



Machine Learning Meetups

Marek Modrý

modry.marek@gmail.com



Learning to Rank





Co je a proč?

Learning to Rank

Learning to Rank

DOTAZ



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



Learning to Rank

DOTAZ



q



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

DOTAZ



q

DOKUMENTY

Dokument d1

Dokument d2

Dokument d3

Dokument d4

Dokument d5

PŘÍZNAKY

(.....vektor příznaků d1.....)

(.....vektor příznaků d2.....)

(.....vektor příznaků d3.....)

(.....vektor příznaků d4.....)

(.....vektor příznaků d5.....)

RELEVANCE

relevance y_1

relevance y_2

relevance y_3

relevance y_4

relevance y_5

Learning to Rank

DOTAZ



q



DOKUMENTY

Dokument d1

Dokument d2

Dokument d3

Dokument d4

Dokument d5

PŘÍZNAKY

(.....vektor příznaků d1.....)

(.....vektor příznaků d2.....)

(.....vektor příznaků d3.....)

(.....vektor příznaků d4.....)

(.....vektor příznaků d5.....)

RELEVANCE

relevance y_1

relevance y_2

relevance y_3

relevance y_4

relevance y_5



**SEŘADIT
DLE
RELEVANCE**

Learning to Rank

DOTAZ



q

DOKUMENTY

Dokument d1
Dokument d2
Dokument d3
Dokument d4
Dokument d5

PŘÍZNAKY

(.....vektor příznaků d1.....)
(.....vektor příznaků d2.....)
(.....vektor příznaků d3.....)
(.....vektor příznaků d4.....)
(.....vektor příznaků d5.....)

~~RELEVANCE~~

~~relevance y_1~~
~~relevance y_2~~
~~relevance y_3~~
~~relevance y_4~~
~~relevance y_5~~



**SEŘADIT
DLE
RELEVANCE**

Learning to Rank

DOTAZ



q

DOKUMENTY

Dokument d1
Dokument d2
Dokument d3
Dokument d4
Dokument d5

PŘÍZNAKY

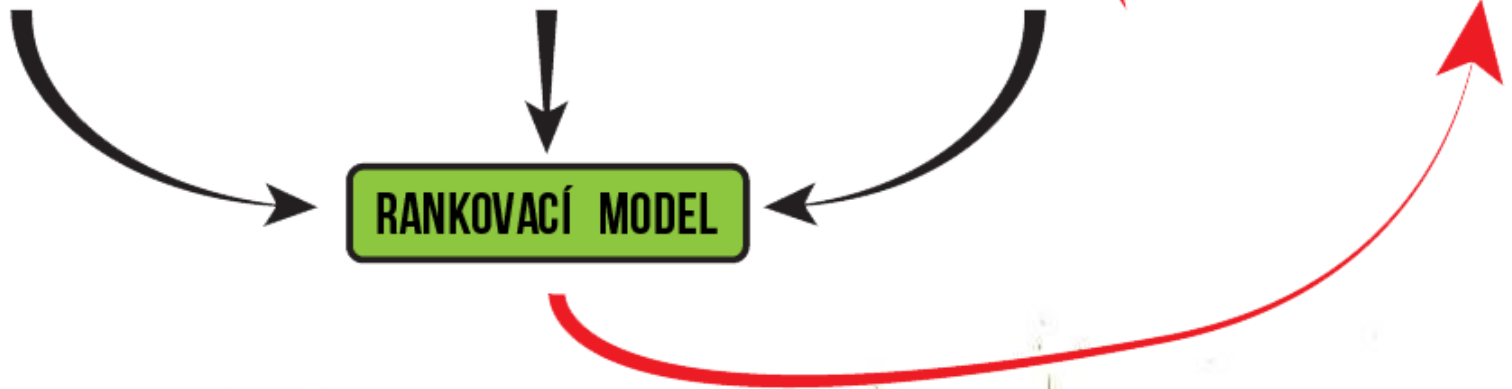
(.....vektor příznaků d1.....)
(.....vektor příznaků d2.....)
(.....vektor příznaků d3.....)
(.....vektor příznaků d4.....)
(.....vektor příznaků d5.....)

