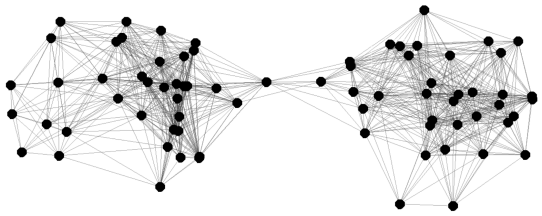
Lecture 18: Wrapping up SSL + Kernel mean embedding

18.1 On Probit for graphical SSL

We ended last lecture with a discussion of the semi-supervised learning problem



Recall. Given $\{\underline{x}_i, y_i \approx \ell(\underline{x}_i)\}_{i=1}^{N}$ we wish to find $\{\ell(\underline{x}_i)\}_{i=N+1}^{M}$ & typically $N \ll M$.

Our proposed solution at the end of last lecture was

$$\underline{f}^* = \underset{\underline{f} \in \mathbb{R}^M}{\operatorname{argmin}} - \sum_{j=1}^{N} \log \Psi(f_j, y_j) + \beta \underline{f}^T C^{-1} \underline{f}$$

(I was missing in Lec 17!)

where Ψ is the CDF of our noise in the probit model & $C \in \mathbb{R}^{M \times M}$ is the matrix

$$C = (L + \tau^2 I)^{-\alpha}$$

where $\beta, \tau^2, \alpha \geq 0$ are hyper-parameters. Let us take a closer look at the matrix C. The claim is that $\underline{f}^T C^{-1} \underline{f}$ here defines "good/meaningful" RKHS norm.

First, suppose L has eig. decomp. $L = U \Sigma U^T$ where $\Sigma = \text{diag}(\underline{\sigma})$, $\underline{\sigma}$ is an increasing seq. of eigenvalues with $\sigma_1 = 0$.

Then we have that $C = U(\Sigma + \tau^2 I)^{-\alpha} U^T$

That is,

$$C = \sum_{j=1}^{M} \frac{1}{(\sigma_j + \tau^2)^\alpha} \underline{u}_j \cdot \underline{u}_j^T$$

where $\lambda_j = \frac{1}{(\sigma_j + \tau^2)^\alpha}$ is a decreasing sequence. Indeed, C is strictly PDS so it defines a kernel & RKHS of functions $f: V \to \mathbb{R}$, ie, $\mathbb{R}^M$.

$$C: V \times V \to \mathbb{R}, \quad C(m,n) = C_{mn}$$

Then the RKHS of C is given by vectors $\underline{f}$ of the form

$$\underline{f} = \sum_{p=1}^{P} \alpha_p C(:, m_p),$$

for a set of indices, $\{m_p\}_{p=1}^P \subset V$.

In otherwords, $\underline{f}$ is a linear comb of the colums of C! we can further write

Further observe that

$$\langle \underline{f}, C(:,\ell) \rangle_{C^{-1}} = \underline{f}^T C^{-1} C(:,\ell)$$

$$= \underline{f}^T \underline{e}_\ell = f_\ell \qquad \text{(Reproducing property!)}$$

(ℓ-unit vec.)

Thus, the RKHS norm $\|\cdot\|_C$ of $\underline{f}$ is given by $\|\underline{f}\|_C = (\underline{f}^T C^{-1} \underline{f})^{1/2}$

Thus, the probit classifier is analgously defined as

$$\underline{f}^* = \underset{\underline{f} \in \mathbb{R}^N}{\operatorname{argmin}} - \sum_{j=1}^{N} \log \Psi(y_j f_j) + \beta \|\underline{f}\|_C^2$$

we can then derive a rep. theorem for this problem & write

(first N col.) (top left NxN block)

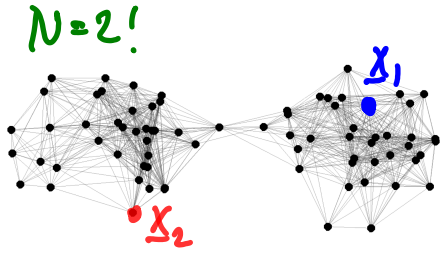
$$\underline{f}^* = C(:, 1:N)\, C(1:N, 1:N)^{-1} \underline{z}^*$$

where $\underline{z}^*$ solves

$$\underline{z}^* = \underset{\underline{z} \in \mathbb{R}^N}{\operatorname{argmin}} - \sum_{j=1}^{N} \log \Psi(y_j z_j) + \beta \underline{z}^T C(1:N, 1:N)^{-1} \underline{z}$$

Observe $\underline{z}^* \in \mathbb{R}^N$ while $\underline{f} \in \mathbb{R}^M$ & assume $N \ll M$ it is much more efficient to solve for $\underline{z}^*$!

Question: where is the geometry?



Consider $L = D - W$ & suppose G has Q connected components $\{G_q\}_{q=1}^Q$. Recall this implies that

$$0 = \sigma_1 = \sigma_2 = \cdots = \sigma_Q < \sigma_{Q+1} < \cdots < \sigma_M$$

& that we may take $\underline{u}_q = \mathbb{1}_{G_q}$ for $q = 1, \ldots, Q$

We then have

$$C = \sum_{q=1}^{Q} \frac{1}{(\sigma_q + \tau^2)^\alpha} \underline{u}_q \underline{u}_q^T + \sum_{j=Q+1}^{M} \frac{1}{(\sigma_q + \tau^2)^\alpha} \underline{u}_j \underline{u}_j^T$$

But $\sigma_q = 0$! so we have

$$C = \tau^{-2\alpha} \sum_{q=1}^{Q} \mathbb{1}_{G_q} \mathbb{1}_{G_q}^T + G(1)$$

Thus, when τ is small, elements of RKHS of C look like the indicator vector $\mathbb{1}_{G_q}$ for $q = 1, \ldots, Q$!

