Lecture 13: Probit continued + Graph models

13.1 Probit model continued

Last lecture we finished with the Probit misfit/loss function. Recall our model & setting

$$X = \{\underline{x}_1, \dots, \underline{x}_n\} \subset \mathcal{X} \subseteq \mathbb{R}^d$$

$$Y = (Y_1, \dots, Y_n) \in \{-1, +1\}^n \leftarrow \text{Possibly noisy labels}$$

$$\ell : \mathcal{X} \to \{-1, 1\} \quad \text{(true label)}$$

Goal: Given x_{n+1} find predict $\ell(\underline{x}_{n+1})$

model $Y_j = \text{sign}(f(\underline{x}_j) + \varepsilon_j)$, $\varepsilon_j \sim \psi$

This model gave rise to the neg. log. likelihood

$$L(f)_{\text{Probit}} = -\sum_{j=1}^{n} \log \Psi(y_j f(\underline{x}_j))$$

where $\Psi(t) = \int_{-\infty}^{t} \psi(s)\, ds$.

Now our plan is to find a function f s.t. $\text{sign}(f(\underline{x}_{n+1}))$ is a good approx to $\ell(x_{n+1})$.

Assuming that ℓ "varies smoothly" over $\mathcal{X}$ ie, label of two similar pts on $\mathcal{X}$ is the same we may choose to look for an f in an RKHS on $\mathcal{X}$. So, let K be a PDS kernel on $\mathcal{X}$.

eg. for MNIST $\mathcal{X} = \mathbb{R}^{784}$ take

$$K(\underline{x}, \underline{x}') = \exp\left(-\frac{\|\underline{x} - \underline{x}'\|_2^2}{2\sigma^2}\right) \equiv \exp\left(-\frac{d(\underline{x}, \underline{x}')^2}{2\sigma^2}\right)$$

& let $(\mathcal{H}_K, \|\cdot\|_K)$ be the RKHS of K.

Then we can solve

$$f^* = \underset{f \in \mathcal{H}_K}{\operatorname{argmin}} \; L_{\text{Probit}}(f) + \frac{\lambda}{2}\|f\|_K^2$$

A first soln may be obtained via the Representer theorem.

$$f^* = K(\cdot, X)\left(K(X,X) + \gamma^2 I\right)^{-1} \underline{z}^*$$

$$\underline{z}^* = \underset{\underline{z} \in \mathbb{R}^n}{\operatorname{argmin}} \; L_{\text{Probit}}(\underline{z}) + \frac{\lambda}{2}\underline{z}^T\left(K(X,X) + \gamma^2 I\right)^{-1}\underline{z}$$

where $\gamma^2 \geq 0$ is nugget if needed.

A second solution may be obtained using random features. Suppose K is shift invariant $K(\underline{x}, \underline{x}') = K(\|\underline{x} - \underline{x}'\|)$ e.g. the Gaussian kernel above. Then we may approximate f^* by

$$\tilde{f} = \frac{1}{\sqrt{N}} \sum_{j=1}^{N} \tilde{\alpha}_j \varphi_j$$

$$\underline{\tilde{\alpha}} = \underset{\underline{\alpha} \in \mathbb{R}^N}{\operatorname{argmin}} L_{\text{probit}}\left(\frac{1}{\sqrt{N}} \sum_{j=1}^{N} \alpha_j \varphi_j(X)\right) + \frac{\lambda}{2} \|\underline{\alpha}\|_2^2$$

where the φ_j are random features generated via

$$\varphi_j(\underline{x}) = \sqrt{2} \cos(\underline{w}_j^T \underline{x} + b_j)$$

$$\underline{w}_j \overset{iid}{\sim} \Lambda, \quad b_j \overset{iid}{\sim} U[0, 2\pi]$$

In the case of the Gaussian kernel we know that $\Lambda(\underline{w}) = \frac{1}{(2\pi/\sigma^2)^{d/2}} \exp\left(-\frac{\sigma^2 \|\underline{w}\|_2^2}{2}\right)$.

Thus, the $\underline{w}_j \overset{iid}{\sim} N(0, 1/\sigma^2)$

Observe, if Ψ is chosen to be log-concave, then its CDF Ψ is also log-concave. in this case, the probit loss $-\sum_{j=1}^{n} \log \Psi(y_j u_j)$ is convex & it is easy to see that both of our formulations lend to convex problems.

13.2 Intro to Graphical Models

We have seen the wide applicability & overall simplicity of RKHS methods. So far the theory has been very abstract & often stated for wide families of kernels. However, the performance of kernel methods is often dependent on the design of our kernel & its features (as we saw in the RF's for eg.)

In many applications we often encounter data $X = \{\underline{x}_1, \ldots, \underline{x}_n\} \subset \mathbb{R}^d$ for a large d (eg MNIST $d = 784$) but this data often

