Factorization Machines

Jakub Pachocki

Carnegie Mellon University

10-805 class talk

FACTORIZATION MACHINES

- A beautiful cross between Matrix Factorization and SVMs
- Introduced by Rendle in 2010

KAGGLE DOMINANCE

#	Δrank	Team Name ‡model uploaded * in the money		Score	Entries	Last Submission UTC (Best – Last Submission)
1	—	3 Idiots ‡ * • guestwalk • mandora • yolicat	(FM)	0.44464	13	Tue, 23 Sep 2014 23:31:16
2	—	Michael Jahrer and Jeong-Yoon Lee *		0.44527	61	Tue, 23 Sep 2014 23:37:47 (-9.7d)
3	—	beile ‡ *		0.44610	67	Tue, 23 Sep 2014 23:07:36 (-0.8h)

#	Δrank	Team Name * in the money		Score	Entries	Last Submission UTC (Best – Last Submission)
1	—	4 Idiots * • guestwalk • Michael Jahrer • yolicat • mandora	(FM)	0.3791384	273	Mon, 09 Feb 2015 19:37:27 (-43.6h)
2	—	Owen *		0.3803652	94	Mon, 09 Feb 2015 02:34:23 (-0.5h)
3	—	Random Walker *		0.3806351	242	Mon, 09 Feb 2015 10:59:10

AD CLASSIFICATION

Clicked?	Country	Day	Ad_type
1	USA	3/3/15	Movie
0	China	1/7/14	Game
1	China	3/3/15	Game

ONE-HOT ENCODING

Clicked?	Country= USA	Country= China	Day= 3/3/15	Day= 1/7/14	Ad_type =Movie	Ad_type =Game
1	1	0	1	0	1	0
0	0	1	0	1	0	1
1	0	1	1	0	0	1

AD CLASSIFICATION

- Very large feature space
- Very sparse samples

- Should we run SGD now?

POLY-2 KERNEL

- Often features are more important in pairs
- e.g. „Country=USA" ^ „Day=Thanksgiving"
- Create a new feature for every pair of features
- Feature space: *insanely* large
- Samples: still sparse

SHARPENING OCCAM'S RAZOR

- We cannot learn a weight for every possible pair of features because of memory constraints

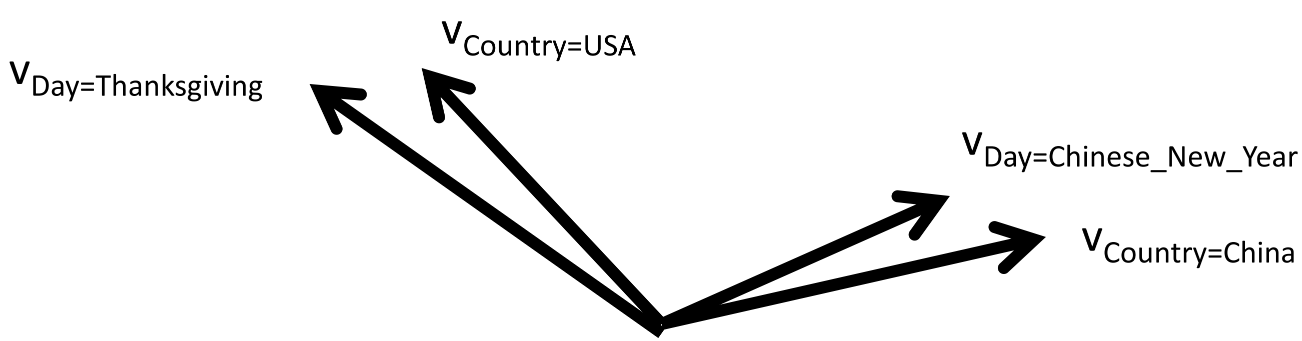
- Even if we could (SVMs?), we might overfit massively

FACTORIZATION MACHINES

- Let $w_{i,j}$ be the weight assigned to feature pair (i,j)
- **Key idea:** Set $w_{i,j} = \langle v_i, v_j \rangle$
- v_is are vectors in *k*-dimensional space

