1 Manifold structure in graph embeddings[1]

Graph embedding is the process of assigning nodes in a graph to points in some high dimensional space. The Manifold Hypothesis states that real-world high dimensional data actually live on a low dimensional manifold embedded in the high dimensional space. So, can we choose an embedding procedure that gives rise to a manifold structure that satisfies this hypothesis?

If it turns out that we have a high dimensional dataset that has a manifold structure such that the data lie close to this structure, then it turns out we can carry out statistical procedures with roughly the same scaling/rates etc as if the data was in the intrinsic dimension of this manifold.

1.1 Latent Position Model

Given a symmetric (kernel) function $f : \mathcal{Z} \times \mathcal{Z} \to [0, 1]$ with $\mathcal{Z} \subseteq R^d$, an undirected graph with n nodes is said to follow a latent position network if it has an adjacency matrix $\boldsymbol{A}_{ij} \mid Z_1, \ldots, Z_n \overset{ind}{\sim} \text{Bernoulli}(f(Z_i, Z_j))$ for $i < j$ where $Z_1, \ldots, Z_n$ are i.i.d replicates of a random vector Z with distribution F_Z supported on $\mathcal{Z}$.

1.2 Spectral Embedding

If we take an undirected graph with (symmetric) adjacency matrix $\boldsymbol{A}$ (from a latent position model) and finite embedding dimension $D > 0$, we compute a spectral (i.e. eigen-) decomposition $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{U}^T + \underbrace{\boldsymbol{U}_\perp \boldsymbol{S}_\perp \boldsymbol{U}_\perp^T}_{Noise}$ where $\boldsymbol{S}$ is a diagonal matrix of the D largest eigenvalues of $\boldsymbol{A}$. Now we can define the adjacency spectral embedding by $\hat{\mathbf{X}} = [\hat{X}_1, \ldots, \hat{X}_n]^T = \boldsymbol{U}|\boldsymbol{S}|^{\frac{1}{2}} \in R^{n \times D}$.

We now need to prove that the embeddings $\hat{X}_i$ actually provide consistent estimates of $X_i = \phi(Z_i)$.

1.2.1 Generalised Random Dot Product Graph

Given a set of points $X_1, \ldots, X_n \in R^D \overset{i.i.d}{\sim} F_X$ we have an adjacency matrix $\boldsymbol{A}_{ij} \overset{ind}{\sim} \text{Bernoulli}(X_i^T \mathbf{I}_{pq} X_j)$ where $\mathbf{I}_{pq} = \begin{pmatrix} \mathbf{I}_p & \mathbf{0} \\ \mathbf{0} & -\mathbf{I}_q \end{pmatrix}$ is a block diagonal matrix of the p-dimensional identity matrix and the negative of the q-dimensional identity, and $p + q = D$. By subsequently introducing a matrix $\boldsymbol{Q} \in O(p, q)$ where $O(p, q) = \{\boldsymbol{M} \in R^{D \times D} \mid \boldsymbol{M}\mathbf{I}_{pq}\boldsymbol{M}^T = \mathbf{I}_{pq}\}$ is the indefinite orthogonal group it can be proved that the embeddings are consistent.

The result is that

$$d_H(S(\hat{\mathbf{X}}), S(\boldsymbol{Q}^{-1}\mathbf{X})) \leq c \left\|\boldsymbol{Q}^{-1}\right\| \max_{i \in [n]} \left\|\boldsymbol{Q}\hat{X}_i - X_i\right\|_2 \to 0$$

with high probability, which effectively states that in some large limit in the number of data points, the embedding *is* a consistent estimator of the points X_i on the low-dimensional manifold.

[1] This section is a rough summary of [2].

1.3 Spectral representation of latent position network models

Now we want to define a map $\phi : R^d \to L^2(R^d)$ that transforms the latent position Z_i into a vector $X_i = \phi(Z_i)$ of which the point $\hat{X}_i$ is an estimator.