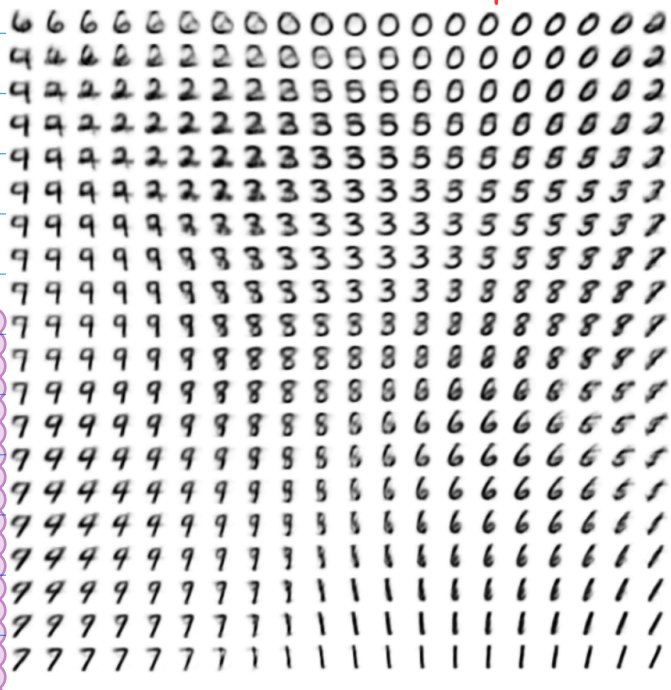
belongs to a low-dim manifold. ie,

$$X = \{x_1, \dots, x_n\} \subset \mathcal{M} \subset \mathbb{R}^d$$

where $\dim(\mathcal{M}) \ll d$.

(for MNIST believed to be 10-15 dim.).

2D-manifold of MNIST learned by a VAE.



This is often referred to as the manifold assumption which underlies the design & analysis of many ML algorithms.

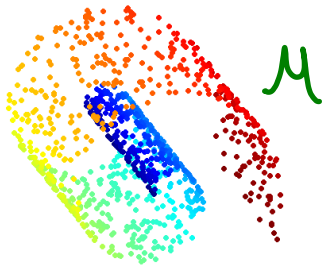
Therefore, it is natural to think of algorithms that can exploit or discover such low-dimensional manifolds/structures. This is essential the subject of study in Manifold Learning.

We will see an example of such a technique in the context of graphical models.

Note: Manifold learning models, at their heart are dimensionality reduction techniques. And, largely fall in the same category as PCA.

Suppose the data set $X \subset M \subset \mathbb{R}^d$

We do not see M but only see pts sampled from it. For simplicity, suppose M is compact & the X are sampled uniformly from M.



(Note we need more assumptions on M for some of the theory but that is besides the point right now).

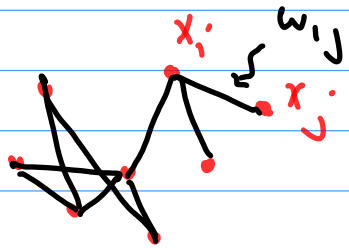
The question is, how do we exploit the structure of the manifold if we cannot see it?

The answer lies at the intersection of graph theory, differential geometry, & diffusion processes.

Given X we can construct a Graph on our data set.

Recall, an undirected weighted graph $G = \{X, W\}$ consists of a set of vertices $X = \{x_1, \ldots, x_n\}$ & associated weighted edges $W_{ij} \geq 0$.

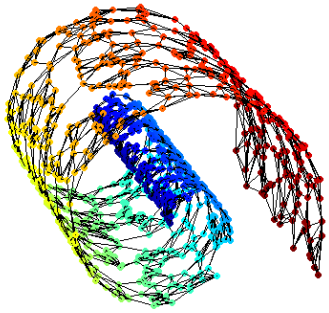
if $w_{ij} = 0$ no edge exists between $\underline{x}_i$ & $\underline{x}_j$
& if $w_{ij} > 0$ then an edge of weight w_{ij} connects $\underline{x}_i$ & $\underline{x}_j$.



Given a point set X we can always construct a

Proximity graph $G(X, W)$

by a simple procedure



- Let X be vertices of G.
- Choose a function

$$\eta : \mathbb{R} \longrightarrow [0, \infty) \quad \text{(Non-increasing)}$$

eg. $\eta(t) = \begin{cases} 1 & \text{if } t \leq r \\ 0 & \text{if } t > r. \end{cases}$ or $\eta(t) = \exp\left(-\frac{t^2}{2r^2}\right)$

- Let $w_{ij} = \eta(\|\underline{x}_i - \underline{x}_j\|)$

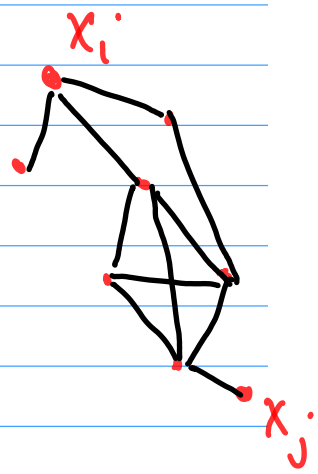
In otherwords, the points in X are connected according to their proximity with respect to their ambient euclidean distance.

To gain intuition why the graph is useful consider doing a discrete random walk on G.

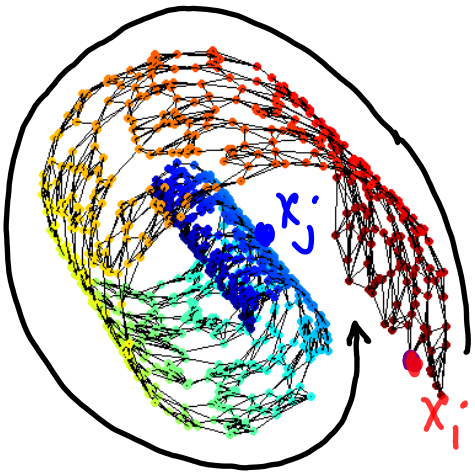
- start from $u(0) = i$
- for $t = 1, 2, 3, \ldots$ do

$u(t) \leftarrow k$ with prob a_k

$$a_k = \frac{w_{u(t),k}}{\sum_{j=1}^{n} w_{u(t),j}}$$



- Compute the expected travel time from x_i to x_j.



This diffusion distance turns out to be a great proxy for the geodesic distance of $\mathcal{M}$. It also allows us to analyze X with respect to the geometry of $\mathcal{M}$.