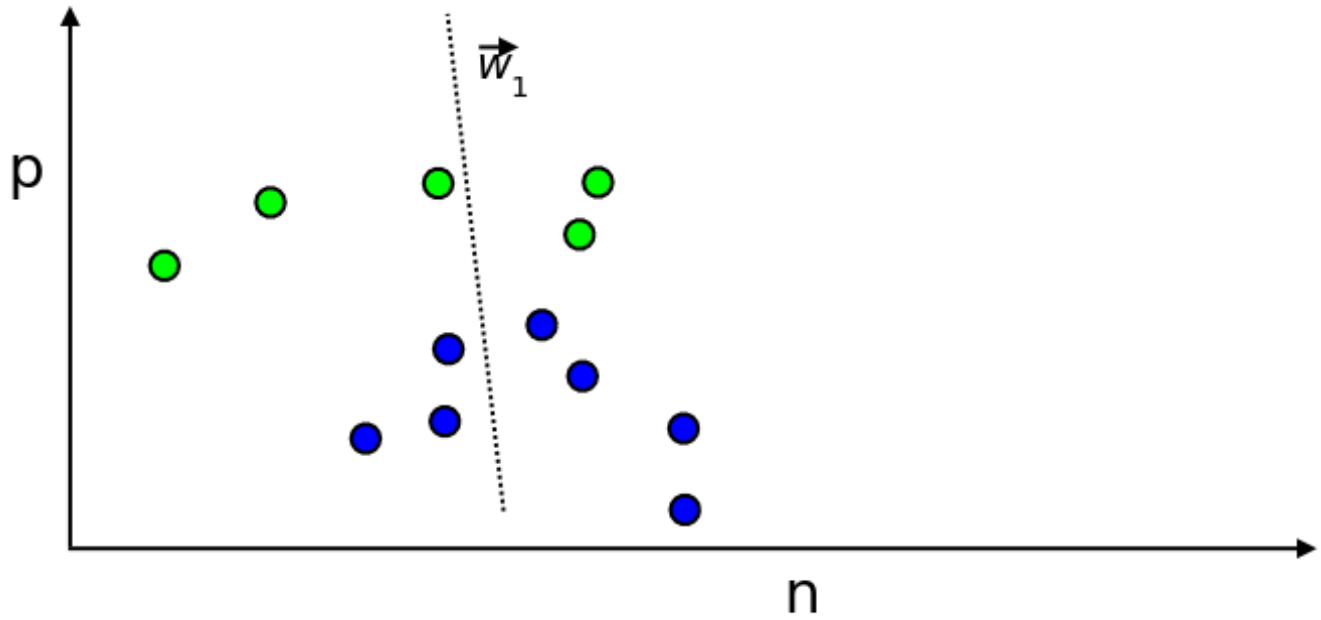
Analogie

Student	Klasifikátor
Chybovost studenta *Err*	**Skutečná chyba klasifikátoru** *Err*
Minimalizovaná výběrem studenta nejlépe ovládajícího látku. V praxi obvykle nevíme, který to je.	*Minimalizovaná výběrem klasifikátoru maximalizující aposteriorní pravděpodobnost. V praxi obvykle nevíme, který to je.*
Chyba u zkoušky *e*	**Trénovací chyba** *e*
Podíl špatně zodpovězených otázek zkoušky. Nevychýlený odhad Err, pokud dle e nevybíráme	*Podíl instancí chybně klasifikovaných instancí v trénovacích datech. Nevychýlený odhad Err, pokud dle e nevybíráme*

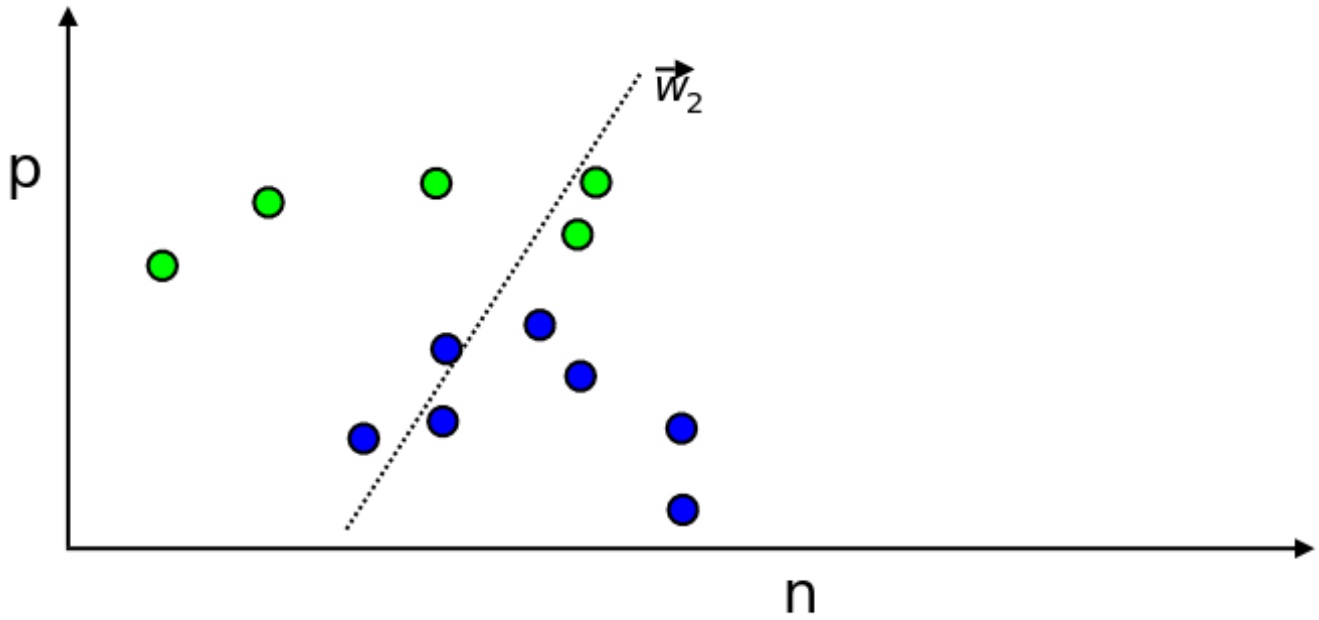
Analogie

Student	**Klasifikátor**
Chybovost studenta *Err* *Minimalizovaná výběrem studenta nejlépe ovládajícího látku. V praxi obvykle nevíme, který to je.*	**Skutečná chyba klasifikátoru** *Err* *Minimalizovaná výběrem klasifikátoru maximalizující aposteriorní pravděpodobnost. V praxi obvykle nevíme, který to je.*
Chyba u zkoušky *e* *Podíl špatně zodpovězených otázek zkoušky. Nevychýlený odhad Err, pokud dle e nevybíráme*	**Trénovací chyba** *e* *Podíl instancí chybně klasifikovaných instancí v trénovacích datech. Nevychýlený odhad Err, pokud dle e nevybíráme*
Výběr studenta	**Trénování (= výběr) klasifikátoru**

