Machine Learning Meetups

Marek Modrý

modry.marek@gmail.com

Learning to Rank

Co je a proč?

Learning to Rank

Learning to Rank

Příklad – Hledání **relevantního dokumentu** k danému **dotazu**

Learning to Rank

DOKUMENTY	PŘÍZNAKY
Dokument d1	(. . . . . vektor příznaků d1. . . . .)
Dokument d2	(. . . . . vektor příznaků d2. . . . .)
Dokument d3	(. . . . . vektor příznaků d3. . . . .)
Dokument d4	(. . . . . vektor příznaků d4. . . . .)
Dokument d5	(. . . . . vektor příznaků d5. . . . .)

Learning to Rank

DOKUMENTY	PŘÍZNAKY	RELEVANCE
Dokument d1	(. . . . . vektor příznaků d1. . . . .)	relevance y1
Dokument d2	(. . . . . vektor příznaků d2. . . . .)	relevance y2
Dokument d3	(. . . . . vektor příznaků d3. . . . .)	relevance y3
Dokument d4	(. . . . . vektor příznaků d4. . . . .)	relevance y4
Dokument d5	(. . . . . vektor příznaků d5. . . . .)	relevance y5

Learning to Rank

DOKUMENTY	PŘÍZNAKY	RELEVANCE
Dokument d1	(. . . . . vektor příznaků d1. . . . . .)	relevance y1
Dokument d2	(. . . . . vektor příznaků d2. . . . . .)	relevance y2
Dokument d3	(. . . . . vektor příznaků d3. . . . . .)	relevance y3
Dokument d4	(. . . . . vektor příznaků d4. . . . . .)	relevance y4
Dokument d5	(. . . . . vektor příznaků d5. . . . . .)	relevance y5

