課題6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 6 の目標は，TAN を用いて欠損値を含んだデータから分類を行うことである．まずはデータセット spam2 の訓練データ LD.csv から，TAN のグラフとパラメータを学習する．ここで LD.csv は欠損値を含んでいないので，課題 24 と全く同じ方法で構造やパラメータを学習できる．学習後は関数 classification で分類を行うが，分類を行うテストデータ TD.csv が欠損値を含んでいる．今，TD.csv 中のある行において欠損値をもつ変数の集合を S とすると，TAN の識別関数は

$$\widehat{x}_0 = \underset{x_0}{\operatorname{argmax}}\, \theta_{x_0} \sum_{S\ \text{の値}} \prod_{i}^{n} \theta_{i\pi_i x_i}$$

である．例えば，テキストの図 2.2 の TAN を用いて，$X_2 = x_2$ が与えられて X_1 と X_3 が欠損したデータの分類を行う場合，識別関数は以下のようになる．

$$\widehat{x}_0 = \underset{x_0}{\operatorname{argmax}}\, \theta_{x_0} \sum_{x_1, x_3} \prod_{i}^{n} \theta_{i\pi_i x_i}$$

$\sum_{x_1,x_3}$ は欠損した変数 X_1, X_3 のとりうる値のパターンに関する総和を示しており，そのパターン数は $(x_1, x_3) = (0,0), (0,1), (1,0), (1,1)$ の $2 \times 2 = 4$ 通りある．この例からわかるように，欠損データが増加すると，S の値 のパターン数が指数的に増加してしまう．したがって，欠損データが多くなると直接 $\sum_{S\ \text{の値}}$ の箇所を計算するのは計算時間的に困難なので，効率的に計算可能なテキスト 4 章のアルゴリズム 1 を用いる．