Next we define an operator

$$Ag(x) = \int_{R^d} f(x,y)g(y)dy$$

for $g \in L^2(R^d)$. Provided f is symmetric, A is self-adjoint (A.1) which means we can define a set of orthonormal eigenfunctions u_j satisfying $Au_j = \lambda_j u_j$. We can extend this set of eigenfunctions to include functions in the null space of A (i.e. with eigenvalue 0) to form a complete basis of $L^2(R^d)$ which allows us to represent a function g in terms of the basis:

$$g = \sum_j \langle g, u_j \rangle u_j, \qquad Ag = \sum_j \lambda_j \langle g, u_j \rangle u_j$$

Now we let $\phi(x) = (|\lambda_j|^{\frac{1}{2}} u_j(x))_{j=1,\dots,D}$ and show that with this representation we can recover the spectral definition for A so that we show that $f(Z_i, Z_j) = X_i^T \mathbf{I}_{pq} X_j$ by showing (A.3):

$$\begin{aligned} \phi(x)^T \mathbf{I}_{pq} \phi(y) = f(x,y) \quad &\text{(almost everywhere)} \\ \implies \phi(z_i)^T \mathbf{I}_{pq} \phi(z_j) &= f(z_i, z_j) \end{aligned}$$

1.4 Proof of $\dim_H(\mathcal{M}) \leq d$

If we cover R^d with sets S_i then $\bigcup \phi(S_i)$ is a cover of $\phi(R^d)$. To guarantee that the Hausdorff dimension of the points defined by X_i is $\leq d$ we require that $|\phi(S_i)|^a \leq c|S_i|^a$ (directly from the definition of Hausdorff dimension, noting that the constant c does not affect the dimensionality).

If we compare distances in both spaces:

$$\begin{aligned} \|\phi(x) - \phi(y)\|^2 &= \langle \phi(x), \phi(x) \rangle + \langle \phi(y), \phi(y) \rangle - 2\langle \phi(x), \phi(y) \rangle \\ &= f(x,x) + f(y,y) - 2f(x,y) \\ &\leq c \|x - y\|^2 \end{aligned}$$

A.3

This shows that distances in the feature space are at most equal to distances in the latent space. This means that if we form a cover in the space R^d with sets S_i, their diameter, i.e. the distance between the two farthest apart points in each set S_i, must be greater than or equal to the diameter of the covering sets $\phi(S_i)$ in the feature space, thereby demonstrating that the dimension in feature space must be less than or equal to d.

2 Optimistic bounds for multi-output prediction[2]

2.1 Multi-output prediction

In multi-output prediction we want to learn a prediction function $f : \mathcal{X} \to \mathcal{V} \subseteq R^q$ that takes some vector of covariates $X \in \mathcal{X}$ and outputs a q-dimensional response.

[2] This section is a rough summary of [1]

For example, we might want to assign X to a subset of q possible labels.

In multi-label classification there is usually only a small subset of labels assigned, i.e. $k \ll q$ so the output Y is a k-sparse binary vector. In multi-output regression, we will output a response $Y \in R^q$.

2.2 Statistical setting

Given a hypothesis class (of finite Rademacher complexity (B.1)) of functions $f \in \mathcal{F} \subseteq \mathcal{M}(\mathcal{X}, \mathcal{V})$, the goal of the learner is to find one such function that minimizes the test error $\mathcal{E}_{\mathcal{L}}(f) = \mathrm{E}_{(X,Y)\sim P}[\mathcal{L}(f(X), Y)]$ where $\mathcal{L} : \mathcal{V} \times \mathcal{Y} \to R$ is an appropriate loss function and P is some unknown probability distribution. In practice we cannot observe the test error, so we instead compute the empirical error $\hat{\mathcal{E}}_{\mathcal{L}}(f) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}(f(X_i), Y_i)$ over some dataset $\mathcal{D} = [(X_i, Y_i)]_{i=1}^{n}$.