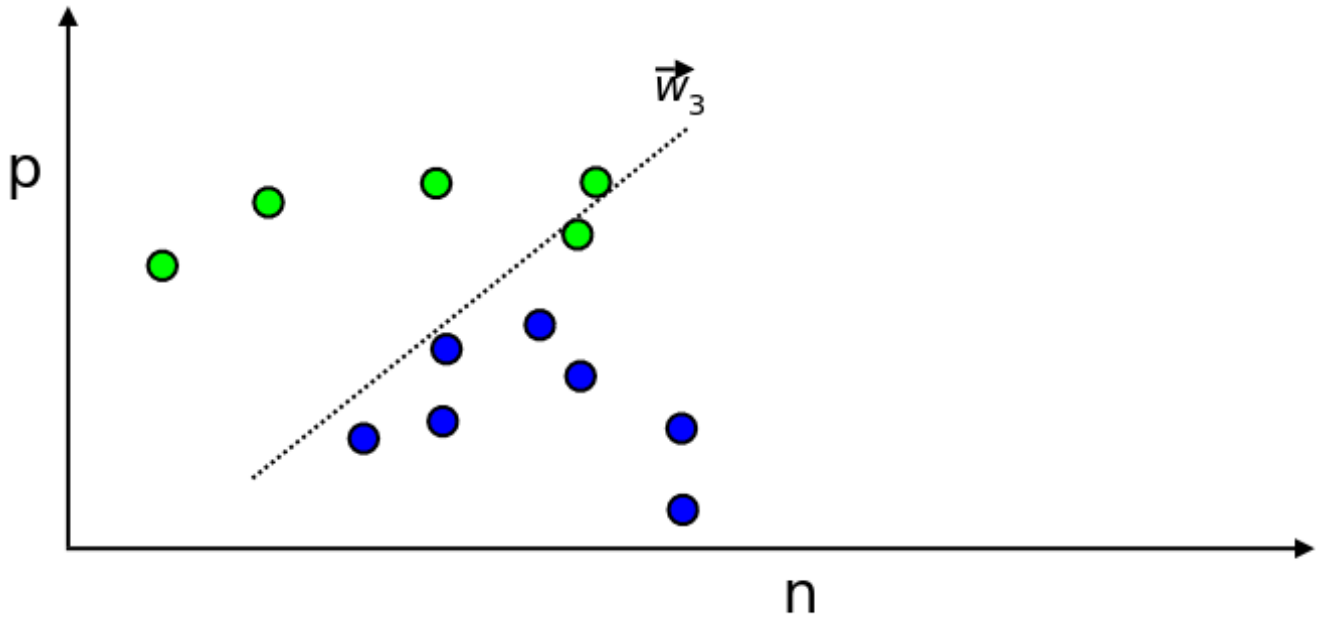
Trénování ≈ výběr



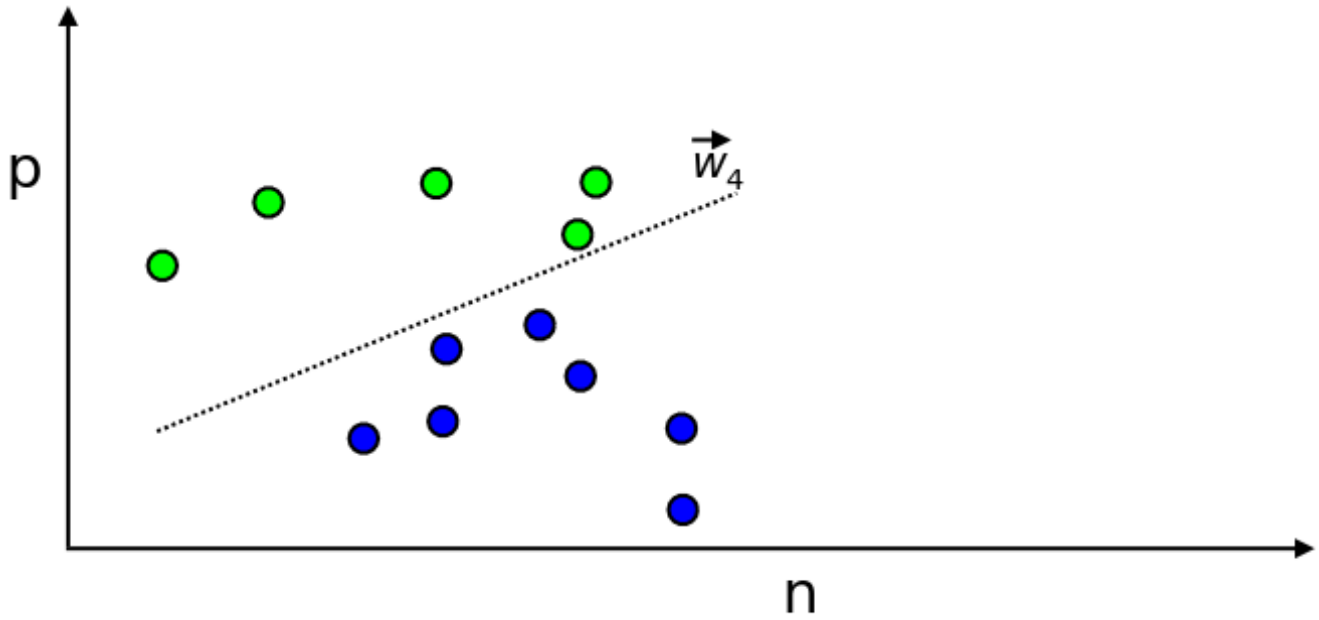
Trénování ≈ výběr



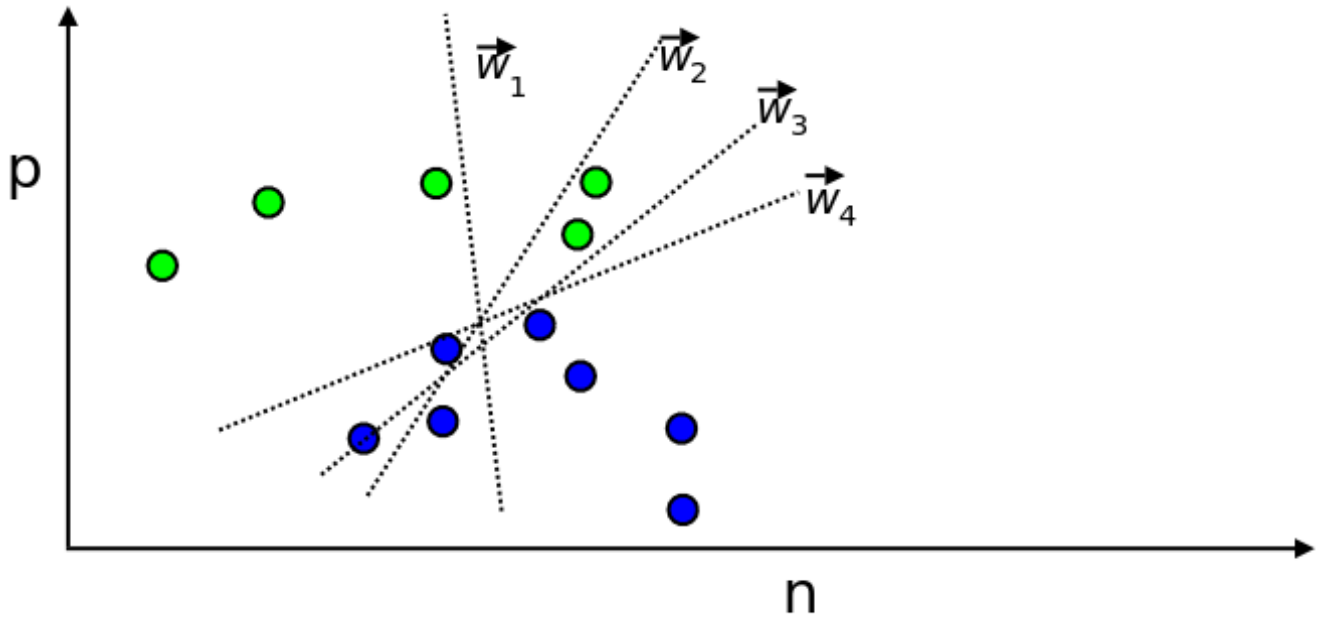
Trénování ≈ výběr



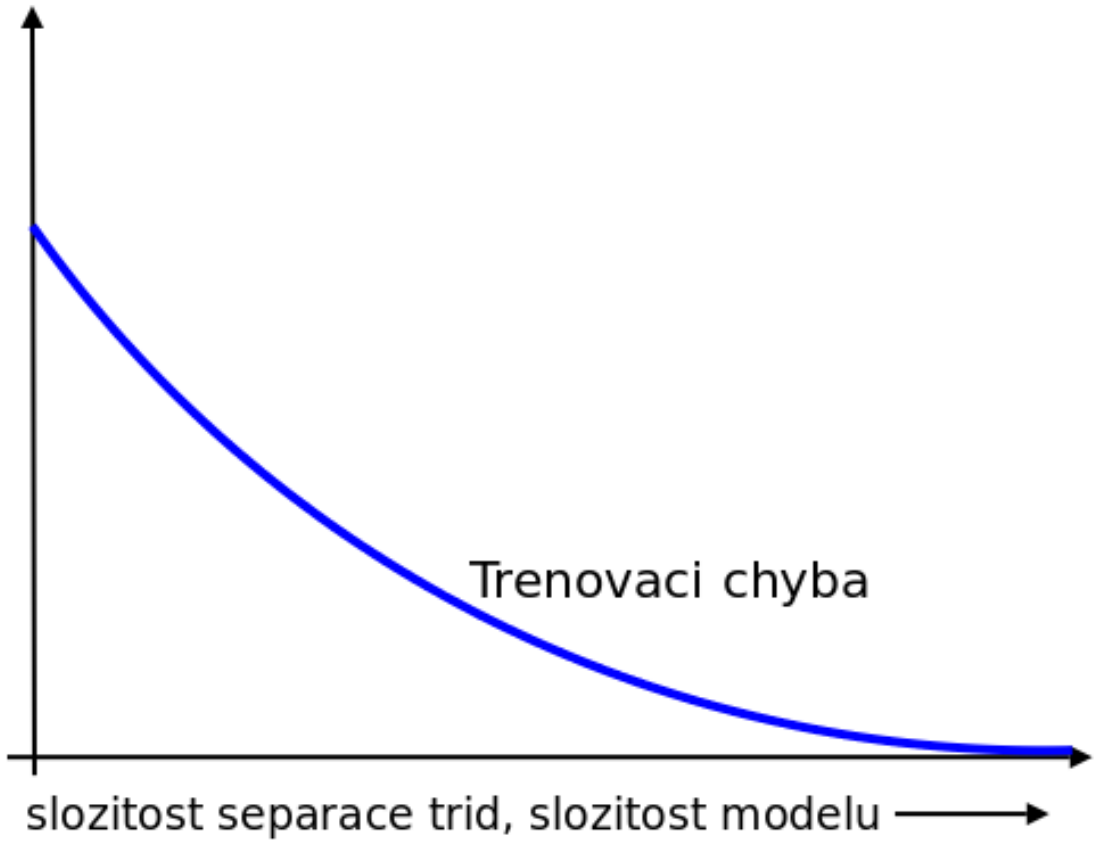
Trénování ≈ výběr



