情報意味論 (課題2)

慶應義塾大学理工学部
櫻井 彰人

レポート課題2

- レポート提出は、前回同様、web を通じて行ってください。
- 2.1a と 2.1b はどちらかを選択して下さい。2.1a は、R のプログラムへの慣れが必要かもしれません。
- 2.4はRのプログラムを動かし、修正する必要がありますが、修正量はわずかで、容易に想像できます。
- レポートは電子的に作成してください。TeX, MsWord で作成して結構です。提出は pdf 形式でも結構です。
- 書くべき内容に関しては特には述べません。すべて常識的に判断してください。
- 締め切りは、4週間+α 後の木曜日 i.e. 1/**14** 一杯とします。

(2015.12.5)

2.1a トピックモデル

計算は密行列で行って結構です。

APデータに tf-idf を導入したら、結果が良くなるか調べてみよう

- tfidf の定義については、例えば、wikipedia の記述を参照してください。
- 文書単語行列を用いる場合、そのまま用いるよりは、tfidf を利用した方が、例えば文書間の類似度を測る場合、よりよい結果が得られることがある。
- tfidf 行列に対し LSA と pLSA を適用して下さい。どのような結果になりますか。
- なお、APデータでは、tf はすでに考慮されている（単語出現頻度を考えている）ので、idf のみを計算し、tfidf行列を作って下さい。ヒント: tfidf <- t(t(tf)*idf)

密行列でも疎行列でも

- また逆に、tf を用いない、つまり、文書に単語があらわれれば 1、そうでなければ0 としたら、LSA/pLSA/LDA の結果はどうなりますか（どう劣化するか、または、劣化しないかを調べて下さい（実際例では、劣化するどころかよくなる場合があります））。

AP$v <- 1は目的に合いません

APが疎行列 simple_triplet_matrix であれば、AP$v[]<- 1 とすればよい。

- これに、idf を導入して下さい。LSA/pLSA の結果はどうなりますか？
- なお結果の評価は、数値的な評価ではなく、トピックがまとまっているかどうかの、極めて主観的な評価で結構です。

2-1b 混合推論

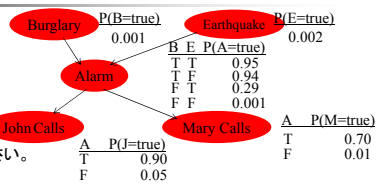
原因間推論と診断推論を同時に

例: “John calls” かつ “Earthquake=false”:

$P(A \mid J, \neg E) = 0.03$

$P(B \mid J, \neg E) = 0.017$

$P(B \mid J, \neg E)$ を求めてください。



ヒント

$$P(J \mid B, \neg E) = ??$$
$$P(J \mid \neg B, \neg E) = ??$$
$$P(B, J, \neg E) = ??$$
$$P(\neg B, J, \neg E) = ??$$
$$P(B \mid J, \neg E) = ??$$

2.2 条件付独立性

(1) 削除

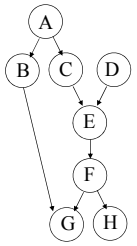
(2) x → y → z は x ← y → z と等価であるのにも関わらず、必ずしも x → y ← z と等価にはならないことを示しなさい。前者は証明（式変形のみ）を、後者は、実際に x, y, z を構成することで示してください。

(3) Explaining away effect が観測される例を構成し、説明して下さい。用いるのは3変数だけで結構です。すなわち、例えば、Earthquake, Burglary, Alarm です。

2.2 条件付独立性（続）

(4) 右の関係があるとき、条件付独立性に関する下記の問に答えなさい

1. (B ⫫ C | A)？
2. (A ⫫ F | E)？
3. (C ⫫ D | F)？
4. (A ⫫ G | B,F)？



2-3 EM

Dempster, Laird and Rubin の論文 "Maximum Likelihood from Incomplete Data via the EM Algorithm" の p.2 に書かれている例題を、論文に書かれているようにEMアルゴリズムを用いて、解いて下さい（http://web.mit.edu/6.435/www/Dempster77.pdf）。

プログラムを作成し、実際に動作させ、論文と同様の結果が得られるか、確認してください。なお、本講義のスライドに少し簡単にした問題の答えが書いてあります。それを参考にして結構です。

2-4 モデル選択

Wikipedia にあるプログラム (http://en.wikipedia.org/wiki/File%3aEm_old_faithful.gif) を少し手直しして混合正規分布の可視化をしてみる。

- プログラムを修正して、3個の正規分布、4個の正規分布の混合分布でfitしてください。よりよいfitになりますか？
  - プログラムを示してください。結果の図を回答して下さい。
- 要素分布の最適な（汎化能力が高いという意味で）個数はどのように決めたらよいでしょうか？（実際に計算しなくて結構です。考え方を述べれば結構です。）

http://en.wikipedia.org/wiki/File%3aEm_old_faithful.gif のうち、「#plot initial contours」の前までは、同様。これ以降を次のものに変える。
HTMLファイル及び画像ファイルは、作業ディレクトリの下に作られる

```
#load library animation
if(!require(animation)){
  install.packages("animation")
}
library(animation)
ani.options(loop=FALSE, interval=0.25)      ##seconds per frame

saveHTML({
##plot initial contours
iter <- 1
plot.em(theta)

#run EM and plot
for (iter in 2:30){
 T <- E.step(theta)
 theta <- M.step(T)
 plot.em(theta)
}
})
```