2.3 Generalisation bounds

Generalisation bounds have two highly useful features: providing information regarding which factors in your problem make it more or less difficult and providing motivation for constructing future algorithms that aim to minimize the generalisation bound (which we can directly compute). Generalisation bounds state that with high probability

$$\mathcal{E}_{\mathcal{L}}(f) \leq \hat{\mathcal{E}}_{\mathcal{L}}(f) + \text{Complexity term} \quad \forall f \in \mathcal{F} \subseteq \mathcal{M}(\mathcal{X}, \mathcal{V})$$

A classical generalisation bound is given by (B.3):

$$\mathcal{E}_{\mathcal{L}}(f) \leq \hat{\mathcal{E}}_{\mathcal{L}}(f) + \underbrace{2\lambda\mathfrak{R}_n(\mathcal{F}) + b\sqrt{\frac{\log(1\delta)}{n}}}_{\text{Complexity term}} \quad \forall f \in \mathcal{F} \subseteq \mathcal{M}(\mathcal{X}, \mathcal{V})$$

with probability at least $1 - \delta$ where $\mathcal{L}$ is a λ-Lipschitz loss function such that $|\mathcal{L}(u, y) - \mathcal{L}(v, y)| \leq \lambda|u - v|$ for $y \in \mathcal{Y}; u, v \in R$. The key element is that with a Lipschitz loss (which many loss functions satisfy) we have a rate of $O(n^{-\frac{1}{2}})$ and that the output function $f \in \mathcal{F}$ has a 1-dimensional output.

What do we want from our bounds?

- Applicable to a broad range of function classes $\mathcal{F}$, with no strong constraints on dimensionality.
- Only limited dependence upon unobservable properties of P, i.e. should only depend on easily computable sample statistics.
- Decays optimally as $n \to \infty$, i.e. our bounds should improve. Optimal improvement in the limit.
- Degrade gracefully as $q \to \infty$, i.e. as the output dimension increases, our bounds should get worse, but not too dramatically in the limit.

2.4 Self-bounding Lipschitz condition

To satisfy the demands for our bounds, we define the following: We say a loss function $\mathcal{L}$ is (λ, θ)-self-bounding Lipschitz with parameters $\lambda, \theta \geq 0$ if for all labels

$y \in \mathcal{Y}$ and outputs $u, v \in \mathcal{V}$

$$|\mathcal{L}(u, y) - \mathcal{L}(v, y)| \leq \lambda \cdot \max\{\mathcal{L}(u, y), \mathcal{L}(v, y)\}^{\theta} \cdot \|u - v\|_{\infty}$$

- $\theta = 0$ corresponds to a Lipschitz condition with respect to the supremememum norm.
- $\theta = \frac{1}{2}$ holds for twice differentiable loss functions when $q = 1$

Obtains bounds of order $O(n^{-\frac{1}{2(1-\theta)}})$, provided the empirical error is small, which interpolates between slow ($O(n^{-\frac{1}{2}})$) and fast ($O(n^{-1})$) rates, based on θ. Note that this condition depends only on the loss function, not the distribution of the data.

2.4.1 Example: Multinomial logistic loss

The multinomial logistic loss is often used for multi-label classification and is given by

$$\mathcal{L}(u, y) = \log\left(\sum_{j \in [q]} \exp(u_j - u_y)\right)$$

This is (λ, θ)-self-bounding Lipschitz with $\lambda = 2$ and $\theta = \frac{1}{2}$ which implies fast rates.

2.4.2 Projected function class

In order to estimate the complexity of our multi-output functions with outputs in R^q, we need to reformulate them as a family of functions with outputs in R (since the Rademacher complexity deals with function classes with outputs in R). This is done by projecting our function class to a new class with an additional argument and outputs in R.

