Lecture 23

9.6 Laplacian Eigenmaps

Taken from [12].

We at last are ready to integrate geometry into data science. We begin with *Laplacian Eigenmaps* [12].

We define "big O" and "little o" notation before beginning. Let $f : \mathbb{R} \to \mathbb{R}$ and $h : \mathbb{R} \to \mathbb{R}$. We say that

$$f(x) = O(h(x)) \text{ as } x \to a$$

if and only if there exists constants $C, \delta > 0$ such that

$$|f(x)| \leq C|h(x)| \text{ when } 0 < |x - a| < \delta$$

In other words, $f(x) = O(h(x))$ as $x \to a$ if

$$\lim_{x \to a} \left| \frac{f(x)}{h(x)} \right| < \infty.$$

We shall use this often when $a = 0$, and in this case simply write $f(x) = O(h(x))$.

"Little o" notation is similar but is a stronger condition. We say

$$f(x) = o(h(x)) \text{ as } x \to a$$

if and only if

$$\lim_{x \to a} \left| \frac{f(x)}{h(x)} \right| = 0.$$

9.6.1 Optimal manifold embeddings

We motivate the Laplacian Eigenmaps algorithm by considering how to embed $\mathcal{M} \subset \mathbb{R}^N$ into a lower dimensional space. Let us start by trying to map $\mathcal{M}$ to the real line $\mathbb{R}$. We are looking here for a map from the manifold to the real line such that points close together on the manifold are mapped close together on the line. Let $f : \mathcal{M} \to \mathbb{R}$ be such a map.

Let $p, q \in \mathcal{M}$. The following proposition shows that $\|\nabla f(p)\|_p$ provides us with an estimate of how far apart f maps the two points.

Proposition 7. *Let $f : \mathcal{M} \to \mathbb{R}$ and $p, q \in \mathcal{M}$. Then*

$$|f(p) - f(q)| \leq d_{\mathcal{M}}(p,q)\|\nabla f(p)\|_p + o(d_{\mathcal{M}}(p,q)).$$

Proof. Let

$$l = d_{\mathcal{M}}(p,q),$$

and let $\gamma : [0,l] \to \mathcal{M}$ be a geodesic curve parameterized by length (meaning that $\|\gamma'(t)\|_{\gamma(t)} = 1$ for all t) such that $\gamma(0) = p$ and $\gamma(l) = q$. Then:

$$f(q) = f(p) + \int_0^l \partial_{\gamma'(t)} f(\gamma(t))\, dt = f(x) + \int_0^l \langle \nabla f(\gamma(t)), \gamma'(t)\rangle_{\gamma(t)}\, dt.$$

Now, by the Cauchy-Schwartz inequality (which holds on Riemannian manifolds too),

$$\langle \nabla f(\gamma(t)), \gamma'(t)\rangle_{\gamma(t)} \leq \|\nabla f(\gamma(t))\|_{\gamma(t)} \|\gamma'(t)\|_{\gamma(t)} = \|\nabla f(\gamma(t))\|_{\gamma(t)}.$$

Now Taylor expand the function $t \mapsto \|\nabla f(\gamma(t))\|_{\gamma(t)}$ around $t = 0$ to get:

$$\begin{aligned}\|\nabla f(\gamma(t))\|_{\gamma(t)} &= \|\nabla f(\gamma(0))\|_{\gamma(0)} + \partial_t \|\nabla f(\gamma(t))\|_{\gamma(t)}\Big|_{t=0} \cdot t + \cdots \\ &= \|\nabla f(p)\|_p + O(t)\end{aligned}$$

Thus:

$$f(q) \leq f(p) + \int_0^l \|\nabla f(p)\|_p + O(t)\, dt = f(p) + l\|\nabla f(p)\|_p + o(l),$$

which gives the result. □

If we want to find a map f that does not distort local geodesic distances on $\mathcal{M}$ too much, Proposition 7 motivates us to look for a map $f : \mathcal{M} \to \mathbb{R}$ that preserves locality on average by trying to find:

$$\operatorname*{arg\,inf}_{\|f\|^2=1} \int_{\mathcal{M}} \|\nabla f\|^2 \tag{62}$$