Trénování ≈ výběr



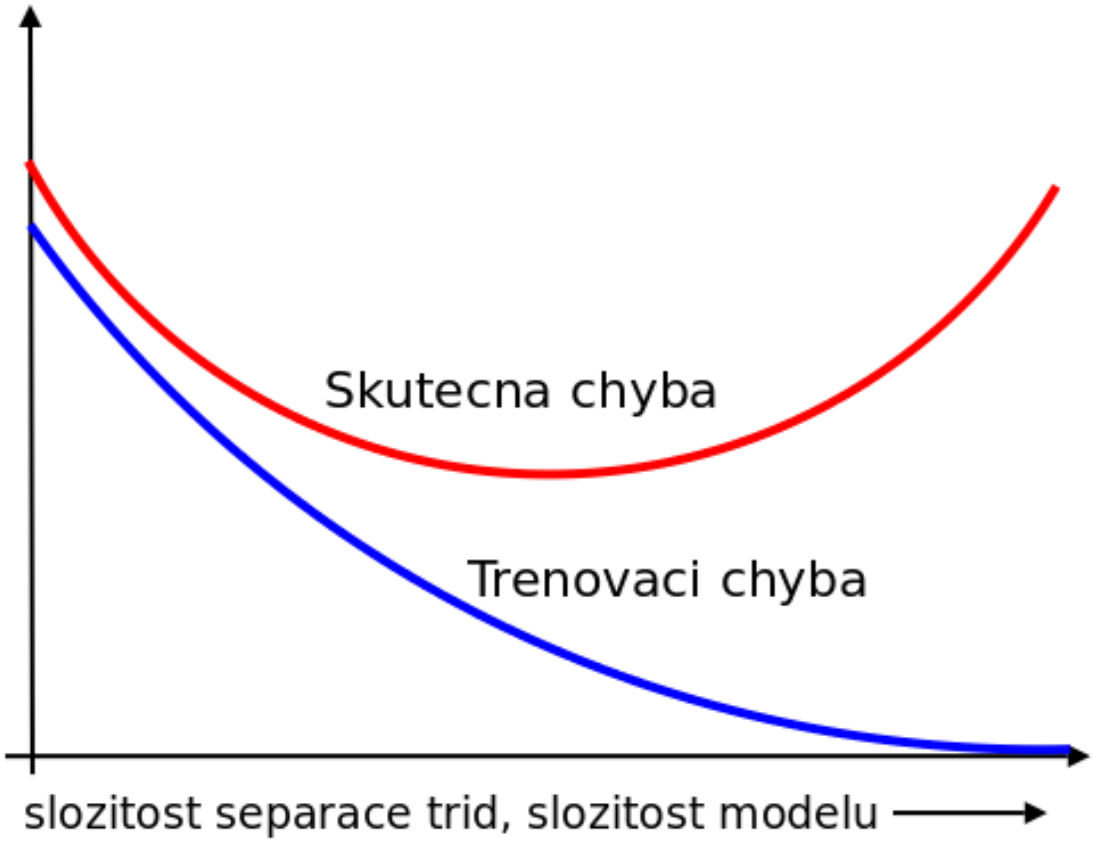
Analogie



Množství studentů, z nichž vybíráme

Množství klasifikátorů, z nichž vybíráme. Obykle úměrné maximální **složitosti** *klasifikátorů, kterou připouštíme*

