

Online VB for LDA in VW

Matt Hoffman, Columbia Dept. of Statistics

John Langford, Yahoo! Research

LDA (Blei et al. 2003) in a tiny nutshell

- Latent Dirichlet Allocation (LDA) is a hierarchical Bayesian model that explains the variation in a set of documents in terms of a set of K latent “topics,” i.e., distributions over the vocabulary
- Each document is assumed to be a mixture of these topics
- Words are drawn by:
 - Choosing a topic z | per-doc mixture weights
 - Sampling from that topic z

Example LDA topics

topic 0:

game	----	0.2027
games	----	0.1311
play	----	0.0525
ball	----	0.0361
score	----	0.0305
points	----	0.0256
rules	----	0.0224
first	----	0.0213
lead	----	0.0211
played	----	0.0188
goal	----	0.0186
card	----	0.0173
minutes	----	0.0163

Example LDA topics

topic 1:

born	---	0.0975
career	---	0.0441
died	---	0.0312
worked	---	0.0287
served	---	0.0273
director	---	0.0209
member	---	0.0176
years	---	0.0167
december	---	0.0164
joined	---	0.0162
college	---	0.0157
january	---	0.0147
university	---	0.0145

Example LDA topics

topic 2:

university	---	0.1471
college	---	0.0584
research	---	0.0412
professor	---	0.0347
science	---	0.0259
studies	---	0.0229
education	---	0.0226
degree	---	0.0210
department	---	0.0141
study	---	0.0136
academy	---	0.0125
sciences	---	0.0123

Example LDA topics

topic 3:

stage	----	0.2467
page	----	0.1115
stages	----	0.0631
murray	----	0.0603
mask	----	0.0528
shadow	----	0.0365
hearts	----	0.0320
finger	----	0.0295
suit	----	0.0280
min	----	0.0227
burn	----	0.0215
arrow	----	0.0206
bow	----	0.0201

Example LDA topics

topic 4:

fire	---	0.0462
attack	---	0.0392
killed	---	0.0391
battle	---	0.0363
gun	---	0.0194
shot	---	0.0185
fight	---	0.0179
shooting	---	0.0171
men	---	0.0165
enemy	---	0.0161
attacks	---	0.0152
fighting	---	0.0143
weapons	---	0.0143

Example LDA topics

topic 5:

due	---	0.0198
effects	---	0.0166
caused	---	0.0132
found	---	0.0125
cause	---	0.0125
reported	---	0.0125
study	---	0.0116
damage	---	0.0114
people	---	0.0113
result	---	0.0113
high	---	0.0113
associated	---	0.0108

Example LDA topics

topic 6:

california	---	0.1872
san	---	0.1705
los	---	0.1066
mexico	---	0.0865
francisco	---	0.0655
santa	---	0.0399
del	---	0.0394
mexican	---	0.0369
city	---	0.0339
las	---	0.0245
juan	---	0.0239
antonio	---	0.0194
orange	---	0.0188
american	---	0.0165

Online VB for LDA

(Hoffman et al., NIPS 2010)

- Until converged:
 - Choose a mini-batch of documents randomly
 - For each document in that mini-batch
 - Estimate approximate posterior over what topics each word in each document came from
 - (Partially) update approximate posterior over topic distributions based on what words are believed to have come from what topics

Online VB for LDA in VW

To learn a set of topics:

```
./vw wiki.dat --lda 10  
--lda_alpha 0.1 --lda_rho 0.1 --lda_D 75963  
--minibatch 256 --power_t 0.5 --initial_t 1  
-b 16  
--cache_file /tmp/vw.cache --passes 2  
-p predictions.dat  
--readable_model topics.dat
```

Online VB for LDA in VW

```
./vw wiki.dat: Analyze word counts in wiki.dat  
--lda 10: Use 10 topics
```

Online VB for LDA in VW

Hyperparameters:

--lda_alpha 0.1: $\theta_d \sim \text{Dirichlet}(\alpha)$

--lda_rho 0.1: $\beta_k \sim \text{Dirichlet}(\rho)$

of documents

--lda_D 75963: We'll analyze a total of 75963 unique documents

Online VB for LDA in VW

Learning parameters:

--minibatch 256: Analyze 256 docs at a time

--power_t 0.5, --initial_t 1: Stepsize schedule

$$\eta_t = (\text{initial_t} + t)^{-\text{power_t}}$$

Online VB for LDA in VW

-b 16: We expect to see at most 2^{16} unique words

Online VB for LDA in VW

To run multiple passes through the dataset:

--cache_file /tmp/vw.cache: Where to cache
parsed word counts

--passes 2: Number of times to go over the
dataset

Online VB for LDA in VW

-p predictions.dat: File to print out the inferred per-document topic weights to

--readable_model topics.dat: We print out the topics in human-readable format to topics.dat

Data Format

No labels, no namespace

| word_id:word_ct word_id:word_ct word_id:word_ct
word_id:word_ct ...

| word_id:word_ct word_id:word_ct word_id:word_ct
word_id:word_ct ...

| word_id:word_ct word_id:word_ct word_id:word_ct
word_id:word_ct ...

...

Output Predictions Format

Each line corresponds to a document d

Each column corresponds to a topic k

$$\begin{array}{ccccccc} Y_{1,1} & Y_{1,2} & \dots & Y_{1,k} & \dots & Y_{1,K} & 1 \\ Y_{2,1} & Y_{2,2} & \dots & Y_{2,k} & \dots & Y_{2,K} & 1 \\ \dots & & & & & & \\ Y_{d,1} & Y_{d,2} & \dots & Y_{d,k} & \dots & Y_{d,K} & 1 \\ \dots & & & & & & \end{array}$$

Output Topics Format

Each line corresponds to a topic k

Each column corresponds to a word w

$\lambda_{1,1}$ $\lambda_{1,2}$... $\lambda_{1,k}$... $\lambda_{1,K}$

$\lambda_{2,1}$ $\lambda_{2,2}$... $\lambda_{2,k}$... $\lambda_{2,K}$

...

$\lambda_{w,1}$ $\lambda_{w,2}$... $\lambda_{w,k}$... $\lambda_{w,K}$

...

$\lambda_{W,1}$ $\lambda_{W,2}$... $\lambda_{W,k}$... $\lambda_{W,K}$

Online VB for LDA in VW

To learn a set of topics:

```
./vw wiki.dat --lda 10  
--lda_alpha 0.1 --lda_rho 0.1 --lda_D 75963  
--minibatch 256 --power_t 0.5 --initial_t 1  
-b 16  
--cache_file /tmp/vw.cache --passes 2  
-p predictions.dat  
--readable_model topics.dat
```