Statistical learning, complex systems, and pickles

Lecture 2, CSSS09

Greg Leibon
Memento, Inc
Dartmouth College
The Plan

• Part 1: Spectral Clustering
• Part 2: Picklology...and beyond!
• Appendix: Proof of the mighty Commute Theorem
Part 1: Spectral Clustering
“Vagabond Clustering”: find a partition $\{A_k\}_{k=1}^K$ that minimizes

$$Vagabondliness = \sum P(X_{t+1} \in \overline{A}_k \mid X_t = A_k)$$

Finding this partition is usually called “NCut” and it is NP-hard.

Since finding this partition is NP-hard, to find the solution we will need to “relax...”

Key Theorem: $P(X_{t+1} \in \tilde{A} \mid X_t \in A) = \langle \chi_A, \Delta \chi_A \rangle$
\[
\chi_A(i) = \begin{cases} 
\frac{1}{\sqrt{\nu w(A)}} & i \in A \\
0 & i \notin A
\end{cases}
\]

for A and B disjoint

\[
\langle \chi_A, \chi_B \rangle = 0
\]

\[
\langle \chi_A, \chi_A \rangle = 1
\]
Key Theorem: \( P(X_{t+1} \in \tilde{A} \mid X_t \in A) = \langle \chi_A, \Delta \chi_A \rangle \)

Approximate the Vagabond Clustering

\[
\min_{\text{partitions } \{A_k\}} \sum_k \langle \chi_{A_k}, \Delta \chi_{A_k} \rangle
\]

by “relaxing” and solving

\[
\min_{\{v_i | \langle v_i, v_j \rangle = \delta_j^i\}} \sum_i \langle v_i, \Delta v_i \rangle
\]

which from the spectral theorem is the first k Laplace eigen functions.
Proof: \[ P(X_{t+1} \in \bar{A} \mid X_t \in A) = \frac{P(X_{t+1} \in \bar{A}, X_t \in A)}{P(X_t \in A)} \]
\[ = \frac{\sum_{j \in \bar{A}, i \in A} P(X_{t+1} = j, X_t = i)}{P(X_t \in A)} \]
\[ = \frac{\sum_{j \in \bar{A}, i \in A} P(X_{t+1} = j \mid X_t = i)P(X_t = i)}{P(X_t \in A)} \]
\[ = \frac{\sum_{j \in \bar{A}, i \in A} P_i^j w^i}{\mu(A)} \]
\[ = \sum_{j \in \bar{A}, i \in A} P_i^j w^i (\chi_A(i) - \chi_A(j))^2 \]
\[ = \frac{1}{2} \sum_{i,j} P_i^j w^i (\chi_A - \chi_A)^2 = \langle \chi_A, \Delta \chi_A \rangle \quad \text{Q.E.D.} \]
First three (non-trivial) Eigenfunctions

Notice, they are not localized

The Vagabond functions
Can use our eigenfunctions to embed our states in Euclidean space!


Spectral Clustering Algorithm

1. Find the M orthonormal eigenvectors corresponding to the M smallest eigenvalues.

2. Using these eigenfunctions, embed our states into Euclidean space and then apply K-means.

Many Variations...


K-means

• Simplest clustering algorithm is *k*-means
• To run requires fixing \( K = \# \text{(Clusters)} \)
• Requires an Euclidean type embedding
• We are attempting to minimizing a loss function:

\[
L = \sum_{k=1}^{K} \sum_{x_i \in C_k} (x_i - \mu_k)^2
\]
K-means algorithm:
1. Randomly choose points in each cluster and compute centroids.

2. Organize points by distance to the centroids.
3. Update centroids
4. Repeat...
...until stable.
A hard part is choosing $K = \#(\text{Clusters})$

**Elbowloggy:**

$$L = \sum_{k=1}^{K} \sum_{x_i \in C_k} (x_i - \mu_k)^2$$

Though in practice this rarely works in a complex multi-scalar system.
Also, need to account for “non-ballish” geometry Classical K-means
Spectral Clustering

\[ L = \sum_{k=1}^{N} \sum_{x_i \in C_k} d(x_i, \mu_k) \approx 488 \]
Spectral clustering in Mat Lab.