Set

$$\mathcal{L} = -\Delta$$

Then by (61) we have:

$$\int_{\mathcal{M}} \|\nabla f\|^2 = \int_{\mathcal{M}} f\mathcal{L}f = \langle f, \mathcal{L}f\rangle$$

Thus the solution to (62) must be an eigenfunction of $\mathcal{L}$. If $\mathcal{M}$ is compact and connected, then $\mathcal{L}$ has a discrete set of eigenvalues $0 = \lambda_1 < \lambda_2 \leq \lambda_3 \leq \cdots$ with Neumann boundary condition. In this case the first eigenfunction $\phi_1 \in \mathbf{L}^2(\mathcal{M})$ is the constant function, which maps $\mathcal{M}$ to a single point. Since this is not terribly interesting and erases all geometric information about $\mathcal{M}$, we replace (62) with

$$\arg\inf_{\substack{\|f\|=1 \\ \langle f,\phi_1\rangle=0}} \int_{\mathcal{M}} \|\nabla f\|^2$$

It follows that the solution to (9.6.1) is the second eigenfunction of $\mathcal{L}$, which is ϕ_2. Thus the map $\phi_2 : \mathcal{M} \to \mathbb{R}$ provides an optimal one dimensional embedding of $\mathcal{M}$. If instead we want an optimal m dimensional embedding, we take $\Phi : \mathcal{M} \to \mathbb{R}^d$, where:

$$\Phi(p) = (\phi_2(p), \ldots, \phi_{m+1}(p)),$$

where ϕ_k is the k^{th} eigenfunction of $\mathcal{L}$ with eigenvalue λ_k.

9.6.2 Optimal graph embeddings

Now let us examine the graph analogue of the previous section. Suppose we have a connected, weighted graph $G = (V, E, w)$ with non-negative weights $w(i, j)$ and with n vertices $V = \{x_1, \ldots, x_n\}$. We consider the problem of mapping the graph G to a line so that connected points stay as close together as possible. Let $\mathbf{y} = (y_1, \ldots, y_n)^T$ be such a map. A reasonable criterion for choosing a "good" map is to minimize the following objective function:

$$\sum_{(i,j)\in E} w(i, j)(y_i - y_j)^2, \tag{63}$$

subject to appropriate constraints. The objective function with our choice of weights $w(i, j)$ incurs a heavy penalty if neighboring points x_i and x_j with large weight $w(i, j)$ are mapped far apart. Therefore, minimizing it is an attempt to ensure if x_i and x_j are "close," then y_i and y_j are close as well.

From our work in spectral graph theory, we know that

$$\sum_{(i,j)\in E} w(i, j)(y_i - y_j)^2 = \mathbf{y}^T L \mathbf{y},$$

where $L = D - A$ is the graph Laplacian with $A_{ij} = w(i,j)$. The minimization problem is then recast as:

$$\underset{\substack{\mathbf{y} \\ \mathbf{y}^T D\mathbf{y}=1}}{\arg\inf}\, \mathbf{y}^T L\mathbf{y}. \tag{64}$$

The constraint $\mathbf{y}^T D\mathbf{y} = 1$ removes an arbitrary scaling factor in the embedding. We view $\mathbf{y}^T D\mathbf{y}$ as the norm of $\mathbf{y}$ over the measure D on the graph. Making the change of variable

$$\mathbf{z} = D^{1/2}\mathbf{y}$$

we get the equivalent optimization problem:

$$\underset{\substack{\mathbf{z} \\ \mathbf{z}^T\mathbf{z}=1}}{\arg\inf}\, \mathbf{z}^T N\mathbf{z},$$

where $N = D^{-1/2}LD^{-1/2}$ is the normalized graph Laplacian. It follows that the vector $\mathbf{z}$ that minimizes the objective function is the eigenvector φ_1 of N with minimum eigenvalue γ_1. We know from our previous work that $\varphi_1 = \mathbf{d}^{1/2}$, where $\mathbf{d}^{1/2}[i] = \sqrt{D_{ii}}$, and that $\gamma_1 = 0$. It follows that the solution to (64) is $\mathbf{y}^\star = D^{-1/2}\varphi_1 = \mathbf{1}$, the constant vector.

