hebin

博客园 首页 新随笔 联系 订阅 管理

随笔 - 4 文章 - 0 评论 - 14

Latent Dirichlet Allocation(LDA)

变量：

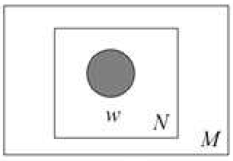
w表示词，z表示主题，$\mathbf{w}=(w_1,w_2,\cdots,w_N)$表示文档，语料库$D=(\mathbf{w}_1,\cdots,\mathbf{w}_M)$，$V$表示所有单词的个数（固定值），$N$表示一个文档中的词数（随机变量），$M$是语料库中的文档数（固定值），$k$是主题的个数（预先给定，固定值）。

在说明LDA模型之前，先介绍几个简单一些的模型。

1.Unigram model:

文档$\mathbf{w}=(w_1,w_2,\cdots,w_N)$，用$p(w_n)$表示词$w_n$的先验概率，生成文档$\mathbf{w}$的概率：$p(\mathbf{w})=\prod_{n=1}^{N}p(w_n)$。

图模型为：



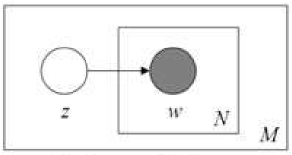
(a) unigram

2.Mixture of unigrams model:

一篇文档只由一个主题生成。该模型的生成过程是：给某个文档先选择一个主题z，再根据该主题生成文档，该文档中的所有词都来自一个主题。假设主题有$z_1,\ldots,z_k$,生成文档$\mathbf{w}$的概率为：

$$p(\mathbf{w})=p(z_1)\prod_{n=1}^{N}p(w_n|z_1)+\cdots+p(z_k)\prod_{n=1}^{N}p(w_n|z_k)=\sum_{z}p(z)\prod_{n=1}^{N}p(w_n|z).$$

图模型为：



(b) mixture of unigrams

LDA模型：

下面说明LDA模型生成一个文档的过程：

1首先要选择一个主题概率分布θ，$\theta=(\theta_1,\ldots,\theta_k)$，$\theta_i$代表第i个主题被选择的概率，即

$p(z=i|\theta)=\theta_i$，且$\sum_{i=1}^{k}\theta_i=1$，$\theta\sim Dir(\alpha)$，$p(\theta|\alpha)=\frac{\Gamma(\sum_{i=1}^{k}\alpha_i)}{\prod_{i=1}^{k}\Gamma(\alpha_i)}\theta_1^{\alpha_1-1}\cdots\theta_k^{\alpha_k-1}$.

公告

昵称：hebin
园龄：3年9个月
粉丝：9
关注：10
+加关注

<			2013年4月			>
日	一	二	三	四	五	六
31	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30	1	2	3	4
5	6	7	8	9	10	11

搜索

常用链接

我的随笔
我的评论
我的参与
最新评论
我的标签

随笔分类

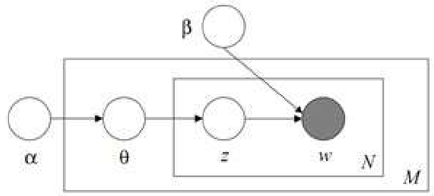
博士记事(1)
机器学习(2)
日常学习(1)

随笔档案

2015年10月 (1)
2014年1月 (1)
2013年4月 (2)

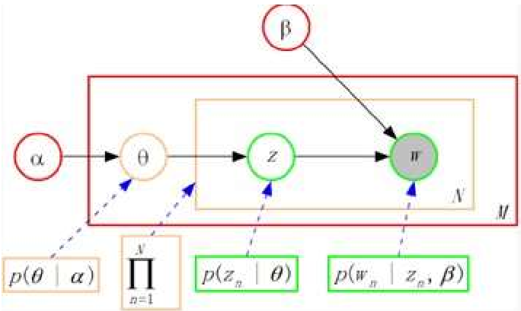
最新评论

1. Re:利用中文数据跑Googl...



由LDA的图模型我们可以清楚得看出变量间的依赖关系。

整个图的联合概率（单个文档）为：$p(\theta, \mathbf{z}, , \mathbf{w}|\alpha, \beta) = p(\theta|\alpha)\prod_{n=1}^{N} p(z_n|\theta)p(w_n|z_n, \beta)$，



生成文档的概率为$p(\mathbf{w}|\alpha, \beta) = \int p(\theta|\alpha)\prod_{n=1}^{N}\sum_{z_n} p(z_n|\theta)p(w_n|z_n, \beta)d\theta$，文本语料库由$M$篇文档组成，$D = (\mathbf{w}_1, \cdots, \mathbf{w}_M)$，故生成文本语料库的概率为

$p(D|\alpha, \beta) = \prod_{d=1}^{M}\int p(\theta_d|\alpha)\prod_{n=1}^{Nd}\sum_{z_{dn}} p(z_{d_n}|\theta_d)p(w_{d_n}|z_{d_n}, \beta)d\theta_d$.