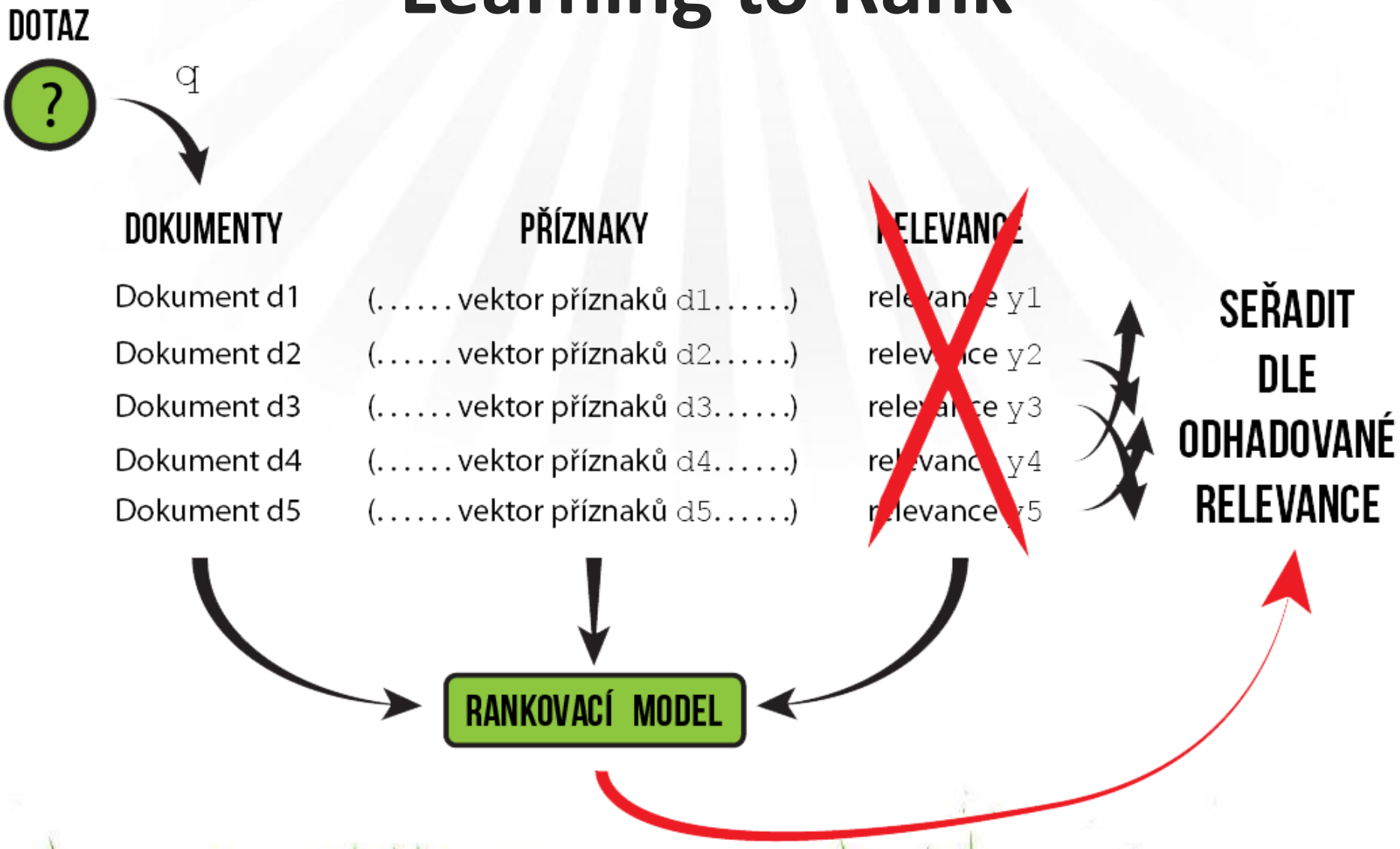
SEŘADIT DLE RELEVANCE

Learning to Rank

DOKUMENTY	PŘÍZNAKY	RELEVANCE
Dokument d1	(. . . . . vektor příznaků d1. . . . . .)	relevance y1
Dokument d2	(. . . . . vektor příznaků d2. . . . . .)	relevance y2
Dokument d3	(. . . . . vektor příznaků d3. . . . . .)	relevance y3
Dokument d4	(. . . . . vektor příznaků d4. . . . . .)	relevance y4
Dokument d5	(. . . . . vektor příznaků d5. . . . . .)	relevance y5

SEŘADIT DLE RELEVANCE

Learning to Rank



Learning to Rank úloha

- Není to klasifikace
- Není to regrese
- **Je to ŘAZENÍ**

Learning to Rank úloha

- Není to klasifikace
- Není to regrese
- **Je to ŘAZENÍ**

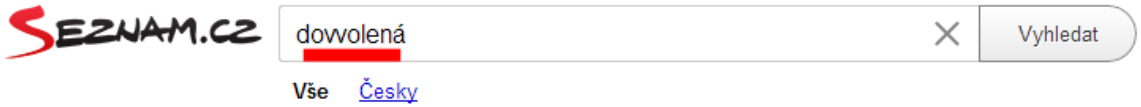
…ALE…

Problémy L2R

Problémy a výzvy

- Používají se **komplikované metriky** pro stanovení kvality řazení
- **Řazení** způsobuje **nediferenciovatelnost**
  - > nelze jednoduše optimalizovat gradientními metodami
  - > obchází se předefinováním problému (zpět ke klasifikaci a regresi) nebo různými triky

Learning to Rank - aplikace



Kaggle – KDD Cup 2013

Completed • $7,500 • 554 teams

KDD Cup 2013 - Author-Paper Identification Challenge (Track 1)

Thu 18 Apr 2013 – Wed 26 Jun 2013 (16 months ago)

- **Cíl**: Stanovit, který autor napsal jaký článek
- **Data**: páry autor-článek -> confirmed/deleted
- **Výstup**: seřadit seznam kandidátských článků tak, aby nahoře byly ty nejpravděpodobněji napsané

Způsoby řešení

Způsoby řešení

- Neuronové sítě
- Boosting (AdaBoost, AdaRank, RankBoost)
- SVM
- Random Forest
- Aktuálně top:
  - Multiple additive regression trees (MART)

MART

- Každým přidaným stromem se posouváme blíže k optimu, opravujeme chyby …
- Až 1000 malých stromů (hloubka 5)
- Data o velikosti 1 000 000 dokumentů

Prostor pro vaše dotazy

Děkuji za pozornost

Marek Modrý, modry.marek@gmail.com