Lecture 13b: Latent Semantic Analysis

CS540 4/25/19

Material borrowed (with permission) from Vasileios Hatzivassiloglou & Evimaria Terzi. Mistakes are mine.

Announcements

NLP Project
- Team in class presentation: Tuesday May 7th
- One week from Tuesday!
- Paper: Friday, May 10th
- Two weeks from tomorrow!

Remember the Mid-term?
- It’s graded (finally)
- Grades have been emailed to you
- No curve: 90-100 is an A, 80-90 a B, etc.
- High score: 100
  - >= 90: 5
  - >= 80: 2
  - >= 70: 2
  - < 70: 1

Midterm Question #1

Consider the following SAT expression:

\[(A \lor B \lor C) \land (\neg A \lor \neg B \lor \neg C) \land (A \lor D) \land (B \lor E) \land (C \lor F) \land (A \lor F)\]

Assume you are using A* search to find a solution, and that your heuristic is the number of subexpressions that are satisfied.

a) What is the first variable that will have a truth value assigned to it, what will that value be, and why? (If two or more options are valid, list them all and explain why)

\{\} \rightarrow \{A=T\}  This is unique (satisfies the most subterms)

#1 (continued)

b) What is the 2nd variable that will have a truth value assigned to it, what will it be, and why? (If two or more options are valid, list them all and explain why)

\{A\} \rightarrow \{A=T, B=T\} or you could add C=T, E=T or F=T. They all satisfy the same number of terms.

Question #2

Now consider using local search to solve the same SAT expression. Assume that the neighborhood function allows you to change the truth value of any one term, and the evaluation function is the number of satisfied subexpressions. Assume, too, that there is no termination predicate.

a) If the initial state is that all variables are assigned to true, how many steps will it take to find a global solution? Give an example of an optimum solution path.

The initial state satisfies 5 of 6 terms. You can find the optimal in one step by negating A, B or C. So an example is \{(A,B,C,D,E,F) \rightarrow \{A,B,C,D,E,F\}\}

#2 (continued)

b) If the initial state is that all variables are assigned to false, how many steps will it take to find a global solution? Give an example of an optimum solution path.

The initial state satisfies only one term. The shortest path is three steps. You can turn A, B and C positive, or F and any two of A, B and C.
#2 (cont.)

c) Give an example of a state that is on a plateau, and explain why it is on a plateau. Would a random plateau walk help in this case?

A plateau state here are all optimal solutions (6 terms satisfied), since there is no termination predicate. The search algorithm will wander around this plateau. A random walk won’t help, because you can’t improve on this solution.

Another plateau state is {A=T, B=T, C=T, D = F, E = F, F = F}. Flipping B, C, D, E or F all results in the same evaluation as the start state (5 terms). Flipping A is downhill. So this is a plateau (if you hit it). Plateau search does help here, because if you negate B and them make E positive, for example, you get to the solution (6 terms satisfied).

#3 (continued)

C) What was the neighborhood relation?

- During Expand: all vertices connected to every vertex in the current state
- During Plateau Search: all vertices connected to every current vertex but one

D) What were the memory states?

Information stored about previous states. Most important: vertex weights. Also, the C' set during plateau search

E) What was the initialization function?

Select a node a random.

Question #3

We read a paper on using dynamic local search to solve the maximum clique problem. Answer the following questions about that paper

A) what was the search space?

The space of subsets of vertices

B) What was the solution set?

All subsets of vertices of size greater than threshold

#3 (cont.)

F) What was the evaluation function?

The size of the current clique. (A lot of you overthought this one)

G) What was the termination predicate?

Maximum clique size threshold (also, max steps is OK)

#3 (cont.)

Question #4

Encode the maximal clique problem as a Genetic Algorithm.

In particular:

A) How are the elements of the population represented?

Many possible solutions. The most common was the bit string representation (1 if vertex in set, 0 otherwise). Explicit list of vertices also good (better for subsequent questions)

B) What cross-over operator do you use?

Depends on your answer to A, but you have to be concerned with generating non-valid population elements (i.e. non-cliques). If you don’t do something to restrict this, I didn’t give you credit.

C) What mutation operator do you use?

Again, depends on part ‘a’. I did not penalize for generating non valid solutions twice.