To eliminate this trivial solution which collapses all vertices in the graph to a single point, we put an additional constraint of orthogonality and look for:

$$\underset{\substack{\mathbf{y} \\ \mathbf{y}^T D\mathbf{y}=1 \\ \mathbf{y}^T D\mathbf{1}=0}}{\arg\inf}\, \mathbf{y}^T L\mathbf{y}.$$

The condition $\mathbf{y}^T D\mathbf{1} = 0$ can be interpreted as removing a translation invariance in $\mathbf{y}$.

By a previous homework exercise, we know that this new optimization problem is equivalent to

$$\underset{\substack{\mathbf{z} \\ \mathbf{z}^T\mathbf{z}=1 \\ \mathbf{z}^T\mathbf{d}^{1/2}=0}}{\arg\inf}\, \mathbf{z}^T N\mathbf{z}.$$

From that same homework we know that the solution is $\mathbf{z}^\star = \varphi_2$, the second eigenvector of N with eigenvalue $\gamma_2 > 0$. The new solution $\mathbf{y}^\star$ is given by $\mathbf{y}^\star = D^{-1/2}\varphi_2$. Notice that this embedding preserves the local geometry of the graph since it minimizes (63), but by Cheeger's Theorem it also let's one cut the data set into two clusters!

If we want to embed the graph into an m-dimensional Euclidean space with coordinate vectors $\mathbf{y}_1, \ldots, \mathbf{y}_m$, we minimize

$$\sum_{(i,j)\in E} w(i,j)\|\mathbf{y}^{(i)} - \mathbf{y}^{(j)}\|^2,$$

where

$$\mathbf{y}^{(i)} = (\mathbf{y}_1[i], \ldots, \mathbf{y}_m[i])^T$$

is the m-dimensional representation of x_i. Let

$$Y = [\mathbf{y}_1 \cdots \mathbf{y}_m]$$

be the $n \times m$ matrix with columns given by $\mathbf{y}_1, \ldots, \mathbf{y}_m$. Then the optimization problem can be recast as:

$$\underset{\substack{Y \\ Y^T D Y = I}}{\arg\inf} \operatorname{Tr}(Y^T L Y). \tag{65}$$

Standard arguments show that we can compute Y by computing the lowest lying eigenvectors $\varphi_2, \ldots, \varphi_{m+1}$ of N, and then setting $\mathbf{y}_k^\star = D^{-1/2}\varphi_k$.

Exercise 35 (optional). Prove that the solutions $\mathbf{y}_1^\star, \ldots, \mathbf{y}_m^\star$ to the optimization (65) are given by the eigenvectors corresponding to the lowest m eigenvalues of the generalized eigenvalue problem:

$$L\mathbf{y} = \lambda D\mathbf{y}$$

Exercise 36 (optional). Implement the Laplacian Eigenmaps algorithm.

References

[1] Bernhard Schölkopf and Alexander J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Adaptive Computation and Machine Learning. The MIT Press, 2002.

[2] Afonso S. Bandeira. Ten lectures and forty-two open problems in the mathematics of data science. MIT course *Topics in Mathematics of Data Science*, 2015.

[3] Jon Shlens. A tutorial on principal component analysis. arXiv:1404.1100, 2014.

[4] Karl Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine, Series 6*, 2(11):559–572, 1901.

[5] V. A. Marchenko and L. A. Pastur. Distribution of eigenvalues in certain sets of random matrices. *Mat. Sb. (N.S.)*, 72(114):507–536, 1967.

[6] J. Baik, G. Ben-Arous, and S. Péché. Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *The Annals of Probability*, 33(5):1643–1697, 2005.

[7] Debashis Paul. Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statistica Sinica*, pages 1617–1642, 2007.

[8] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer-Verlag New York, 2nd edition, 2009.

[9] Athanasios Tsanas and Angeliki Xifara. Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools. *Energy and Buildings*, 49:560–567, 2012.

[10] Daniel A. Spielman. Spectral graph theory. *Yale Course Notes*, Fall, 2009.

[11] Jianzhong Wang. *Geometric Structure of High-Dimensional Data and Dimensionality Reduction*. Springer, 2012.

[12] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396, 2003.

[13] Ronald R. Coifman and Stéphane Lafon. Diffusion maps. *Applied and Computational Harmonic Analysis*, 21:5–30, 2006.