%Sort by eigenvalue the eigenbasis [Eig, O]
Eig = diag(Eig);
[Eig Srt] = sort(Eig);
O = O(:, Srt);

% Now apply K-means 'Rep' times
Emb = O(:, 2:N);
[IDX, C, sumd, D] = kmeans(Emb, K, 'emptyaction', 'drop');

for i = 1:Rep
    [IDX0, C, sumd0, D] = kmeans(Emb, K, 'emptyaction', 'drop');
    if (sum(sumd0) < sum(sumd))
        IDX = IDX0;
        sumd = sumd0;
    end;
end;

why repeat?
Yet another tricky part....

Himalayas

\[ L = \sum_{k=1}^{K} \sum_{x_i \in C_k} (x_i - \mu_k)^2 \]
Watch Out! 100 convergent runs of the spectral k-means were performed. \( \sigma = 1 \)

\[
L = \sum_{k=1}^{N} \sum_{x_i \in C_k} d(x_i, \mu_k) \approx 536
\]

\[
L = \sum_{x_i \in C_k} \sum_{k=1}^{N} d(x_i, \mu_k) \approx 538
\]

\[
L = \sum_{x_i \in C_k} \sum_{k=1}^{N} d(x_i, \mu_k) \approx 534
\]
5.1.3 **Theorem.** If $P$ is an ergodic transition matrix, then the inverse matrix $Z = (I - (P - A))^{-1}$ exists, and

(a) $PZ = ZP$
(b) $Z\xi = \xi$
(c) $\alpha Z = \alpha$
(d) $(I - P)Z = I - A.$
Define the pickle embeddings of a chain: $Z_a^g(i)$

$$\Delta Z_a^g = \frac{e_a}{w_a} - \frac{e_g}{w^g}$$

$$\sum_i Z_a^g(i) = 0$$
What is $Z$...

Green's Function

$G(x, y)$

Potential at $x$ for point source at $y$.

$\Delta G(x, y) = \delta y$
(1.) The function which represents the sum of all the electric particles acting on a given point divided by their respective distances from this point, has the property of giving, in a very simple form, the forces by which it is solicited, arising from the whole electrified mass.—We shall, in what follows, endeavour to discover some relations between this function, and the density of the electricity in the mass or masses producing it, and apply the relations thus obtained, to the theory of electricity.

Firstly, let us consider a body of any form whatever, through which the electricity is distributed according to any given law, and fixed there, and let \( x', y', z' \), be the rectangular co-ordinates of a particle of this body, \( \rho' \) the density of the electricity in this particle, so that \( d\rho'dx'dy'dz' \) being the volume of the particle, \( \rho'dx'dy'dz' \) shall be the quantity of electricity it contains: moreover, let \( r' \) be the distance between this particle and a point \( p \) exterior to the body, and \( V \) represent the sum of all the particles of electricity divided by their respective distances from this point, whose co-ordinates are supposed to be \( x, y, z \), then shall we have

\[
r' = \sqrt{(x' - x)^2 + (y' - y)^2 + (z' - z)^2},
\]

and

\[
V = \frac{\rho'dx'dy'dz'}{r'};
\]

the integral comprehending every particle in the electrified mass under consideration.

Example (\( \mathbb{R}^3 \)):

\[
G(x, y) = \frac{1}{4\pi |x - y|}
\]

....with a second source at infinity
Laplace has shown, in his Méc. Céleste, that the function $V$ has the property of satisfying the equation

$$0 = \frac{d^2 V}{dx^2} + \frac{d^2 V}{dy^2} + \frac{d^2 V}{dz^2},$$

and as this equation will be incessantly recurring in what follows, we shall write it in the abridged form $0 = \delta V$; the symbol $\delta$ being used in no other sense throughout the whole of this Essay.

In order to prove that $0 = \delta V$, we have only to remark, that by differentiation we immediately obtain $0 = \delta \frac{1}{r^2}$, and consequently each element of $V$ substituted for $V$ in the above equation satisfies it; hence the whole integral (being considered as the sum of all these elements) will also satisfy it. This reasoning ceases to hold good when the point $p$ is within the body, for then, the coefficients of some of the elements which enter into $V$ becoming infinite, it does not therefore necessarily follow that $V$ satisfies the equation

$$0 = \delta V,$$

although each of its elements, considered separately, may do so.