下面叙述训练过程：

首先设定目标函数

$$\ell(\alpha, \beta) = \log p(D|\alpha, \beta) = \log\prod_{d=1}^{M} p(\mathbf{w}_d|\alpha, \beta) = \sum_{d=1}^{M}\log p(\mathbf{w}_d|\alpha, \beta).$$

我们参数训练的目标是求使$\ell(\alpha, \beta)$最大的参数α^*, β^*。我们把$p(\mathbf{w}|\alpha, \beta)$展开得 $p(\mathbf{w}|\alpha, \beta) = \frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)}\int(\prod_{i=1}^{k}\theta_i^{\alpha_i - 1})(\prod_{n=1}^{N}\sum_{i=1}^{k}\prod_{j=1}^{V}(\theta_i\beta_{ij})^{w_n^j})d\theta$，由于$\theta$和$\beta$的耦合，对$\ell(\alpha, \beta)$用极大似然估计难以计算。下面我们用变分EM算法来计算最优参数α, β。

E步骤：我们用$L(\gamma, \phi; \alpha, \beta)$来近似估计$\log p(\mathbf{w}|\alpha, \beta)$，给定一对参数值$(\alpha, \beta)$，针对每一文档，求得变分参数$\{\gamma_d^*, \phi_d^* : d \in D\}$，使得$L(\gamma, \phi; \alpha, \beta)$达到最大。

M步骤：求使$\mathscr{L} = \sum_d L(\gamma_d^*, \phi_d^*; \alpha, \beta)$达到最大的$\alpha, \beta$。

重复E、M步骤直到收敛，得到最优参数α^*, β^*。

E步骤的计算方法：

这里用的是变分推理方法(variational inference)，文档的似然函数

$$\begin{aligned}
&\log p(\mathbf{w} \mid \alpha, \beta) \\
&= \log\int\sum_{\mathbf{z}} p(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta)d\theta \\
&= \log\int\sum_{\mathbf{z}} q(\theta, \mathbf{z})\frac{p(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta)}{q(\theta, \mathbf{z})}d\theta \\
&= \log E_q\left[\frac{p(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta)}{q(\theta, \mathbf{z})}\right] \\
&\geq E_q\left[\log\frac{p(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta)}{q(\theta, \mathbf{z})}\right] \\
&= E_q[\log p(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta) - \log q(\theta, \mathbf{z})] \\
&= E_q[\log p(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta)] - E_q[\log q(\theta, \mathbf{z})].
\end{aligned}$$

嘻嘻

--Janvn

5. Re:利用中文数据跑Googl...

@smile_tina你试试另外几个小点的数据集行不行，我的完整版的数据删了...

--hebin

阅读排行榜

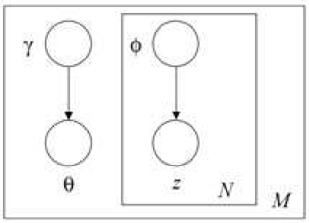
1. 利用中文数据跑Google开...
2. Latent Dirichlet Allocati...
3. 杨锦锋师兄博士毕业答辩(...
4. 博客里编公式(55)

评论排行榜

1. 利用中文数据跑Google开...

推荐排行榜

1. 利用中文数据跑Google开...
2. Latent Dirichlet Allocati...



γ为狄利克莱分布的参数，$\phi = (\phi_{ni})_{n\times i}, n = 1, \cdots, N, i = 1, \cdots, k$ ϕ_{ni}表示第n个词由主题i生成的概率，$\sum_{i=1}^{k}\phi_{ni} = 1$。

下面我们求使$L(\gamma, \phi; \alpha, \beta)$达到极大的参数$\gamma^*, \phi^*$。

将$L(\gamma, \phi; \alpha, \beta)$中的$p$和$q$分解，得

$$\begin{aligned} L(\gamma, \phi; \alpha, \beta) &= E_q[\log p(\theta \mid \alpha)] + E_q[\log p(\mathbf{z} \mid \theta)] + E_q[\log p(\mathbf{w} \mid \mathbf{z}, \beta)] \\ &\quad - E_q[\log q(\theta)] - E_q[\log q(\mathbf{z})]. \end{aligned}$$