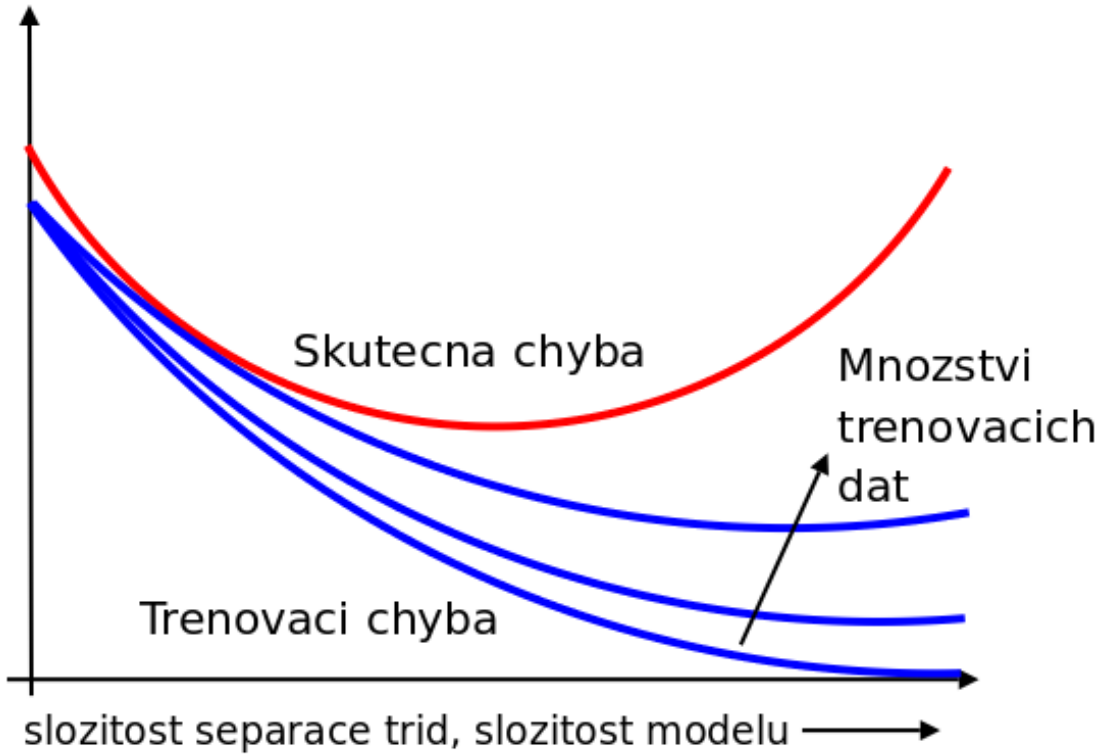
Analogie



Nová zkouška pro nevychýlený odhad Err vybraného studenta

Nová data, pro nevychýlený odhad skutečné chyby Err vybraného klasifikátoru

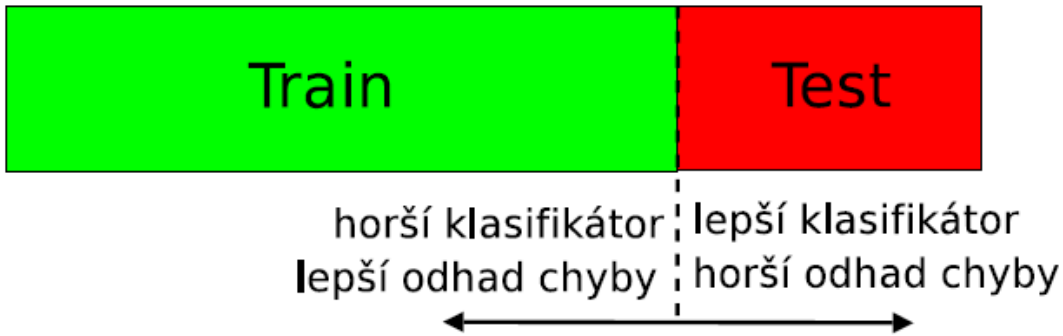
Analogie



Počet otázek ve zkoušce *Množství trénovacích dat*

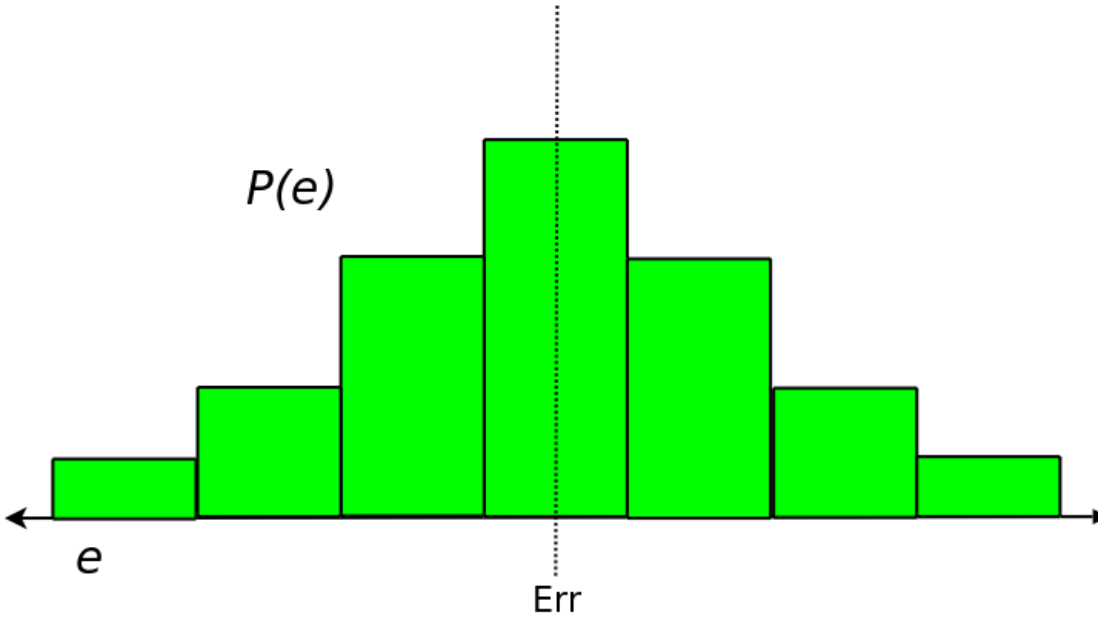
Testovací data

- Nová data pro nevychýlený odhad (tzv. testovací data) obvykle nemáme k dipozici
- Musíme použít část původních dostupných dat (a tím přijít kus trénovacích)



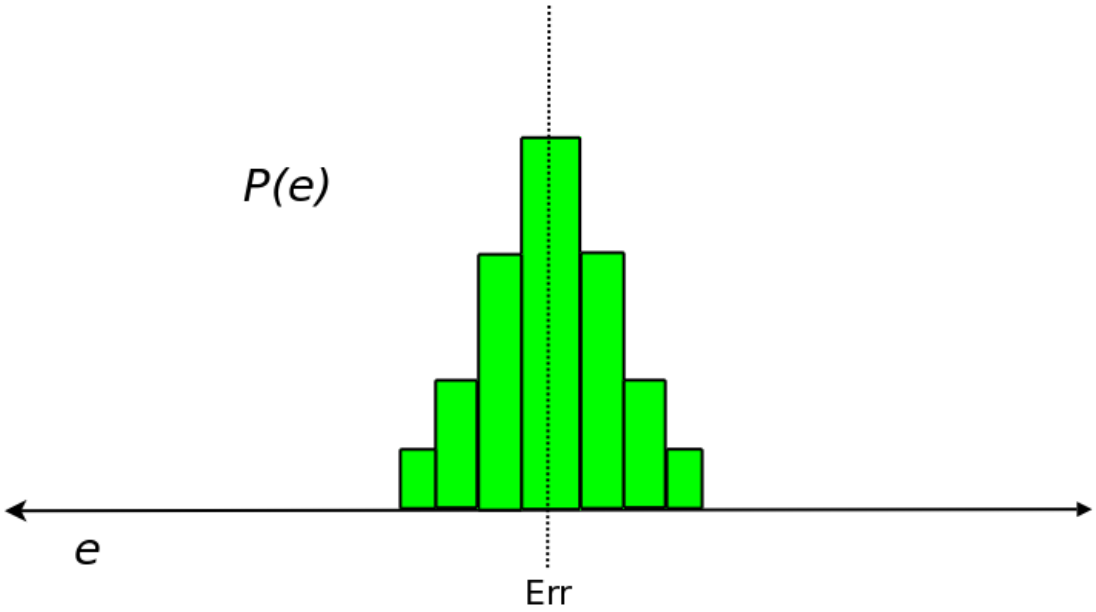
- Obvykle např. 70% vs. 30%
- "Horší" odhad: stále nevychýlený, ale větší rozptyl

Rozptyl odhadu



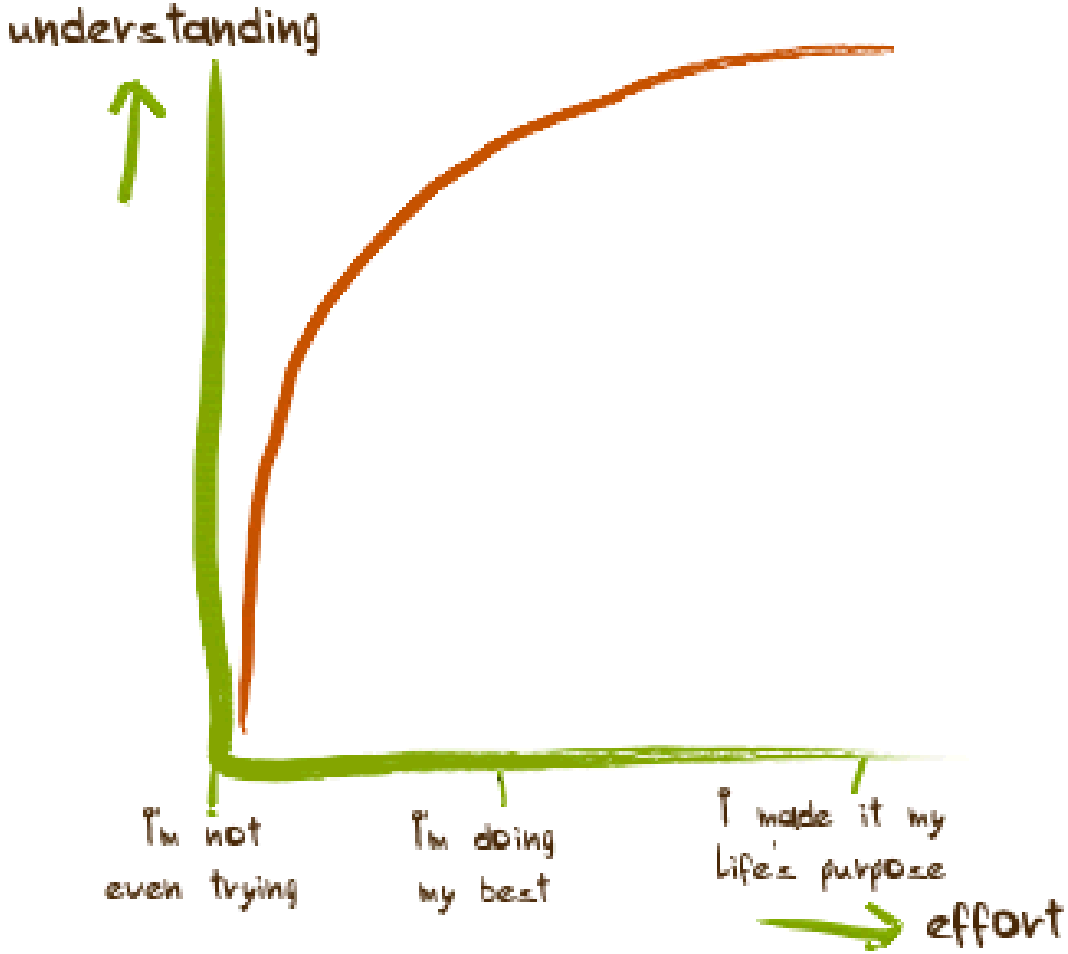
Velký rozptyl: **horší** nevychýlený odhad