... 

Hence, throughout the interior of the mass

$$0 = \delta V + 4\pi \rho;$$

of which, the equation $0 = \delta V$ for any point exterior to the body is a particular case, seeing that, here $\rho = 0.$
The Pickle Embedding
(reduced to 3 dimensions)

\[ \alpha \rightarrow \mathbb{Z}_a^g \]
Applications

The Netflix Challenge

Welcome!

The Netflix Prize seeks to substantially improve the accuracy of predictions about how much someone is going to love a movie based on their movie preferences. Improve it enough and you win one (or more) Prizes. Winning the Netflix Prize improves our ability to connect people to the movies they love.

Read the Rules to see what is required to win the Prizes. If you are interested in joining the quest, you should register a team.

You should also read the frequently-asked questions about the Prize. And check out how various teams are doing on the Leaderboard.

Good luck and thanks for helping!
Random-Walk Computation of Similarities between Nodes of a Graph with Application to Collaborative Recommendation

François Fouss, Alain Pirotte, Member, IEEE, Jean-Michel Renders, and Marco Saerens, Member, IEEE

Abstract—This work presents a new perspective on characterizing the similarity between elements of a database or, more generally, nodes of a weighted and undirected graph. It is based on a Markov-chain model of random walk through the database. More precisely, we compute quantities (the average commute time, the pseudoinverse of the Laplacian matrix of the graph, etc.) that provide similarities between any pair of nodes, having the nice property of increasing when the number of paths connecting those elements increases and when the “length” of paths decreases. It turns out that the square root of the average commute time is a Euclidean distance and that the pseudoinverse of the Laplacian matrix is a kernel matrix (its elements are inner products closely related to commute times). A principal component analysis (PCA) of the graph is introduced for computing the subspace projection of the node vectors in a manner that preserves as much variance as possible in terms of the Euclidean commute-time distance. This graph PCA provides a nice interpretation to the “Fiedler vector,” widely used for graph partitioning. The model is evaluated on a collaborative-recommendation task where suggestions are made about which movies people should watch based upon what they watched in the past. Experimental results on the MovieLens database show that the Laplacian-based similarities perform well in comparison with other methods. The model, which nicely fits into the so-called “statistical relational learning” framework, could also be used to compute document or word similarities, and, more generally, it could be applied to machine-learning and pattern-recognition tasks involving a relational database.
6 Experiments
6.1 Experimental Methodology

Remember that each element of the people and the movie sets corresponds to a node of the graph. Each node of the people set is connected by a link to each movie watched by the corresponding person. Notice that, in this special case, the graph is bipartite. The results shown here do not take into account the numerical value of the ratings provided by the persons but only the fact that a person has or has not watched a movie (i.e., entries in the person-movie matrix are 0s and 1s). Moreover, our experiments do not take the movie_category set into account so that comparisons between the various scoring algorithms remain fair. Indeed, three standard scoring algorithms (i.e., maximum frequency, cosine, and nearest-neighbor algorithms) cannot naturally use the movie_category set to rank the movies.

6.1.1 Data Set

Our experiments were performed on a real movie database from the Web-based recommender system MovieLens (http://www.movielens.umn.edu). We used a sample of their database as suggested in [59]: Enough people (i.e., 943 people) were randomly selected to obtain 100,000 ratings (considering only persons that had rated 20 or more movies on a total of 1,682 movies).
A Wiki Adventure!
<table>
<thead>
<tr>
<th>Clique (graph theory)</th>
<th>Germany</th>
<th>Hungarian language</th>
<th>Pierre de Fermat</th>
<th>Star Wars</th>
<th>Theory of relativity</th>
<th>1989</th>
</tr>
</thead>
</table>

Mark: 7.6/10 | Mark: 7.0/10 | Mark: 6.2/10 | Mark: 7.3/10 | Mark: 7.4/10 | Mark: 8.1/10 | Mark: 5.4/10
Abstract

We introduce a new method for finding nodes semantically related to a given node in a hyperlinked graph: the Green method, based on a classical Markov chain tool. It is generic, adjustment-free and easy to implement. We test it in the case of the hyperlink structure of the English version of Wikipedia, the on-line encyclopedia. We present an extensive comparative study of the performance of our method versus several other classical methods in the case of Wikipedia. The Green method is found to have both the best average results and the best robustness.