把参数(α, β)和(γ, ϕ)代入$L(\gamma, \phi; \alpha, \beta)$，再利用公式$E_q[\log(\theta_i)|\gamma] = \Psi(\gamma_i) - \Psi(\sum_{j=1}^{k}\gamma_j)$（$\Psi$是$\log\Gamma$的一阶导数，可通过泰勒近似来计算），我们可得到

$$\begin{aligned} L(\gamma, \phi; \alpha, \beta) &= \log\Gamma(\sum_{j=1}^{k}\alpha_j) - \sum_{i=1}^{k}\log\Gamma(\alpha_i) + \sum_{i=1}^{k}(\alpha_i - 1)(\Psi(\gamma_i) - \Psi(\sum_{j=1}^{k}\gamma_j)) \\ &\quad + \sum_{n=1}^{N}\sum_{i=1}^{k}\phi_{ni}(\Psi(\gamma_i) - \Psi(\sum_{j=1}^{k}\gamma_j)) \\ &\quad + \sum_{n=1}^{N}\sum_{i=1}^{k}\sum_{j=1}^{V}\phi_{ni}w_n^j\log\beta_{ij} \\ &\quad - \log\Gamma(\sum_{j=1}^{k}\gamma_j) + \sum_{i=1}^{k}\log\Gamma(\gamma_i) - \sum_{i=1}^{k}(\gamma_i - 1)(\Psi(\gamma_i) - \Psi(\sum_{j=1}^{k}\gamma_j)) \\ &\quad - \sum_{n=1}^{N}\sum_{i=1}^{k}\phi_{ni}\log\phi_{ni}. \end{aligned}$$

然后用拉格朗日乘子法（即变量的拉格朗日函数对变量求偏导等于零，求出变量对应的等式）来计算可得

$$\phi_{ni} \propto \beta_{iv}\exp(\Psi(\gamma_i) - \Psi(\sum_{j=1}^{k}\gamma_j)),$$

$$\gamma_i = \alpha_i + \sum_{n=1}^{N}\phi_{ni}.$$

由$\sum_{i=1}^{k}\phi_{ni} = 1$归一化求得$\phi_{ni}$。由于解$\phi_{ni}$和$\gamma_i$相互影响，可用迭代法来求解，算法如下：

```
initialize φ_ni^0 := 1/k for all i and n
initialize γ_i := α_i + N/k for all i
repeat
    for n = 1 to N
        for i = 1 to k
            φ_ni^{t+1} := β_{iw_n} exp(Ψ(γ_i^t))
        normalize φ_n^{t+1} to sum to 1.
    γ^{t+1} := α + Σ_{n=1}^N φ_n^{t+1}
until convergence
```

最终可以得到收敛的参数γ^*, ϕ^*。这里的参数γ^*, ϕ^*是在给定一个固定的文档$\mathbf{w}$下产生的，因此γ^*, ϕ^*也可记为$\gamma^*(\mathbf{w}), \phi^*(\mathbf{w})$，变分分布$q(\theta, \mathbf{z}|\gamma^*(\mathbf{w}), \phi^*(\mathbf{w}))$是后验分布$p(\theta, \mathbf{z}|\mathbf{w}, \alpha, \beta)$的近似。文本语料库$D = (\mathbf{w}_1, \cdots, \mathbf{w}_M)$，用上述方法求得变分参数$\{\gamma_d^*, \phi_d^* : d \in D\}$。

M步骤的计算方法：

将$\{\gamma_d^*, \phi_d^* : d \in D\}$代入$\sum_d L(\gamma_d, \phi_d; \alpha, \beta)$得$\mathscr{L} = \sum_d L(\gamma_d^*, \phi_d^*; \alpha, \beta)$，我们用拉格朗日乘子法求$\beta$，拉格朗日函数为$l = \mathscr{L} + \sum_{i=1}^{k}\lambda_i(\sum_{j=1}^{V}\beta_{ij} - 1)$，求得$\beta_{ij} \propto \sum_{d=1}^{M}\sum_{n=1}^{N_d}\phi_{dni}w_{dn}^j$，由$\sum_{j=1}^{V}\beta_{ij} = 1$归一化求得$\beta_{ij}$。

下面求α 我们对拉格朗日函数l对α_i求偏导 得

分类：机器学习

« 上一篇：博客里编公式
» 下一篇：利用中文数据跑Google开源项目word2vec

posted @ 2013-04-25 22:12 hebin 阅读(1780) 评论(0) 编辑 收藏