課題4の補足

担当教授：植野真臣, TA: 木下 涼

e-mail: kinoshita@ai.is.uec.ac.jp

0.1　全体の流れ

課題 4 では，TAN を用いて二つのデータセット spam と sentiment の分類精度 (正答率) を測定する．課題 2 で用いた Naive Bayes は学習データに依らずグラフが確定していたが，課題 4 では TAN のグラフを学習データから推定しなければならない．具体的には，テキストの図 2.2 に示される $X_1 \leftarrow X_2 \rightarrow X_3$ のように，説明変数間で構成する木構造をデータから推定する．木構造の推定方法として，テキスト 31〜32 ページに記載されているように，各説明変数間の条件付き相互情報量を重みとした最大全域木を求める．最大全域木を求めるアルゴリズムとして，プリム法やクラスカル法がある．なお，TAN.java では変数 X_0(LD.csv, TD.scv で一番左側の列) を木構造のルートノードと定めていることに注意する．条件付き相互情報量の計算は関数 getConditionalMutualInformation に実装し，それを用いた説明変数間の木構造推定は関数 getMaximumSpanningTree に実装する．TAN.java では木構造の格納先を int 型配列 str_tan としており，求めた木構造において X_i の目的変数以外の親が X_j の時，str_tan$[i] = j$ として木構造を表している．

TAN のグラフを推定したら，そのグラフにしたがって各変数のパラメータを求め，さらにその後分類精度を測定する．木構造のルートノード以外の説明変数は目的変数以外にも親変数を持つため，それに対応するように関数 setFrequencyTable, getParameters, classification を実装する必要がある．

0.2　TAN が int 型の一次元配列で保持できる理由

TAN のグラフの情報を保持するには，各変数の親変数をそれぞれ保持すれば良い．例えばテキストの図 2.2 の TAN の構造情報を保持するには，
X_1 の親変数は X_2 と目的変数，
X_2 の親変数は目的変数，
X_3 の親変数は X_2 と目的変数
という情報を保持すれば良い．実際には，TAN は全ての説明変数が目的変数を親にもつことがわかっているため，この情報は保持しなくても良く，
X_1 の目的変数以外の親変数は X_2，
X_3 の目的変数以外の親変数は X_2
という情報だけを保持すれば良い．結果として，図 2.2 の TAN の構造は以下のように int 型の一次元配列で保持できる．
$int[]str\_tan = new\ int[3]$; (説明変数が 3 つの TAN の構造情報の格納先)
$str\_tan[0] = 1$; (X_1 の目的変数以外の親変数は X_2)
$str\_tan[1] = -1$; (X_2 の目的変数以外の親変数は存在しない．存在しないことを -1 で表現．)
$str\_tan[2] = 1$; (X_3 の目的変数以外の親変数は X_2)

0.3 N_{ijk} について

TAN のパラメータの最尤推定量 $\widehat{\theta}_{ijk} = \dfrac{N_{ijk}}{N_{ij}}$ で使う N_{ijk} が非常にわかりにくいので，求め方を以下で紹介する．N_{ijk} とは，$X_i = k$ かつその親変数集合が j 番目の状態をとる頻度である．親変数集合の状態とは，X_i の全ての親変数がとる値の組合せのことである．具体的には，X_i の親が X_p と X_0 であり，それぞれの変数が 2 値をとる時，X_p, X_0 の状態とは $(X_p, X_0) = (0,0), (0,1), (1,0), (1,1)$ の 4 つである．例えばテキストの図 2.2 の TAN において，

N_{130} =「$X_1 = 0$ かつその親変数集合の状態が $(1,0)$ となる頻度」=「$X_1 = 0$ かつ $X_2 = 1$ かつ $X_0 = 0$ となる頻度」，

N_{141} =「$X_1 = 1$ かつその親変数集合の状態が $(1,1)$ となる頻度」=「$X_1 = 1$ かつ $X_2 = 1$ かつ $X_0 = 1$ となる頻度」，

N_{311} =「$X_3 = 1$ かつその親変数集合の状態が $(0,1)$ となる頻度」=「$X_3 = 1$ かつ $X_2 = 0$ かつ $X_0 = 1$ となる頻度」

である．一方，X_2 の親は X_0 のみであるから，親状態は 0 か 1 の二つである．したがって，

N_{200} =「$X_2 = 0$ かつその親変数集合の状態が 0 となる頻度」=「$X_2 = 0$ かつ $X_0 = 0$ となる頻度」，

N_{201} =「$X_2 = 1$ かつその親変数集合の状態が 0 となる頻度」=「$X_2 = 1$ かつ $X_0 = 0$ となる頻度」，

N_{210} =「$X_2 = 0$ かつその親変数集合の状態が 1 となる頻度」=「$X_2 = 0$ かつ $X_0 = 1$ となる頻度」，

N_{211} =「$X_2 = 1$ かつその親変数集合の状態が 1 となる頻度」=「$X_2 = 1$ かつ $X_0 = 1$ となる頻度」

である．

0.4 分類精度のめやす

正しく実装すると，データセット sentiment の分類精度は約 0.6 から 0.7 になる．