We do this by defining $(\Pi \circ f)(x, j) = \pi_j(f(x))$ where $\pi_j : R^q \to R$ maps the j^{th} coordinate projection and hence

$$\Pi \circ \mathcal{F} = \{\Pi \circ f \mid f \in \mathcal{F}\} \subseteq \mathcal{M}(\mathcal{X} \times [q], R)$$

Often the complexity of this new class is

$$\mathfrak{R}_{nq}(\Pi \circ \mathcal{F}) = \tilde{O}\left(\frac{1}{\sqrt{nq}}\right)$$

[3]

2.5 New generalisation bound

With the self-bounding constraint imposed a new generalisation bound (B.5) is derived with the implication for empirical risk minimisation that

$$\mathcal{E}_{\mathcal{L}}(\hat{f}) \leq \tilde{O}\left(\left(\frac{\lambda}{\sqrt{n}}\right)^{\frac{1}{1-\theta}} + \epsilon\right)$$

where $\epsilon = \inf_{f \in \mathcal{F}}\{\mathcal{E}_{\mathcal{L}}(f)\}$.

The bound is optimal in two senses, one that the range over θ, $[0, 1/2]$ cannot be improved upon in terms of observed rates and in that it cannot be improved upon upto polylogarithmic factors.

[3] $\tilde{O}$ denotes Big-O notation that ignores logarithmic factors

Appendices

A Manifold structure in graph embeddings

A.1 Self-adjointness

We say an operator is *self-adjoint* if it satisfies the following relation:

$$\langle Ax, y\rangle = \langle x, Ay\rangle$$

i.e. $A^\dagger = A$.

So given our operator $Ag(x) = \int_{R^d} f(x,y)g(y)dy$ we can easily show its self-adjointness provided f is symmetric:

$$\begin{aligned}\langle Ag, h\rangle &= \int f(x,y)g(y)h(x)dxdy \\ \langle g, Ah\rangle &= \int g(x)f(x,y)h(y)dxdy \\ &= \int g(y)f(y,x)h(x)dxdy\end{aligned}$$

where we have simply swapped labels x and y in the last line. Clearly the inner products are the same, provided $f(x,y) = f(y,x)$.

A.2 Properties of Sets

A.2.1 Diameter of a set

We define the diameter of a set as

$$|S| = \sup\{\rho(x,y) \mid x, y \in S\}$$

for $S \subseteq X$ where X is a metric space with metric $\rho(\cdot,\cdot)$.

A.2.2 Hausdorff Dimension

In order to demonstrate that our data lie on a low-dimensional set after embedding, we need an appropriate definiton of the dimension of a set. In this vein, we first define the a-dimensional *Hausdorff content* of a set as:

$$\mathcal{H}^a(\mathcal{M}) = \inf\left\{\sum_i |S_i|^a \mid \mathcal{M} \subseteq \bigcup_i S_i\right\}$$

which is a measure of the size of the set $\mathcal{M}$ in the a^{th} dimension. The union of S_i sets forms a cover of $\mathcal{M}$ and $|S_i|$ denotes the diameter of the set S_i. As we decrease the diameter of the sets S_i, we need more of them to cover $\mathcal{M}$. If for example, we take a curve C of unit length and cover it with a finite set of n balls, each with diameter $\delta = 1/n$, then we have $\mathcal{H}^a(C) = \inf\left\{\sum_i^n \frac{1}{n^a}\right\} = \inf\{\frac{n}{n^a}\}$. As we shrink the diameter as much as possible, i.e. $\lim_{\delta\to 0}$, then we have $\mathcal{H}^a(C) = \lim_{n\to\infty} n^{1-a}$ which equals 0 for any $a > 1$.

From this we can then define the *Hausdorff dimension* by

$$\dim(\mathcal{M}) = \inf\{a \mid \mathcal{H}(\mathcal{M}) = 0\}$$

i.e. the maximum value of a such that the next larger value of a would give a Hausdorff content of 0. For the example above, then we see that $\dim(C) = 1$.