Evaluation Methodology. We carried out a blind evaluation of the methods on 7 different articles, chosen for their diversity: (i) Clique (graph theory): a very short, technical article. (ii) Germany: a very large article. (iii) Hungarian language: a medium-sized, quite technical article. (iv) Pierre de Fermat: a short biographical article. (v) Star Wars: a large article, with an important number of links. (vi) Theory of relativity: a short introductory article pointing to more specialized articles. (vii) 1989: a very large article, containing all the important events of year 1989. It was unreasonable to expect our testers to evaluate more articles. In order to avoid any bias, we did not run the methods on these 7 articles before the evaluation procedure was launched.

People were asked to assign a mark between 0 and 10 (10 being the best) to the list of the first 20 results returned by
Why do we like the Pickle Embedding?

\[ ||Z_a^g - Z_b^g|| \] has a natural interpretation via the Commute Time from i to j (denoted \( T_{ab} \)) is the expected number up step needed to get from i to j and back to i.

\[[\text{states Time}] = \text{Commute}(15,25,1000);\]
The Commute Theorem

\[ \|Z_a^g - Z_b^g\|_{Dir}^2 = \|Z_a^b\|_{Dir}^2 = T_{ab} \]

Note: linear combinations simply translate the image, and we can form

\[ Z = \sum_{n=0}^{\infty} (P^n - P^\infty) \]

For a reversible, ergodic Markov chain.

---


A Wiki Adventure, done correctly! i.e. a possibly non-reversible chain
The Commute Theorem

\[ \|Z_a^g - Z_b^g\|_{Dir}^2 = \|Z_a^b\|_{Dir}^2 = T_{ab} \]

and we can use

\[ Z = \sum_{n=0}^{\infty} (P^n - P^\infty) \]

is true for any ergodic Markov chain!
Geometry and conformal invariance

\[ (z_1; z_2; z_3; z_4) = \frac{(z_1 - z_3)(z_2 - z_4)}{(z_2 - z_3)(z_1 - z_4)} \]

\[ N_{z_1; z_2; z_3; z_4} = \frac{-1}{2\pi} \log(||(z_1; z_2; z_3; z_4)||) \]

need 4th point to avoid infinity

This is the distance in the Hyperbolic plane (Klein Model)

\[ d(z, w) = cN_{z, w, z^\infty, w^\infty} \]

Why should we care...

Theorem (Doyle, Steiner 2009) N determines the conformally invariant part of a Markov chain and with w the whole chain.
Appendix: Proof of the Commute Theorem

\[ \left\| Z_a^b \right\|^2 = T_{ab} \]

Why we are doing it.... to do some mathematics (yeah!)....

- Visit the almighty *maximum principle*
- Dwell on the *fundamental mysteries of probability theory*
- Directly interact with the single most important fact about infinity
Commute Theorem:

\[ \| Z_a^b \|^2 = T_{ab} \]

Proof: Let \( p_e \) be the probability of reaching \( j \) before \( i \) when starting at \( i \).

Proposition 1 (Renewal Theory)

\[
T_{ab} = \frac{1}{p_{esc} w^a}
\]

Proposition 2 (Maximum Principle, Potential Theory)

\[
\| Z_a^b \|^2 = \frac{1}{p_{esc} w^a}
\]
The second fundamental mystery of probability theory is that:

$$E(X) = E(E(X \mid Y)).$$

As an application let

$$X = \sum_{k=1}^{N} R_k$$

where $N$ is a positive integer valued random variable and let the $\{R_i\}$ share the expected value $E(R)$. Then

$$E \left( \sum_{k=1}^{N} R_k \right) = E \left( E \left( \sum_{k=1}^{N} R_k \mid N = n \right) \right)$$

$$= \sum_{n=1}^{\infty} E \left( \sum_{k=1}^{n} R_k \right) P(N = n) = \sum_{n=1}^{\infty} nE(R)P(N = n)$$

$$= E(R) \sum_{n=1}^{\infty} nP(N = n) = E(R)E(N)$$

Wald’s Theorem (Renewal Theory)
Proposition 1

\[ T_{ab} = \frac{1}{p_{esc}w^a} \]

\[ R = \# \text{ Steps taken going from } a \text{ to } a \]

\[ N = \# \text{ Returns to } i \text{ on commute} \]