This observation generalizes to weakly connected graphs (see Hoffmann et al (2020) for eg).

18.2 Kernel mean embedding of distributions

* Based on the excellent review article of Muandet et al. "Kernel Mean Embedding of distributions: A Review & Beyond"

Let us think about the idea of the diffusion distance for a moment.

Given X & diff. map $\mathcal{F}_t : X \to \mathbb{R}^m$ (feature map of heat kernel) define

$$D_t(\underline{x}_i, \underline{x}_j) = \| \mathcal{F}_t(\underline{x}_i) - \mathcal{F}_t(\underline{x}_j) \|_2$$

In otherwords, first we embed X in a feature space (here $\mathbb{R}^m$). Second, compute pairwise dist. in that feature space.

This idea of mapping to feature space & then manipulate/analyze data is fundamental to kernel methods.

A very good example is Kernel PCA (see Mika et al "Kernel PCA & denoising in Feature Space".)

What we aim to do now is to use kernels to devise a notion of "distance" between probability distributions.

why are we doing this?
suppose we are given two data sets X, X' both drawn/generated independently. We wish to know if X & X' are drawn from the same distribution? (This is called a two sample hypothesis test)

Another perspective, with diff maps we computed distance of points. Now we want to compute distance between prob. measures.

To achieve this goal we first need to map a prob. measure to our feature space. This is easily done for empirical measures.

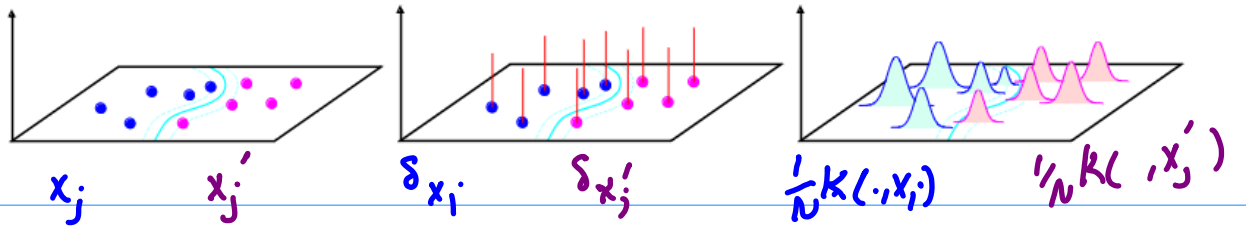
Suppose $\mu^N = \frac{1}{N}\sum_{j=1}^{N} \delta_{x_j}$, where δ_{x_j} are point masses at points $x_j \in \mathcal{X}$ (generic state space) such that $x_j \overset{iid}{\sim} \mu$ (some prob. measure on $\mathcal{X}$).

Now pick a kernel $K: \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ & consider

duality pairing

$$\mu_K^N(x) := \mathbb{E}_{\xi \sim \mu^N} K(x, \xi) = \frac{1}{N} \sum_{j=1}^{N} [\delta_{x_j}, K(x, \cdot)]$$

$$= \frac{1}{N} \sum_{j=1}^{N} K(x, x_j)$$



Now observe that $\mu_K^N \in \mathcal{H}_K$, ie the RKHS of K. This is called the kernel mean embedding of μ^N.

We can naturally extend this idea to prob. measures on $\mathcal{X}$:

Def^n: Let $\mu \in \mathbb{P}(\mathcal{X})$ (Borel prob. measure on $\mathcal{X}$). We then define the kernel mean embedding of μ via K as

$$\mu_K := \mathbb{E}_{\xi \sim \mu} K(\cdot, \xi) = \int K(\cdot, \xi) \mu(d\xi)$$

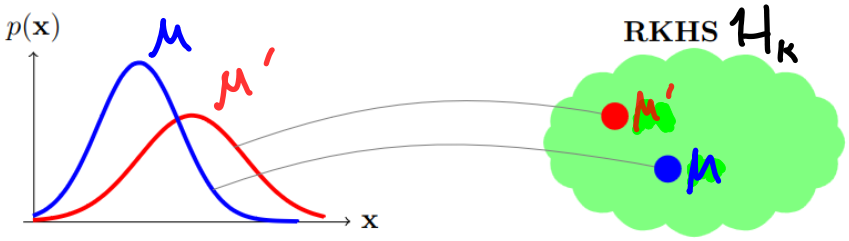
which defines an embedding $\mathbb{P}(\mathcal{X}) \to \mathcal{H}_K$.

Using this embedding we define a notion of "discrepancy/divergence" between prob. measures on $\mathcal{X}$.

Def^n (Maximum Mean Discrepancy (MMD))

Given $\mu_1, \mu_2 \in \mathbb{P}(\mathcal{X})$ & kernel $k: \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ with RKHS $\mathcal{H}_k, \|\cdot\|_k$ we define

$$MMD_k(\mu, \mu') = \| \mu_k - \mu'_k \|_k$$



Why is it called a "discrepancy"?

First of all, it may not satisfy triangle ineq. so it is not generally a metric.

Second, depending on choice of k we may have $MMD(\mu, \mu') = 0$ while $\mu \neq \mu'$ for some pairs of measures! (