股票数据分析汇总报告

朱恬骅

September 9, 2012

1 Topic Model revisited

In tranditional topic models, we assume that each observation is a consequence of several random variables, called 'topics'. Each 'topic' represents a distribution of some results, denoted as r_i. The topics are thus hidden variables, as we have been familiar with in HMMs. For the set of topics $t_1, t_2, ..., t_m$ and the set of possible results $r_1, r_2, ..., r_n$, we have a matrix $P = (p_{ij})_{m \times n}$, where $p_{ij} = P(r_j|t_i)$. We call this matrix P as topic-result matrix.

2 Towards observations from two perspectives

Tranditional topic model may also assume that, two observations, or two sets of feature properties, are results of independent topics. Further more, any two topics are irrelevant. However, in the reality, such constraints seldom hold. Two topics may cover a mutual set of conditions, and two obseravtions, if they are correspondent to one same object, may interfere each other, or display some mutual features. In such situations, we cannot adopt independent identical distribution assumption any longer.

Instead, we introduce a matrix, called topic co-occurence matrix. This matrix, demoted as A in the following, describes the posibility of two topics occuring in the same context. That is, the relevance between some two observation results, respectively denoted as Xs and Ys, from two different perspectives, are assumed to attribute to the relevance between their own generative topics.

Given some observed vectors $X_1, X_2, ..., X_n \in \mathbb{R}^p$ and $Y_1, Y_2, ..., Y_n \in \mathbb{R}^q$, each X_i and Y_i are features for one object, we want to reduce the number of representation dimensions, by using its topic information. For the reason that topic variables are hidden, thus inaccessible to the observers, we can only estimate the matrix A by factorizing the expectation matrix of the product of the two observation vectors, i.e. $E = \mathbf{E}XY^\top$. In our model, we have two sets of topics, say T_X and T_Y, and note $s = |T_X|$, $t = |T_Y|$, generating results according to the topic-result possiblity matrices P_X and P_Y (and sometimes, generating several results, and the observation is the sum). Then we have the following equation and constraints holding:

$$E = R_X P A Q^\top R_Y^\top \tag{1}$$
$$\sum_{i,j} A_{ij} = 1 \tag{2}$$
$$\sum_j P_{ij} = 1, \forall i \tag{3}$$
$$\sum_j Q_{ij} = 1, \forall i \tag{4}$$

where $P_{p\times s} = P_X, Q_{q\times t} = P_Y, A \in \mathbb{R}^{s\times t}$, and p, q are the length of observation vectors from X and Y, repectively. R_X and R_Y are matricies composed of all result column vectors.

We may use normal topic models to cope with $R_X P$ and $R_Y Q$, and then solving the first equation. In the siduations where only two results, like $0, r(r > 0)$, are possible for one component of the result, we can compute r's automatically by simply factorizing E into three matricies, like $E = P' A Q'^\top$, and compute r's by summing up the rows of P' and Q'.

3 What we want to do

Note: All the following sections are obsolete and requires rewriting.

3.1 Ideas

- Build three matrices, P, A and Q, so that $\mathbf{E}XY^\top \approx PAQ^\top$.
- $P \in \mathbb{R}^{p\times s}, Q \in \mathbb{R}^{q\times t}, A \in \mathbb{R}^{s\times t}$

- So that $\bar{X} = P^\top X = \arg\min_{x \in \mathfrak{X}} [\mathfrak{L}_1(Px, X) + R_1(x)] \in \mathbb{R}^s$.
  $\mathfrak{L}_1$ is a loss function, and R_1 is a regularization parameter to encourage sparsity. $\mathfrak{X}$ is a subset of $\mathbb{R}^s$
  Similar conditions hold for $\bar{Y} = Q^\top Y$.

3.2 Relation to two-view topic model

- Consider two hidden random vectors $W \in \mathbb{R}^s$ and $Z \in \mathbb{R}^t$ that are unobserverd.
  Due to symmetry, discussions about Y is omitted here.

- W is a posibility vector which means that X is generated from a mixture W over s topics. Denote the probability of topic l as $p_l(X)$, then we have

- $p(X|W) = \sum_{l=1}^{s} W_l p_l(X)$

- $p(X|W)$ is independent of Y and Z, but W and Z are correlated.

3.3 Explaining P, A, and Q

So, we have:

$$\mathbf{E}_{X,Y} XY^\top = \sum_{l,l'} \mathbf{E} W_l Z_{l'} \int X p_l(X) \int Y^\top q_{l'}(Y) \tag{5}$$

$P \in \mathbb{R}^{p \times s}$ satifies that $P[l,:] = \int X p_l(X)$
$Q \in \mathbb{R}^{q \times t}$, $Q[l,:] = \int Y q_{l'}(Y)$.
$A \in \mathbb{R}^{s \times t}$, $A_{l,l'} = \mathbf{E}_{W,Z} W_l Z_{l'}$