Second Fundamental Mystery of Probability Theory (see the appendix)

\[ E(T_{ab}) = E\left( \sum_{k=1}^{N} R_k \right) \]

\[ = E(N)E(R) \]

Very well known!

\[ E(R) = \frac{1}{w^a} \]
\[ E(N) = 1(p_e) + 2(p_e(1 - p_{esc})) + \ldots \]
\[ = p_e \sum_{k=0}^{\infty} k(1 - p_e)^{k-1} \]
\[ = p_e \frac{1}{(1 - (1 - p_e))^2} = \frac{1}{p_e} \]

Fundamental Theorem of Analysis

\[ \sum_{i=0}^{\infty} r^i = \frac{1}{1 - r} \]
\[ \sum_{i=0}^{\infty} ir^{i-1} = \frac{d}{dr} \left( \sum_{i=1}^{\infty} r^i \right) \]
\[ = \frac{1}{(1 - r)^2} \]

Here's one argument....

\[ N = \#\text{Loops} = \#(S = i) - 1 \]

Law of Large Numbers

\[ E(R) = \lim_{T \to \infty} \frac{1}{N} \sum_{k=1}^{N} R_k \]
\[ = \lim_{T \to \infty} \frac{T}{N} \]
\[ = \lim_{T \to \infty} \frac{1}{N} \sum_{k=1}^{N} \frac{1}{k} \]
\[ = \frac{1}{\omega^i} \]

\[ T_{ab} = \frac{1}{p_{esc}\omega^a} \]
Proposition 2

\[ \Delta \tilde{Z}^b_a = 0 \quad i \neq a, b \]
\[ \tilde{Z}^b_a(a) = 0 \]
\[ \tilde{Z}^b_a(b) = 1 \]

Dirichlet Problem

Unique solution up to constant from Maximum Principle

\[ Z^b_a(i) = C \tilde{Z}^b_a(i) + D \]

\[ \tilde{Z}^b_a(i) \] equals the probability starting at \( i \) of reaching \( b \) before \( a \).

\[ \tilde{Z}^b_a(a) \] is minus the probability of escaping from \( a \) to \( b \) starting at \( a \).

\[ \Delta \tilde{Z}^b_a(b) = \frac{w^a p_e}{w^b} \]

constant function in kernel
\[Z_a^b = \frac{1}{w^a p_e} \tilde{Z}_a^b\]

\[
\|\tilde{Z}_a^b\|_{Dir}^2 = \langle \tilde{Z}_a^b, \Delta \tilde{Z}_a^b \rangle = w^b \frac{w^a p_e}{w^b} = w^a p_e
\]

\[
\|Z_a^b\|_{Dir}^2 = \frac{1}{(w^a p_e)^2} \|\tilde{Z}_a^b\|_{Dir}^2 = \frac{1}{w^a p_e}
\]
On the 2-d Euclidean lattice a drunk Lord will always find his way home. In three dimensions he may not be so lucky!

“I have not had a moment's peace or happiness in respect to electromagnetic theory since November 28, 1846. All this time I have been liable to fits of ether dipsomania, kept away at intervals only by rigorous abstention from thought on the subject.”
~Lord Kelvin~ (to FitzGerald 1896)

* Dipsomania is a term USUALLY related to an incontrollable craving for alcohol.... the obsession is so compulsive that the dipsomaniac will ingest whatever intoxicating liquid is at hand, whether it is fit for consumption or not. Dipsomania differs from alcoholism in that it is an uncontrollable periodic lust for alcohol, with, in the interim, no desire for alcoholic beverages.

Ether Dipsomania is a term related to an incontrollable craving for a consistent and appealing theory of something in the form of an analogon to the theory of electromagnetism...the obsession is so compulsive that the ether dipsomaniac will....