Rozptyl odhadu



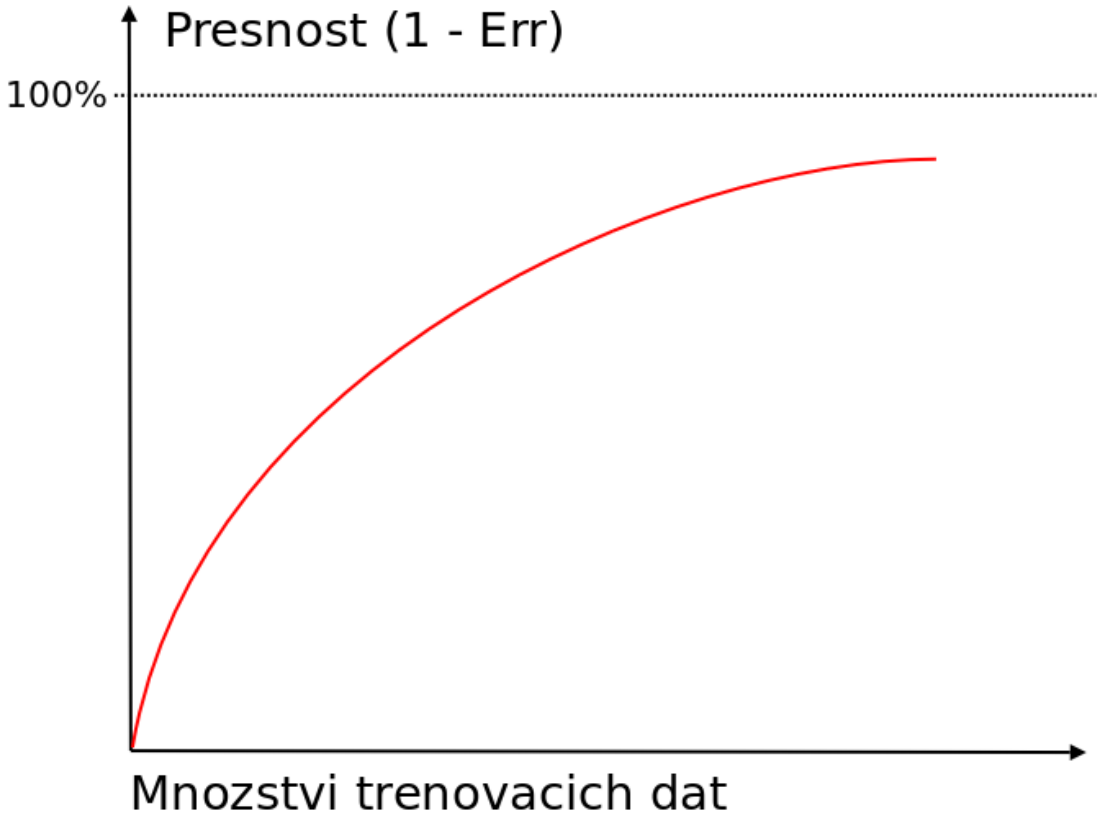
Malý rozptyl: **lepší** nevychýlený odhad

Křivka učení



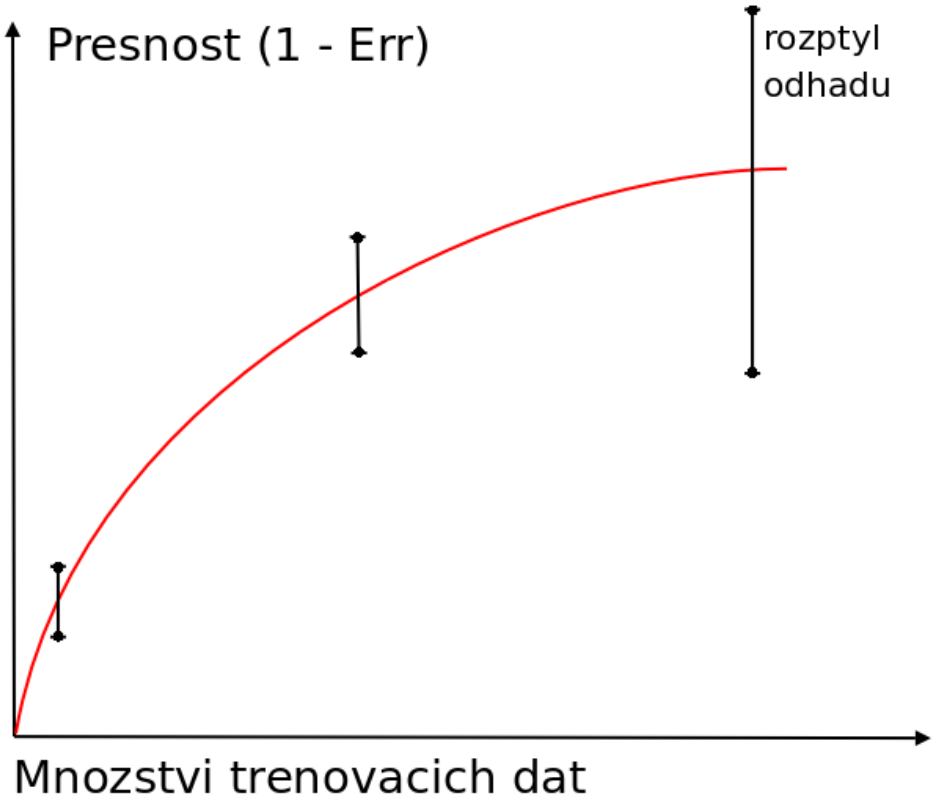
- Populární pojetí

Křivka učení



- Skutečná přesnost $1 - Err$ klasifikátoru v závislosti na množství trénovacích dat

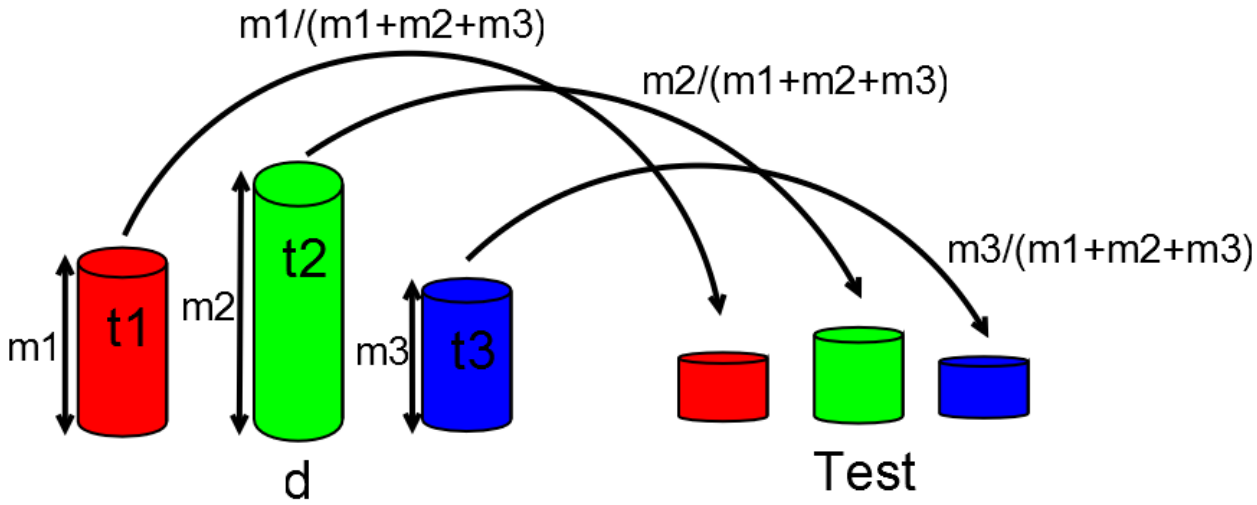
Křivka učení



- Odhadnutná přesnost $1 - e$ klasifikátoru (e na testovacích datech)
- Čím více trénovacích, tím méně testovacích, tím vyšší rozptyl e