A.2.3 Hausdorff Distance

Hausdorff distance measures how far apart two subsets of a metric space are. Two sets X and Y are considered to be within *Hausdorff distance* r of one another iff every point of X is within r of some point in Y and vice versa. It is defined by:

$$d_H(X,Y) = \inf\{r > 0 \mid X \subseteq N_r(Y), Y \subseteq N_r(X)\}$$
$$N_r(X) = \{y \mid \rho(x,y) < r \text{ for some } x \in X\}$$

where N_r is the neighbourhood of points around X that are no further than r from any point in X.

A.3 Spectral representation of latent position network models

Indefinite inner product on the feature map under the operator A:

$$\begin{aligned}
&\int (\phi(x)^T \mathbf{I}_{pq} \phi(y)) g(y) dy \\
&= \int \sum_j \lambda_j u_j(x) u_j(y) g(y) dy \\
&= \sum_j \lambda_j u_j(x) \int u_j(y) g(y) dy \\
&= \sum_j \lambda_j \langle g, u_j \rangle u_j(x) \\
&= Ag(x) \\
&= \int f(x,y) g(y) dy \\
\implies &\phi(x)^T \mathbf{I}_{pq} \phi(y) = f(x,y) \quad \text{(almost everywhere)} \\
\implies &\phi(z_i)^T \mathbf{I}_{pq} \phi(z_j) = f(z_i, z_j)
\end{aligned}$$

Usual inner product on the feature map under the operator A:

$$\begin{aligned}
&\int \langle \phi(x), \phi(y) \rangle g(y) dy \\
&= \int \sum_j |\lambda_j| u_j(x) u_j(y) g(y) dy \\
&= \sum_j |\lambda_j| \langle g, u_j \rangle u_j(x) \\
&= |A| g(x) \\
&= \int |f(x,y)| g(y) dy \\
\implies &\langle \phi(x), \phi(y) \rangle = |f(x,y)| \quad \text{(almost everywhere)}
\end{aligned}$$

Now if f is positive definite, then $|f(x,y)| = f(x,y)$.

This means that

$$\begin{aligned}\|\phi(x)-\phi(y)\|^2 &= \langle\phi(x),\phi(x)\rangle + \langle\phi(y),\phi(y)\rangle - 2\langle\phi(x),\phi(y)\rangle \\ &= f(x,x)+f(y,y)-2f(x,y)\end{aligned}$$

Finally, we make the assumption that $f(x,x)+f(y,y)-2f(x,y) \leq c\,\|x-y\|^2$. (?) I don't understand this assumption.

A.3.1 Indefinite inner product

If we define the indefinite inner product as

$$[g,h] = \langle Hg, h\rangle$$

where H is some Hermitian(/self adjoint) matrix (i.e $H^\dagger = H$) and use the basis expansions of $g = \sum_j \langle g, u_j\rangle u_j$ and h, we get

$$[g,h] = \sum_{jk} \langle g,u_j\rangle\langle Hu_j,u_k\rangle\langle h,u_k\rangle$$

Now if we set $H = A$, then we get

$$[g,h] = \sum_j \lambda_j \langle g,u_j\rangle\langle h,u_j\rangle$$

Note that this scalar product is not positive definite.

B Optimistic bounds for multi-output prediction

B.1 Rademacher complexity

Rademacher complexity is a measure of the size of some function class $\mathcal{G} \subseteq \mathcal{M}(\mathcal{Z}, R)$. We do this by assessing how well functions in that class can correlate with random data, given by sequences of coin flips $\boldsymbol{\sigma} = (\sigma_i)_{i\in[n]} \in \{-1,+1\}^n$. The idea is that larger function classes will be able to correlate better with random sequences, hence this is a proxy for the size of the class.