~~RELEVANCE~~

~~relevance y_1~~
~~relevance y_2~~
~~relevance y_3~~
~~relevance y_4~~
~~relevance y_5~~

**SEŘADIT
DLE
ODHADOVANÉ
RELEVANCE**

RANKOVACÍ MODEL



Learning to Rank úloha

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

Learning to Rank úloha

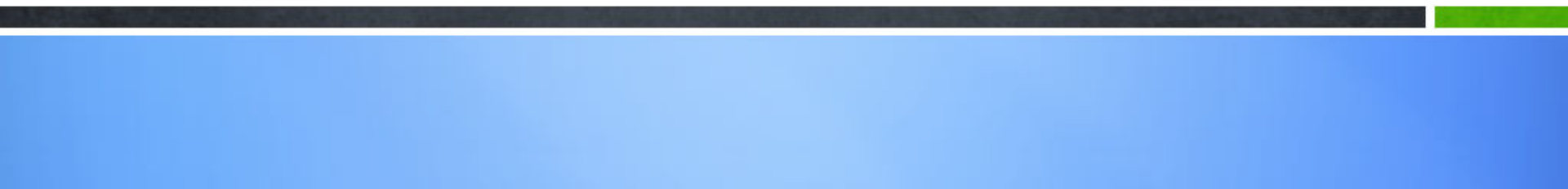
- Není to klasifikace
- Není to regrese
- **Je to ŘAZENÍ**

...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

SEZNAM.CZ

dovolená



Vyhledat

Vše [Česky](#)

Google
Česká republika

Hledali jsme **dovolená**
Přesto můžeme hledat dovolená

amazon
Try Prime

Recommendations for You in Books



Bayesian Reasoning and Machine Learning

> David Barber

Hardcover

★★★★★ (10)

\$92.00 \$82.80

Why recommended?

> [See more recommendations](#)



Programming Collective Intelligence...

> Toby Segaran

Paperback

★★★★★ (105)

\$30.99 \$26.67

Why recommended?

Expedia.ie™



Sa...

Ven...
(01) 524 5005 [\(Opening hours\)](#)
Booked in the last 14 hours



Hilton Garden Inn Calabasas ★★★

Calabasas
4.6 out of 5 (286 reviews)
(01) 524 5005 [\(Opening hours\)](#)
Booked in the last 11 hours



Palm Garden Hotel ★★★

Thousand Oaks
4.0 out of 5 (692 reviews)
(01) 524 5005 [\(Opening hours\)](#)
Booked in the last 10 hours

Kaggle – KDD Cup 2013



Completed • \$7,500 • 554 teams

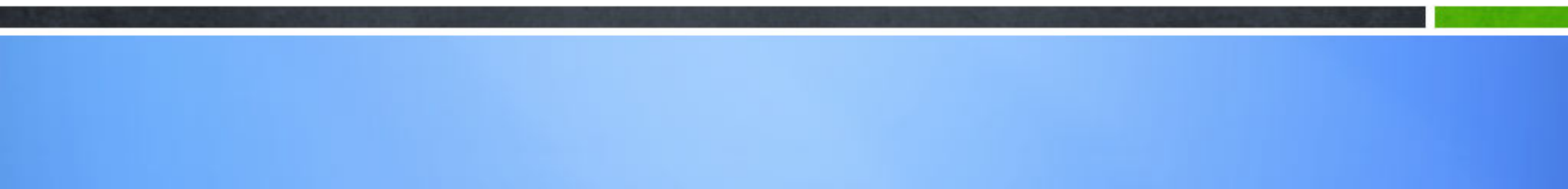
KDD Cup 2013 - Author-Paper Identification Challenge (Track 1)

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 přidáním 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