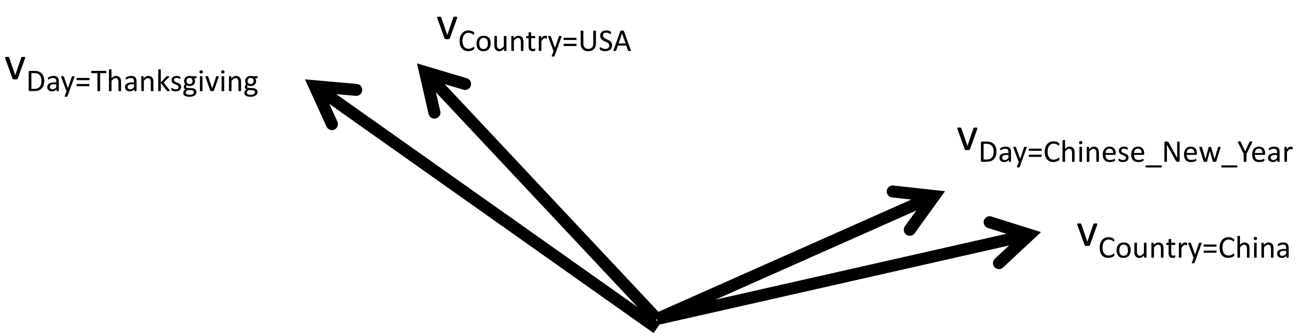
FACTORIZATION MACHINES

- Let $w_{i,j}$ be the weight assigned to feature pair (i,j)
- **Key idea:** Set $w_{i,j} = \langle v_i, v_j \rangle$



SVMS MEET FACTORIZATION

- The idea is that weights between different pairs of features are not entirely independent
- Their dependence is described by latent factors



MATRIX FACTORIZATION RECAP

$$
n \text{ users} \left[\overbrace{\begin{pmatrix} r11 & \dots & & \\ \dots & \dots & & \\ & rij & & \\ & & \dots & \\ & & & rnm \end{pmatrix}}^{m \text{ movies}} \right.
$$

