GENERALIZED LINEAR MODELS IN COLLABORATIVE FILTERING

HAO WU

STANFORD UNIVERSITY

Generalized Linear Models

Generalization of linear regression from normal distribution to exponential family

Model components

- **distribution from the exponential family**
- **linear parameter**
- **link function**

$$g(y) = \eta = x^T \beta$$

- **sufficient statistic**

$$(y_1,\ y_2,\ \ldots y_n \,|\, x) \equiv (t,\ n \,|\, x)$$

Example: linear regression, logistic regression

COLLABORATIVE FILTERING

	Alt. Linear Regression	**Alt. Logistic Regression**
distribution	normal	binomial
link function	$\mu = \mathbf{UM}^T$	$\log\left(\mathbf{P}/(1-\mathbf{P})\right) = \mathbf{UM}^T$
loss function	square error	logistic loss
sufficient stat.	sum	# of 1s
application	direct feedbacks (rating)	indirect feedbacks (click)

$$\min_{\mathbf{U},\mathbf{M}} \quad \sum_{i=0}^{N} L(y_i, \mathbf{u}_{u_i}^T \mathbf{m}_{m_i})$$

$$+ \lambda\big[\alpha(\sum_{i}^{n_U} \|\mathbf{u}_i\|_1 + \sum_{i}^{n_M} \|\mathbf{u}_i\|_1) + (1-\alpha)(\|\mathbf{U}\|_F + \|\mathbf{M}\|_F)\big]$$

$$\text{s.t.} \qquad \mathbf{U} \in \mathbb{R}^{n_U \times k}, \ \mathbf{M} \in \mathbb{R}^{n_M \times k}$$

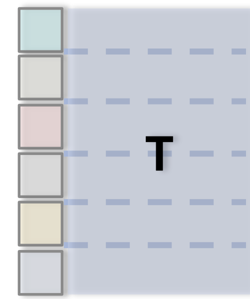
DISTRIBUTED ALG.

(assuming n_m × k fits in a single machine)

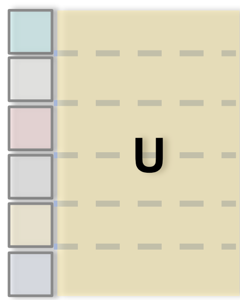
RDD[u, m, y]
Input entries

reduceByKey
+
aggregateByKey

RDD[u, map(m -> (t, n))]
Sufficient Stat.



RDD[u, vector]
User features



Array[vector]
Movie features

for each iteration:

- Join **U** with **T** to from **D** (co-partitioned join)
- Update **M**
- Broadcast **M** (communication: log(p)(n_Mk))
- Update **U**

Update M

for each movie:

- prepare dataframe by filter() and map()
- distributed logistic regression
    - LogisticRegression()
    - all-to-one and one-to-all of size k

Update U

Map local logistic regression to users

- added local training method to LogisticRegression()
- no communication of data

Summary

- Sparsity is preserved
- Scales in n_U, but not n_M or k
- Communication cost: $\log(p)(n_M k)$
- Computational depth: $\log(n_U)(n_M k)$