Learning Preferences with Millions of Parameters by Enforcing Sparsity

Presenter:

Xi Chen[1], Bing Bai[2], Yanjun Qi[2], Qihang Lin[1], Jaime G. Carbonell[1]

1. Machine Learning Department, Carnegie Mellon University
2. NEC Lab America

Motivation

- Ranking: match a query to document
  (find documents that are most relevant to the query)
- Most of the current methods use hand-coded features
- Our Goal: learning to rank directly from words [B. Bai et al. 09]

- Vector Space Model (query/document as vector)

$$\boxed{q, d = [w_1, \ldots, w_{\mathcal{D}}] \in \mathbb{R}^{\mathcal{D}}}$$

D: vocabulary size

w_i: normalized weight (tf-idf) of i-th word

- Cosine similarity (relevance of a document to a query)

$$\boxed{f(q, d) = q^T d}$$

- Does not deal with synonyms ☹
- No machine learning ☹

Basic Model

[D. Grangier 08]
[B. Bai et al. 09]

- Similarity Score (relevance of documents to a query)

$$f(q,d) = q^{\top} W d = \sum_{i,j} W_{ij}(q_i \cdot d_j)$$

W_{ij}: relationship/correlation between q_i and d_j

- Goal: Learn W matrix

- Generalization ability: $W \in \mathbb{R}^{\mathcal{D}\times\mathcal{D}}$ $\mathcal{D}$: vocabulary size

  $\mathcal{D} = 10,000$, $\mathcal{D}^2 = 10^8$ free parameters to model

- Memory issues: $\mathcal{D} = 10,000$: W needs 1GB Memory
- Computational cost: $q^T W d$

W: Sparse Matrix ! ! !

Training Framework

- Data:

  Tuples $\mathcal{R}$: query q, related doc. d^+, unrelated doc. d^-.

- Learn *W* such that:

$$f(q, d^+) = q^\top W d^+ > f(q, d^-) = q^\top W d^-$$

- Margin rank loss:

$$L_W(q, d^+, d^-) = \max(0, 1 - q^T W d^+ + q^T W d^-)$$

- Model:

$$W^* = \arg\min_W \frac{1}{|\mathcal{R}|} \sum_{(q,d^+,d^-)\in\mathcal{R}} L_W(q, d^+, d^-) + \lambda \|W\|_1.$$

Sparsity Enforcing Regularization

Training Algorithm

- Stochastic (sub)Gradient Descent:

$$\nabla L_{W^t}(q, d^+, d^-) = \begin{cases} -q(d^+ - d^-)^\top & \text{if} \quad q^\top W^t (d^+ - d^-) < 1 \\ 0 & \text{otherwise} \end{cases}.$$

$$W^{t+1} = W^t - \eta_t \nabla L_{W^t}(q, d^+, d^-)$$

$\eta_t = \frac{C}{\sqrt{t}}$: decaying rate

$\eta_t = C$: fixed rate

- Mini-batch Shrinkage Strategy (every T iterations):

$$\widehat{W}^t = \text{argmin}_W \frac{1}{2} \|W - W^t\|_F^2 + \lambda \sum_{k=t-T+1}^{t} \eta_t \|W\|_1$$

$$\Longrightarrow \widehat{W}^t_{ij} \to 0$$

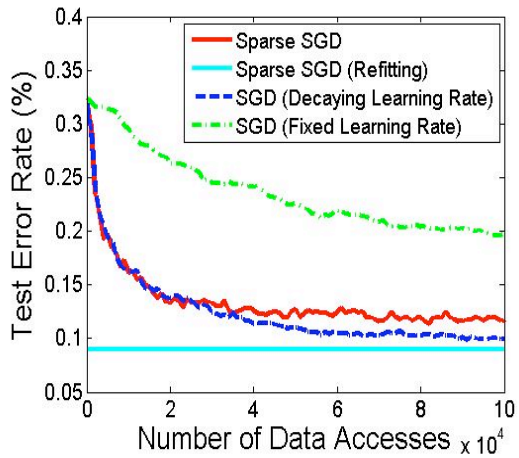
- Refitting Step (Reduce the bias of ℓ_1 regularization) :

