CSC412 – Assignment #4 Solution

April 7, 2002

- Everybody did well on this assignment.
- There are only two serious mistakes that people made. Some people did not calculate the log likelihoods correctly (please refer to the lecture notes to make sure you know how to calculate this for various graphical models), and some people were not careful with probabilities being zero when calculate the mutual information weights.
- Everybody got the maximum likelihood tree correctly.
- Full marks is 75, and the marks are distributed as follows:
  - 5 marks – Output a representation of the optimal undirected tree.
  - 5 marks – Draw the histogram of log likelihoods.
  - 5 marks – Calculate the min, max, mean and median of the log likelihoods.
  - 5 marks – Figure out which training case has worst log likelihood.
  - 5 marks – Print out keywords in the worst training case, and calculate the percentage of postings mentioning "god" and "ftp".
  - 5 marks – Output a representation of the optimal directed tree.
  - 10 marks – Calculate the counts and marginal and joint probabilities.
  - 15 marks – Calculate mutual information weights.
  - 5 marks – Call `mwst.m`.
  - 15 marks – Calculate log likelihoods.
- Here is a sample solution:

```
load a4newsgroups

% general parameters
[numvar numdoc] = size(documents);
% values(i) is the i'th value of the variable.
values = [0 1];
numval = length(values);

% minprob is the minimum non-zero probability of the joint probabilities
% calculated from the document.
minprob = 1/numdoc;

% collect counts and calculate joint probabilities
% pij(x1,x2,v1,v2) = prob(x1=values(v1),x2=values(v2) in documents)
pij = zeros(numvar,numvar,numval,numval);
for v1=1:numval,
  for v2=1:numval,
    pij(:,:,v1,v2) = full((documents==values(v1))*(documents==values(v2))');
  end;
```

```
end;
pij = pij/numdoc;

% calculate marginal probabilities
pi = reshape(pij,numvar^2,numval^2);
pi = pi(find(eye(numvar)),find(eye(numval)));

% calculate entropies and weights
% Take care of 0 probabilities by max(prob,minprob).
% hi and hij are entropies of marginal and joint probabilities respectively,
% and are used to calculate the weights.
hi  = sum(pi.*log(max(pi,minprob)),2);
hi  = hi(:,ones(1,numvar));
hij = sum(sum(pij.*log(max(pij,minprob)),3),4);
wij = hij-hi-hi';

% calculate maximum weight spanning tree
tree = mwst(-wij);

% calculate log-likelihoods of documents
lik = zeros(1,numdoc);
logpij = log(pij);
logpi  = log(pi);
for vv=1:numvar,
  if tree(vv)~=0,
    index = 0;
    doc = full(documents([vv tree(vv)],:));
    for v1=1:numval,
      index = index + v1*(doc==values(v1));
    end;
    lik = lik + reshape(logpij(vv,tree(vv),index(1,:)+numval*(index(2,:)-1)),...
                        1,numdoc) - logpi(tree(vv),index(2,:));
  else
    index = 0;
    doc = full(documents(vv,:));
    for v1=1:numval,
      index = index + v1*(doc==values(v1));
    end;
    lik = lik + logpi(vv,index);
  end;
end;
```