The definition of the empirical Rademacher complexity is

$$\hat{\mathfrak{R}}_{\mathbf{z}} = \sup_{\tilde{\mathcal{G}}\subseteq\mathcal{G}:|\tilde{\mathcal{G}}|<\infty} \mathrm{E}_{\boldsymbol{\sigma}}\left(\sup_{g\in\tilde{\mathcal{G}}}\frac{1}{n}\sum_{i\in[n]}\sigma_i\cdot g(z_i)\right)$$

Given $n \in N$ we define the worst case Rademacher complexity by

$$\mathfrak{R}_n(\mathcal{G}) = \sup_{\mathbf{z}\in\mathcal{Z}^n} \hat{\mathfrak{R}}_{\mathbf{z}}(\mathcal{G})$$

B.2 Rademacher contraction inequality

We define a new class of functions, by the composition of a loss function with the original class of functions

$$\mathcal{L}\circ\mathcal{F} = \{(x,y)\mapsto\mathcal{L}(f(x),y) \mid f\in\mathcal{F}\} \subseteq \mathcal{M}(\mathcal{X}\times\mathcal{Y}, R)$$

then the Rademacher complexity of this composed function class is related to the Rademacher complexity of the original function class by

$$\mathfrak{R}_n(\mathcal{L}\circ\mathcal{F}) \leq \lambda\mathfrak{R}_n(\mathcal{F})$$

B.3 Rademacher concentration inequality

The Rademacher concentration inequality is an analog of Hoeffdings inequality and tells us that the expectation of some function g in our class $\mathcal{G}$ is bounded above by the empirical average + complexity terms:

$$\mathrm{E}[g(Z)] \leq \frac{1}{n}\sum_{i=1}^{n} g(Z_i) + 2\mathfrak{R}_n(\mathcal{G}) + b\sqrt{\frac{\log(1/\delta)}{n}}$$

B.4 Proof of the classical generalisation bound

$$\begin{aligned}
\mathcal{E}_{\mathcal{L}}(f) &= \mathrm{E}_{(X,Y)\sim P}[\mathcal{L}(f(X),Y)] \\
&\leq \frac{1}{n}\sum_{i} \mathcal{L}(f(X_i),Y_i) + 2\mathfrak{R}_n(\mathcal{L}\circ\mathcal{F}) + b\sqrt{\frac{\log(1/\delta)}{n}} \quad \text{(concentration ineq.)} \\
&\leq \frac{1}{n}\sum_{i} \mathcal{L}(f(X_i),Y_i) + 2\lambda\mathfrak{R}_n(\mathcal{F}) + b\sqrt{\frac{\log(1/\delta)}{n}} \quad \text{(contraction ineq.)} \\
&= \hat{\mathcal{E}}_{\mathcal{L}}(f) + 2\lambda\mathfrak{R}_n(\mathcal{F}) + b\sqrt{\frac{\log(1/\delta)}{n}}
\end{aligned}$$

B.5 New generalisation bound

The new derived generalisation bound is given by

$$\mathcal{E}_{\mathcal{L}}(f) \leq \hat{\mathcal{E}_{\mathcal{L}}}(f) + C_0\left(\sqrt{\hat{\mathcal{E}}_{\mathcal{L}}(f)\cdot\Gamma^{\lambda\theta}_{nq\delta}(\mathcal{F})} + \Gamma^{\lambda\theta}_{nq\delta}(\mathcal{F})\right)$$

with

$$\Gamma^{\lambda\theta}_{nq\delta}(\mathcal{F}) = \left(\left(\sqrt{q}\log^{3/2}(e\beta nq)\mathfrak{R}_{nq}(\Pi\circ\mathcal{F}) + \frac{1}{\sqrt{n}}\right)\right)^{\frac{1}{1-\theta}} + b\frac{\log(n/\delta)}{n}$$

References

[1] Henry WJ Reeve and Ata Kaban. Optimistic bounds for multi-output prediction, 2020.

[2] Patrick Rubin-Delanchy. Manifold structure in graph embeddings, 2020.