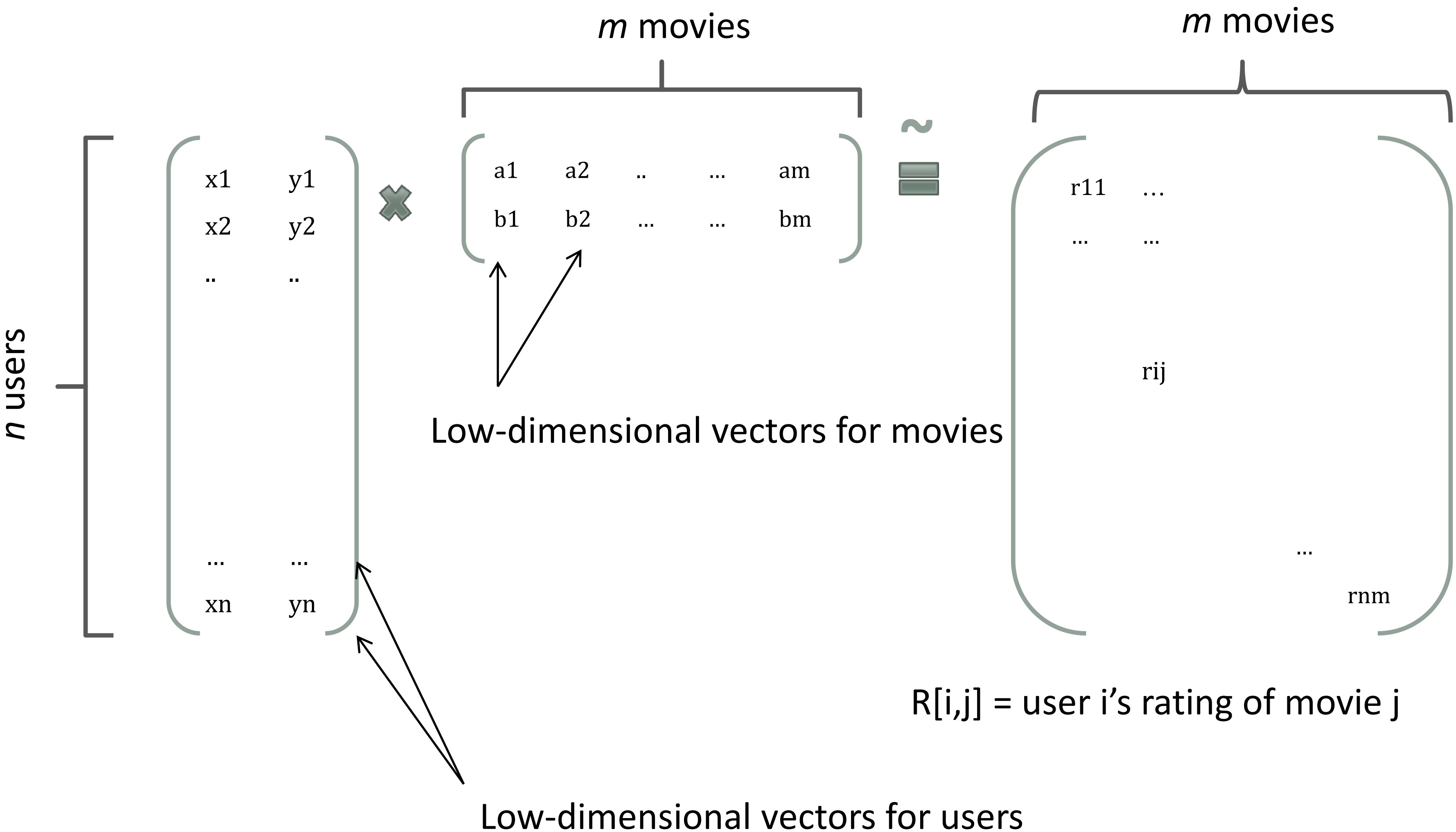
R[i,j] = user i's rating of movie j

MATRIX FACTORIZATION RECAP

$$
n \text{ users} \left[\begin{pmatrix} x1 & y1 \\ x2 & y2 \\ .. & .. \\ & \\ ... & ... \\ xn & yn \end{pmatrix} \right. \times \overbrace{\begin{pmatrix} a1 & a2 & .. & ... & am \\ b1 & b2 & ... & ... & bm \end{pmatrix}}^{m \text{ movies}} \approx \overbrace{\begin{pmatrix} r11 & ... & & & \\ ... & ... & & & \\ & rij & & & \\ & & & ... & \\ & & & & rnm \end{pmatrix}}^{m \text{ movies}}
$$

R[i,j] = user i's rating of movie j

MATRIX FACTORIZATION RECAP



FMS AND MATRIX FACTORIZATION

Rating?	User=Alice	User=Bob	User=Jane	Movie= Titanic	Movie= Avatar
1	1	0	0	0	1
0	0	1	0	1	0
1	0	0	1	1	0

FMS AND MATRIX FACTORIZATION

Rating?	User=Alice	User=Bob	User=Jane	Movie= Titanic	Movie= Avatar
1	1	0	0	0	1
0	0	1	0	1	0
1	0	0	1	1	0

- **Equivalent!** The latent factors for user and movie feature weights yield the factorization.

FMS AND SVMS

Clicked?	Country= USA	Country= China	Day= 3/3/15	Day= 1/7/14	Ad_type =Movie	Ad_type =Game
1	1	0	1	0	1	0
0	0	1	0	1	0	1
1	0	1	1	0	0	1

FMS AND SVMS

Clicked?	Country= USA	Country= China	Day= 3/3/15	Day= 1/7/14	Ad_type =Movie	Ad_type =Game
1	1	0	1	0	1	0
0	0	1	0	1	0	1
1	0	1	1	0	0	1

- What if we set *k* very large?

FMS AND SVMS

Clicked?	Country= USA	Country= China	Day= 3/3/15	Day= 1/7/14	Ad_type =Movie	Ad_type =Game
1	1	0	1	0	1	0
0	0	1	0	1	0	1
1	0	1	1	0	0	1

- **Equivalent!** For *k* large enough we can express any pairwise interactions between features.

FACTORIZATION MACHINES

- Generalize SVMs and Matrix Factorization (and many other models)
- Can be learned in linear time using SGD

BONUS: BUT HOW DO I WIN AT KAGGLE?

- (other than months of feature engineering)
- Use knowledge of the original fields the features come from!
- e.g. Country might have a different relationship to Date than to Ad_type

FIELD-AWARE FACTORIZATION MACHINES

- Learn a *different* set of latent factors for every pair of fields
- Instead of $<v_i, v_j>$ we use $<v_{i,f(j)}, v_{j,f(i)}>$, where f(i) is the field feature i comes from

MISCELLANEOUS

- Use hash trick!
- ... and regularization

- Generating more features: GBDT, neural nets, etc...

THANK YOU!

Questions?