D) Describe what a hyper-plane represents, given you answer to part ‘a’

 Depends on ‘a’, but pretty much any answer to ‘a’ implies that a hyperplane represents a (generally non-maximal) clique.
Question #5

Assume you want to implement a Kalman filter to track 2D points in an image that implements a constant acceleration model instead of a constant velocity model. (Assume that the observations are still \((x, y)\) positions at time \(t\).

A) How many terms are in the state vector, and what do they represent?

6. They represent position \((x, y)\), velocity \((x', y')\) and acceleration \((x'' \text{ and } y'')\)

B) What is the \(F\) Matrix? (note: I am not asking for a definition. I am asking for a matrix with actual values in it.) Explain why.

\[
\begin{bmatrix}
1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1
\end{bmatrix}
\]

The role of \(F\) is to project the model one step forward. If acceleration is constant, then velocity becomes velocity plus acceleration, while position is still position plus velocity.

Some of you did integrals and put 1/2s in the \([3,5]\) and \([4,6]\) positions. Good.

C) What is the \(H\) matrix? (note: same as above) Explain why.

\[
\begin{bmatrix}
1 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 1
\end{bmatrix}
\]

Because its role is to keep the observable part of the state, i.e. \((x, y)\), and remove everything else.

Homonyms

Crane
- A bird with long legs and a long neck
- A tool used to lift large objects
- A motion that stretches (e.g. crane your neck)

Date
- A fruit of the palm tree
- A romantic evening
- A day of the month

Foil
- A material used to wrap things (sometimes made of aluminium)
- To ruin a plan or scheme
- A dramatic foil

Resolving Homonyms

Part of speech tagging
- Distinguishes between meanings with different grammatical roles
- E.g. crane/bird (noun) vs crane/motion (verb)
- But all the homonyms on the previous slide have 2 noun meanings

Direct modifiers (adjectives/adverbs/phrases)
- “12 ton crane” – probably not a bird
- Anyone seen the movie Rampage?
- Not always available
- Require parsing and de-referencing
- To know which word is being modified

Meaning as Association

Context
- Different meanings (word senses) occur in different contexts
- Crane/bird often occurs with fish, marsh, water, rare, ...
- Crane/tool often occurs with construction, building, lift, collapse, ...
- Use surrounding text to select word meanings

Bag of Words
- Ignore syntax altogether
- Treat text as an unordered set of words
- Two texts are similar if their words are similar
- Problem: two texts may use different terms for the same thing
- We’ve gone from homonyms to synonyms
Association as information

Given a random variable \( X \), entropy is

\[
H(X) = \sum_i p(X_i) \log p(X_i)
\]

Mutual information is the reduction in entropy from knowing another variable

\[
I(X,Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)
\]

Calculating mutual information

\[
I(X,Y) = \sum_i \sum_j p(X,Y) \log \frac{p(X,Y)}{p(X)p(Y)}
\]

Based on Kullback-Leibler distance \( D(p|q) \)

Specific Mutual information

Church and Hanks, 1989; Smadja 1990

Only the 1-1 term

\[
SI(X,Y) = \log \frac{P(X,Y)}{P(X)P(Y)}
\]

Association as conditional probabilities

The Dice coefficient (Dice, 1945)

\[
D(X,Y) = \frac{2P(X,Y)}{P(X) + P(Y)}
\]

Similar to Jaccard coefficient

Datasets in the form of matrices

We are given \( n \) objects and \( d \) features describing the objects. (Each object has \( d \) numeric values describing it.)

Dataset

An \( n \)-by-\( d \) matrix \( A \), \( A_{ij} \) shows the “importance” of feature \( j \) for object \( i \).

Every row of \( A \) represents an object.

Goal

1. Understand the structure of the data, e.g., the underlying process generating the data.
2. Reduce the number of features representing the data

Market basket matrices

\( d \) products
(e.g., milk, bread, wine, etc.)

\( n \) customers

\[
A
\]

\( A_{ij} \) = quantity of \( j \)-th product purchased by the \( i \)-th customer

Find a subset of the products that characterize customer behavior
### Social-network matrices

- **Matrix Structure:**
  - **n** users
  - **d** groups
  - \( A_{ij} = \) participation of the \( i \)-th user in the \( j \)-th group

- **Objective:** Find a subset of the groups that accurately clusters social-network users

### Document matrices

- **Matrix Structure:**
  - **n** documents
  - **d** terms
  - \( A_{ij} = \) frequency of the \( j \)-th term in the \( i \)-th document

- **Objective:** Find a subset of the terms that accurately clusters the documents