4 Constraints on these three matrices

$A_{i,j} \geq 0$ and $\sum_{i,j} A_{i,j} = 1$, measuring how topics are correlated.
$P, Q \geq 0. ||P[l,:]||_1 = 1, ||Q[l',:]||_1 = 1.$
Consider the following optimization:
$L_1 = \lambda_1 \mathbf{E}_X \inf_x [\mathfrak{L}_1(Px, X) + R_1(x)]$
$L_2 = \lambda_2 \mathbf{E}_Y \inf_y [\mathfrak{L}_2(Qy, Y) + R_2(y)]$
$L_0 = \mathfrak{L}_0(PAQ^\top, \mathbf{E}_{X,Y} XT^\top) + R_0(P, A, Q)$
$min_{P,A,Q}[L_1 + L_2 + L_0]$

4.1 Algorithm

$$\mathbf{E}_{X_j,Y_k} X_j Y_k = P_j^\top A Q_k, 1 \le j \le p, 1 \le k \le q, A, P_j, Q_k \ge 0 \tag{6}$$

$$[P, A, Q] = \arg \min_{P,A,Q} \sum_{j,k} (P_j^\top A Q_k - \mathbf{E} X_j Y_k)^2 \tag{7}$$

and $||P||_0 \le u, ||Q||_0 \le v$

Alternating least squares: Fix P, A, and solve for Q; then fix P, Q for A, and A, Q for P.

Consider diagonal A at first.

Least square Denote $\Delta = PAQ^\top - \mathbf{E}XY^\top$, so we have $L = \sum_{j,k}(P_j^\top A Q_k - \mathbf{E}X_j Y_k)^2 = \mathbf{trace}(\Delta\Delta^\top)$

$\frac{\partial L}{\partial \Delta} = 2\Delta$, $\frac{\partial \Delta}{\partial P} = -AQ^\top$.

Similar conclusions apply to Q.

We have two ways to optimize L. One is to use Newton's method to solve the equation, solve alternatively for the three variables:

$$L(P, A, Q) = 0 \tag{8}$$

Another method is to solve equations $\frac{\partial L}{\partial *} = 0$, where $*$ stands for P, Q and A.

Lagrange multiplier Rewrite our target function F as:

$$F = L + R + \mu(\sum_{i,j} A_{ij} - 1) + \sum_l \nu_l(||P[l,:]||_1 - 1) + \sum_{l'} \kappa_{l'}(||Q[l',:]||_1 - 1) \tag{9}$$

So, we have:

$$\frac{\partial F}{\partial P} = \frac{\partial (L+R)}{\partial P} + \sum_l \nu_l sgn(P[l,:]) \mathbf{1}_l^{p\times s} = 0 \tag{10}$$

$$\frac{\partial F}{\partial Q} = \frac{\partial (L+R)}{\partial Q} + \sum_{l'} \kappa_{l'} sgn(Q[l',:]) \mathbf{1}_{l'}^{q\times t} = 0 \tag{11}$$

$$\frac{\partial F}{\partial A} = \frac{\partial (L+R)}{\partial A} + \mu \mathbf{1}_{t\times s} = 0 \tag{12}$$

with all these constraints are simutaneous.

Newton's Method Take P for example.

$$\mathbf{d}L = \sum_{j,k} 2(P_j^\top AQ_k - \mathbf{E}X_jY_k)\mathbf{d}(P_j^\top AQ_k - \mathbf{E}X_jY_k) \tag{13}$$

$= \sum_{j,k} 2\Delta_{jk}\mathbf{d}(P_j^\top AQ_k)$
$= \sum_{j,k} 2\Delta_{jk}\mathbf{d}P_j^\top AQ_k = L$
$\Rightarrow 2\Delta_{jk}\mathbf{d}P_j^\top AQ_k = -\Delta_{jk}^2$
$\Rightarrow \mathbf{d}P_j^\top AQ_k = -1/2\Delta_{jk}$
$\Rightarrow \mathbf{d}P_j^\top = -pinv(AQ_k)\Delta_{jk}/2$
Note: here Q_k is the transpose of the k-th row in Q.

How to meet with the constraints? Constraints.

$$A_{i,j} \geq 0, \sum_{i,j} A_{i,j} = 1 \tag{14}$$

$$P, Q \geq 0 \tag{15}$$

$$||P[l,:]||_1 = 1 \tag{16}$$

$$||Q[l',:]||_1 = 1 \tag{17}$$

The last two constraints are relatively easy to hold. We calculate 1-norm for each line of P or Q, and then divide this line by its 1-norm value, in each iteration. Such criteria will influence the accuracy (i.e., the value of L).

Initialize P and Q with $\frac{1}{pq}\mathbf{1}_{p\times s}$ and $\frac{1}{qt}\mathbf{1}_{q\times t}$, respectively, will hold the second constraint in our experiment. **Note:** $P \in [0,1]^p$, and $Q \in [0,1]^q$.

$$A_{i,j} \geq 0, \sum_{i,j} A_{i,j} = 1 \tag{18}$$

The first part, in our experiment, also holds. But the second part, for the reason that we always use A to hold the last two constraints, will not hold.

This requires us to preprocess our inputs or our model to satisfy this criterion. Here we denote λ^2 for $\sum_{i,j} A_{i,j}$, so our model should be:

$$\mathbf{E}_{X,Y}XY^\top = \lambda^2 PAQ^\top. \tag{19}$$

And when projection is applied, we use $\lambda P^\top X$ for $\bar{X}$, and $\lambda Q^\top X$ for $\bar{Y}$.