Stratifikované dělení

Např. pro data se třemi třídami



- Stejné poměry velikostí tříd v trénovacích i testovacích datech

Souhrn

- Předpokládejme, že všechny chyby jsou **stejně drahé**.
- *Přednáška 2:* Optimální je klasifikace dle maximální aposteriorní pravděpodobnosti $P_{u|\vec{p}}(y|\vec{x})$.
- $(u, \vec{p}$ ... n.v. třídy, resp. příznaků)
  - Minimalizujeme tak skutečnou chybu *Err*

Souhrn

- Předpokládejme, že všechny chyby jsou **stejně drahé**.
- *Přednáška 2:* Optimální je klasifikace dle maximální aposteriorní pravděpodobnosti $P_{u|\vec{p}}(y|\vec{x})$.
- ($u, \vec{p}$ ... n.v. třídy, resp. příznaků)
  - Minimalizujeme tak skutečnou chybu *Err*
- V praxi nedosažitelné, obvykle neznáme rozdělení $P_{u|\vec{p}}$
- *Přednáška 3 a 4:* Vybíráme tedy z množiny klasifikátorů (např. všech lineárních) ten s nejmenší chybou e na trénovacích datech
  - Skutečnou chybu *Err* pak odhadneme na testovacích datech

Odhad rizika

- Předpokládejme, že chyby jsou **různě drahé** a máme zadánu ztrátovou funkci L.
- *Přednáška 2:* Optimální je klasifikace dle minimálního rizika

$$r_{u|\vec{p}} = \sum_t L(t, y) P_{u|\vec{p}}(t|x)$$

- V praxi nedosažitelné, obvykle neznáme rozdělení $P_{u|\vec{p}}$

Odhad rizika

- Na trénovací množině můžeme odhadnout riziko analogicky jako odhadujeme chybu

$$\frac{1}{m}\sum_{i=1}^{m} L(y_i, y_i')$$

  - y_i ... skutečná třída instance i,
  - y_i ... třída instance i určená klasifikátorem
  - m ... počet trénovacích dat
- I další postup analogický
  - vybíráme z množiny klasifikátorů (např. všech lineárních) ten s nejmenším odhadnutým rizikem na trénovacích datech
  - skutečné riziko pak odhadneme na testovacích datech.

Složky chyby při binárním problému

- Předpokládejme binární klasifikační problém: právě dvě třídy: ano, ne.
- Odhad rizika je potom:

$$\frac{1}{k}\sum_{i=1}^{m} L(y_i, y_i') = \frac{1}{k}\sum_{i=1}^{k} L(\text{ano}, \text{ne}) + \frac{1}{m-k}\sum_{i=k+1}^{m} L(\text{ne}, \text{ano})$$

- kde instance $1 \ldots k$ jsou klasifikované jako ano, ale ve skutečnosti jsou ne
  - tzv. **false positives (FP)**
- a instance $k+1 \ldots m$ jsou klasifikované jako ne, ale ve skutečnosti jsou ano
  - tzv. **false negatives (FP)**
- Analogicky: **true positives** (ano, ano) a **true negatives** (ne, ne). Ty se v riziku neprojeví, neboť $L(y, y) = 0$.

Matice záměn

	klasifikace	
skutečnost	ano	ne
ano	TP	FN
ne	FP	TN

- *TP* počet true positives
- *FP* počet false positives
- *TN* počet true negatives
- *FN* počet false negatives

Poměrné veličiny v [0;1]

	klasifikace	
skutečnost	ano	ne
ano	TP	FN
ne	FP	TN

- $TPr = TP/(TP + FN)$ true positive rate (recall, sensitivity) 
- $FPr = FP/(FP + TN)$ false positive rate 
- $TNr = TN/(TN + FP)$ true negative rate (specificity) 
- $FNr = FN/(TN + FP)$ false negative rate 
- $Pre = TP/(TP + FP)$ precision 

Výběr klasifikátoru

- Při binárních klasifikačních problémech vybíráme model na základě kompromisu mezi dvěma veličinami z $\{TPr, FPr, TNr, FNr, Pre\}$. Např.
  - odhad chyby Err je $FPr + FNr$
  - odhad rizika r je $a \cdot FPr + b \cdot FNr$, kde a a b jsou koeficienty úměrné $L(\text{ano}, \text{ne})$ resp $L(\text{ne}, \text{ano})$.
- Proč "kompromis"?
  - Každou jednotlivou veličinu z $\{TPr, FPr, TNr, FNr, Pre\}$ lze minimalizovat triválně!
  - Jak?
- tzv. míra F1 (F1 measure)

$$\frac{2 \cdot Pre \cdot Tpr}{Pre + Tpr}$$

  - Kompromis mezi Precision a Recall, vhodná pro data s velmi nerovnoměrným rozdělením tříd

Přizpůsobení klasifikátoru

- *TPr* naučeného klasifikátoru lze obvykle snadno zvýšit (snížit) na úkor (ve prospěch) FPr
- Např pro lineární klasifikátor namísto

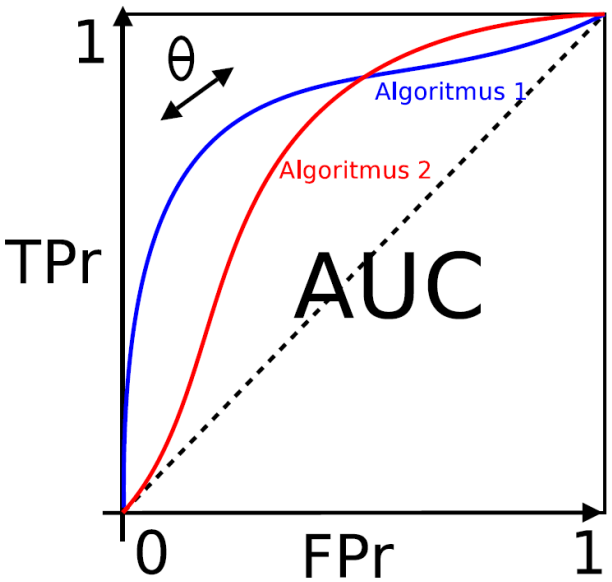
$$ax_1 + bx_2 + c > 0$$

uvažujeme

$$ax_1 + bx_2 + c > \theta$$

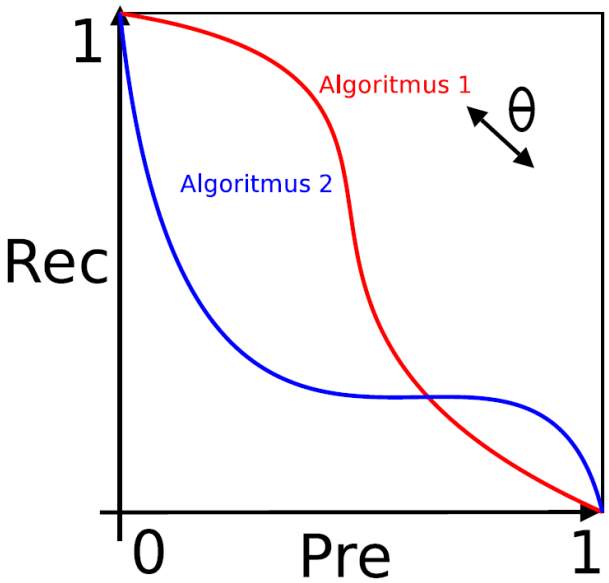
- Dostaneme parametrizovaný klasifikátor $k_\theta(\vec{x}) \mapsto \{\text{ano,ne}\}$
- Mějme dva různé parametrizované klasifikátory k_θ, k'_θ
- Který z nich je lepší?
  - Pro každé θ to může být jinak!