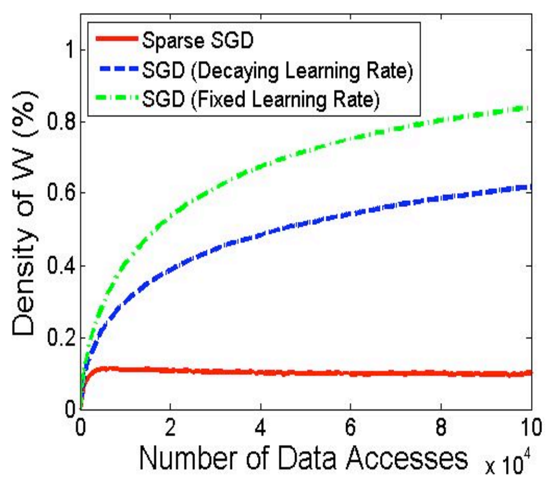
Fixing zeros elements and training the remaining elements without L1-regularization

Experimental Results

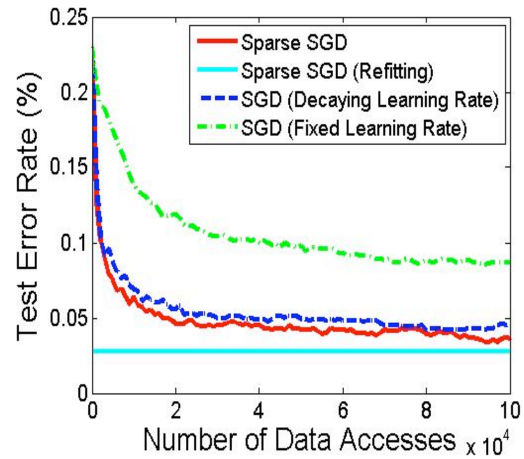
- Ranking Performance (multi-class classification data, relate doc. are in the same class)

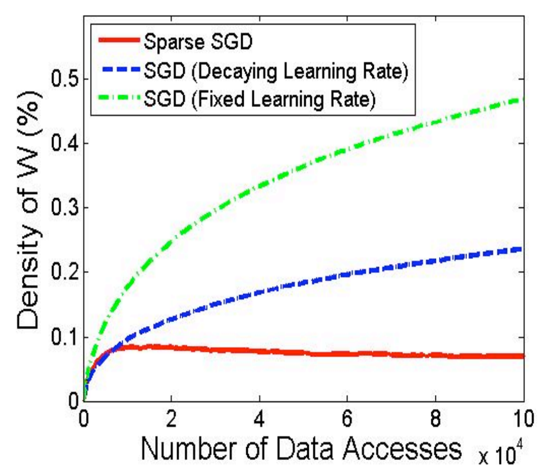
20 News Group (20NG)





RCV1





$\mathcal{D} = 10,000$

Experimental Results

20NG

	MAP	Test Error (%)	Memory (MB)
Identity	0.185	32.3	0.2
Diagonal	0.190	31.8	0.2
SGD (fixed learning rate)	0.258	19.7	1294
SGD (decaying learning rate)	0.399	9.9	943.1
Sparse	0.360	11.4	154.2
Sparse (refitting)	0.426	9.0	154.2

RCV1

	MAP	Test Error (%)	Memory (MB)
Identity	0.380	23.0	0.2
Diagonal	0.390	22.3	0.2
SGD (fixed learning rate)	0.451	8.7	717.2
SGD (decaying learning rate)	0.453	4.6	360.2
Sparse	0.463	3.6	105.4
Sparse (refitting)	0.501	2.9	105.4

$\mathcal{D} = 10,000$

Experimental Results

- Learned word-relationship from Sparse W (20 News Group)

Query word	Most related document words
atheism	keith atheists god caltech
clinton	government health people gay
cpu	mac drive scsi card
graphics	tiff image color polygon
handgun	gun weapons militia fbi
hockey	game espn colorado team
religions	god bible christian jesus