Analýza ROC



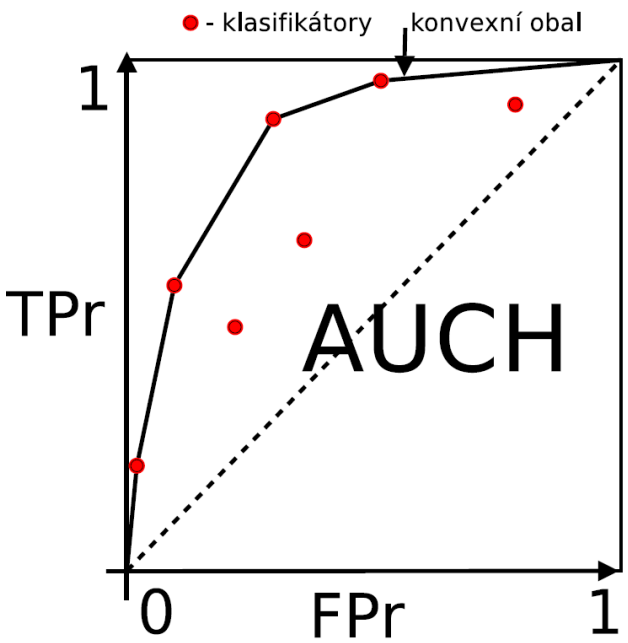
- "Receiver Operating Characteristics"
- Křivka složená ze všech bodů (Fpr, Tpr) pro měnící se θ
- AUC (Area Under ROC Curve) = měřítko nezávislé na θ

Analýza Precision-Recall



- Analogicky křivka Precision vs. Recall

Analýza ROC pro konečnou množinu klasifikátorů



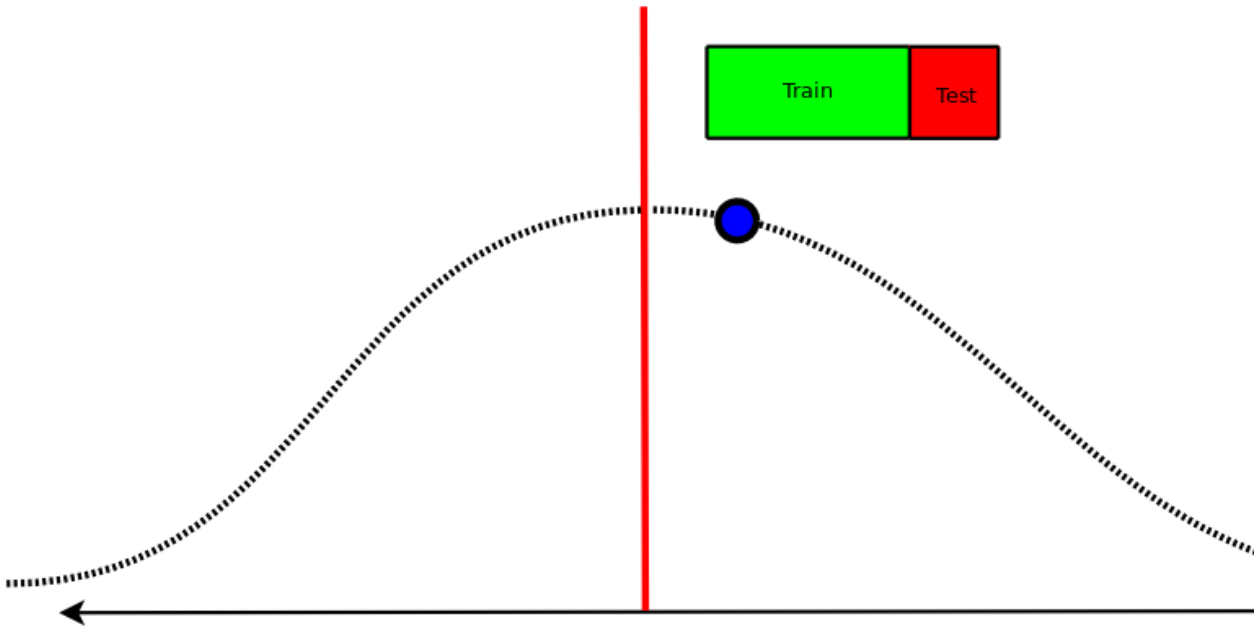
- Kvalita množiny modelů $\approx$ plocha pod konvexním obalem bodů
- AUCH – Area Under ROC Convex Hull

Připomínka

- Všechny dosud jmenované veličiny
  - FP, TP, TN, FN
  - TPr, FPr, TNr, FNr, Pre
  - e, F1
  - ROC křivka, AUC, AUCH, Precision-Recall křivka

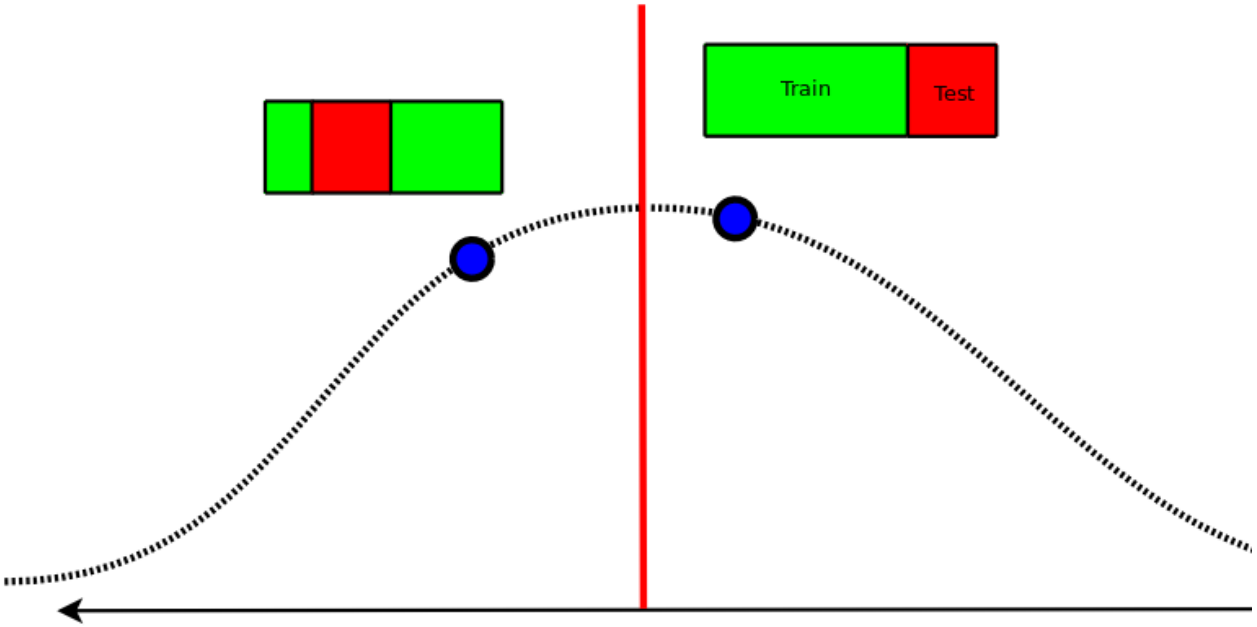
  můžeme počítat jak na trénovacích, tak na testovacích datech.
- Na **trénovacích**: **vychýlené** odhady skutečnosti!
  - Využijeme pouze pro výběr klasifikátoru
- Na **testovacích**: **nevychýlené** odhady skutečnosti!
  - Využijeme pro odhad skutečné kvality klasifikátoru

Rozptyl odhadu



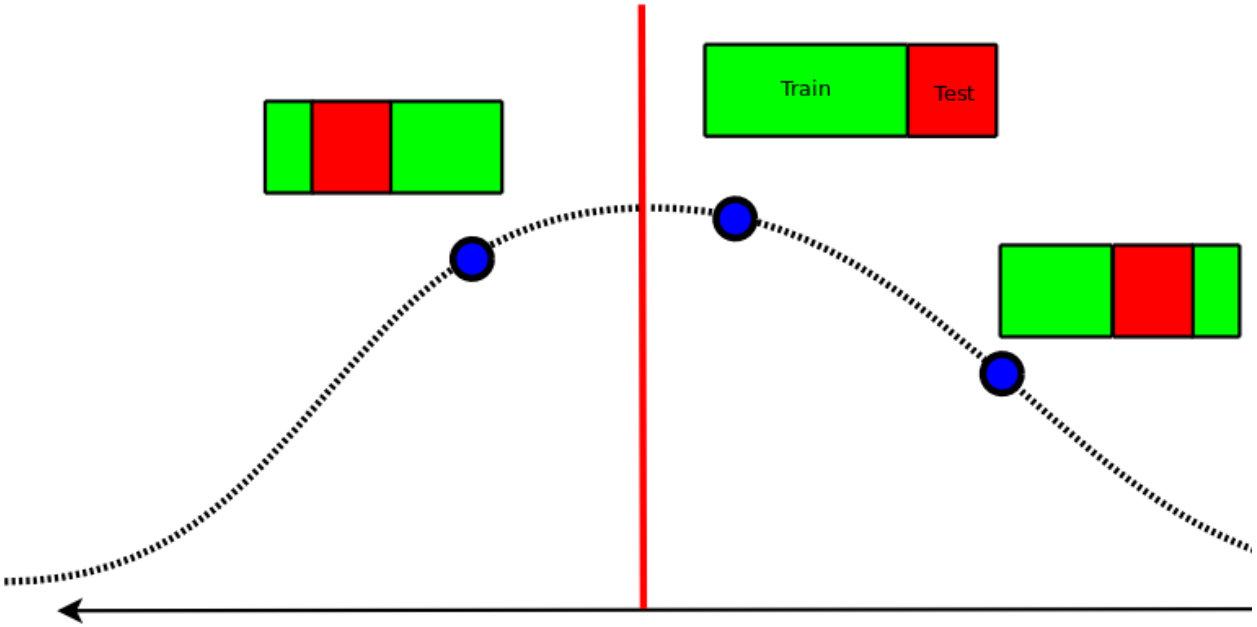
Náhodné rozdělení na trénovací a testovací data

Rozptyl odhadu



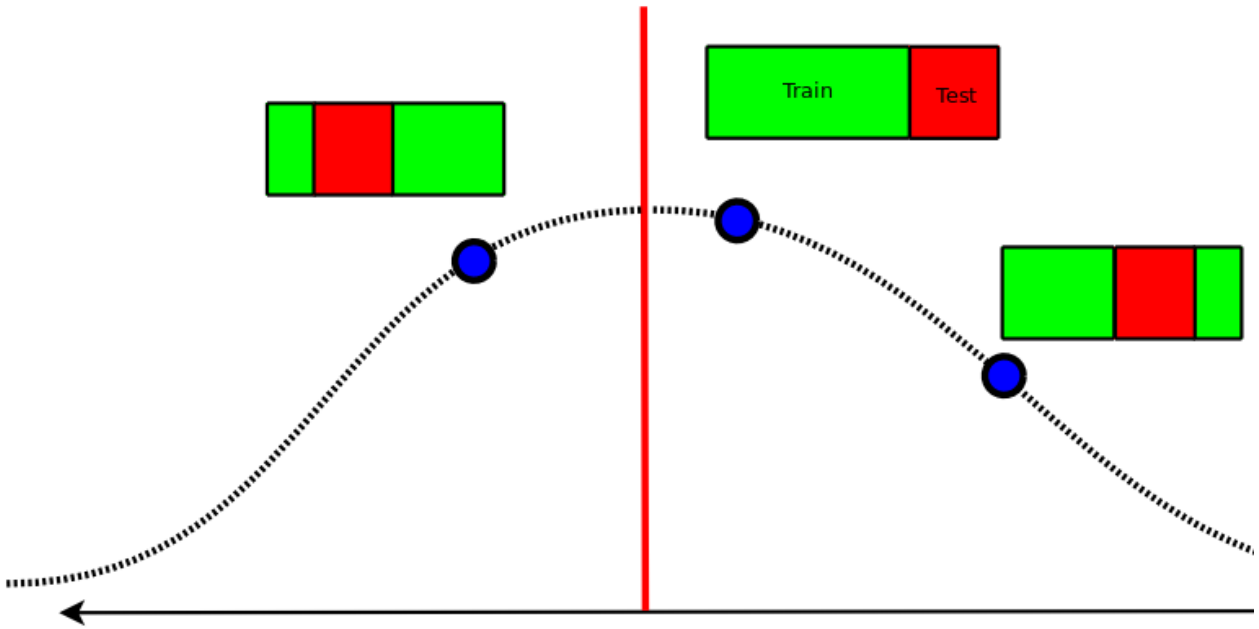
Jiné rozdělení na trénovací a testovací data, jiný odhad

Rozptyl odhadu



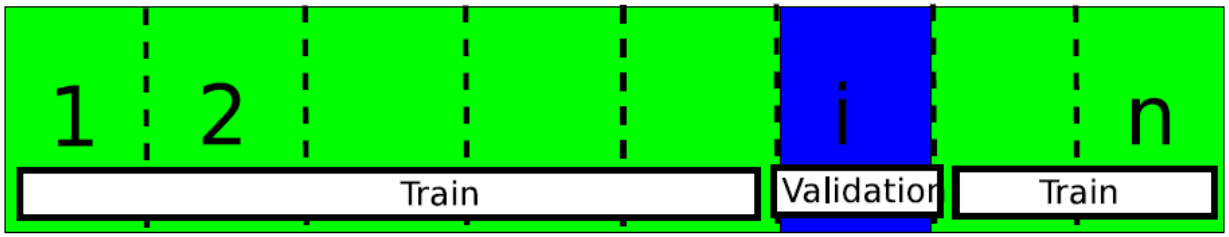
Jiné rozdělení na trénovací a testovací data, jiný odhad

Rozptyl odhadu



Šlo by snížit rozptyl zprůměrováním odhadů z několika různých train/test rozdělení?

Křížová validace



- n-složková křížová validace
- Rozlož trénovací množinu X na n stejně velkých složek X_i (folds) náhodným výběrem

$$X = \cup_i^n X_i$$

$$X_i \cap X_j = \emptyset$$

- Pro $i = 1 \ldots n$
  - Sestroj klasifikátor na $X_1 \cup \ldots X_{i-1} \cup X_{i+1} \ldots X_n$
  - Spočítej veličinu na X_i
- Zprůměruj výsledky: $1/n \sum_i X_i$

Křížová validace

- Pro každé i jiný klasifikátor!
  - Neodhadujeme veličinu pro konkrétní klasifikátor, ale pro

Křížová validace

- Pro každé *i* jiný klasifikátor!
  - Neodhadujeme veličinu pro konkrétní klasifikátor, ale pro algoritmus, který klasifikátory konstruuje
- *Stratifikovaná* křížová validace
  - Rozdělení četnosti tříd totožné v každé složce
  - Viz stratifikované rozdělení na train/test
- Křížová validace *leave-one-out*
  - extrémní případ: počet složek = počet dat
- Otázky
  - Odhadujeme-li chybu křížovou validací, jak závisí odhad na počtu složek?
  - Je odhad chyby křížovou validací nevychýleným odhadem chybovosti klasifikátorů konstruovaných validovaným algoritmem?

Výběr a testování

K.V. můžeme použít pro výběr algoritmu resp. parametru

- např. pro stanovení nejlepšího k pro klasifikaci dle sousedů
- nebo pro stanovení stupně polynomu pro polynomiální klasifikaci
- nebo kombinace obojího atd.

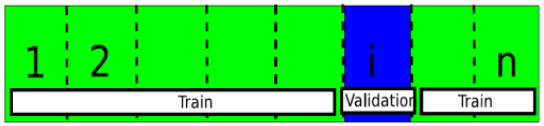
Provedeme K.V. pro všechny 'soutěžící' verze a vybereme verzi s nejlepším výsledkem K.V.

Výběr a testování

Úplný postup pro výběr algoritmu křížovou validací, získání klasifikátoru a odhad jeho kvality

1. Rozdělíme data na Train / Test

2. Křížovou validací na Train vybereme algoritmus



3. Zvoleným algoritmem sestrojíme klasifikátor na Train

4. Jeho kvalitu odhadneme na Test