Kernel Based Regression

- Generalized notion of predictor variables
- Use ridge on generalized variables
- Kernels useful in extending classical regression tools to non traditional data.
  - pictures - graphs

Goal: Find $\hat{h}: V \longrightarrow \mathbb{R}$ describing how X vary across V.

Def: $K: V \times V \longrightarrow \mathbb{R}$ is a psd kernel iff for all submatrices is (a) symmetric (b) psd

Rmk: Kernels create similarity matrices

- Assume our function comes from the class

$$\mathcal{H}_K = \{ h \in \mathbb{R}^{N_V} : \underbrace{h = I\beta}_{\text{linearity}} \text{ and } \underbrace{\beta^T \Delta^{-1} \beta < \infty}_{\text{penalty}} \}$$

$K = I \Delta I^T$ - build $\mathcal{H}_K$ from eigen decomp of K.

Rmk: Choose K which induces class of functions $\mathcal{H}_K$

Solve this problem via

$$\min_{\beta} \left[\sum_{i \in \text{obs}} C(x_i, (I\beta)_i) + \lambda \beta^T \Delta^{-1} \beta \right]$$

Representer Thm: (Don't need to evaluate K.) $\hat{h}$ takes the form $h = K^{(N_V, n)}_{\text{obs.}} \alpha$

when $K^{(N_V,n)}$ is an evaluation on $V \times U^{obs}$

Reduces the optimization

$$\min_{\alpha} \left[\sum_{i \in U^{obs}} C(x_i, (K^{(n)}\alpha)_i) + \lambda \alpha^T K^{(n)} \alpha \right]$$

when $K^{(n)}$ is $n \times n$

Under squared-error loss

$$\min_{\alpha} \left[(x - K^{(n)}\alpha)^T (x - K^{(n)}\alpha) + \lambda \alpha^T K^{(n)} \alpha \right]$$

Change of variables $M = \mathcal{E}_n \Delta_n^{1/2}$ $\quad \theta = \Delta_n^{1/2} \mathcal{E}_n^T \alpha$ for $K^n = \mathcal{E}_n \Delta_n \mathcal{E}_n^T$

$$\min_{\theta} \left[(x - M\theta)^T (x - M\theta) + \lambda \theta^T \theta \right]$$

$$\hat{\theta}_{RR} = (M^T M + \lambda I)^{-1} M^T x^{obs.}$$

$$\rightarrow \hat{\alpha} = \mathcal{E}_n \Delta_n^{-1/2} \hat{\theta}_{RR}$$

and our final solution is $\boxed{\hat{h} = K^{(N_V, n)} \hat{\alpha}}$

In designing a kernel: - captures similarity among vertices

- p.s.d.

Ex: Laplacian Kernel: $K = L^-$

under this kernel, the penalty β has the form

$$\beta^T \Delta^{-1} \beta = \beta^T I^T I \Delta^{-1} I^T I \beta$$
$$= h^T K^- h$$
$$= h^T L h$$
$$= \sum_{(i,j) \in E} (h_i - h_j)^2$$

So this kernel gives a smooth penalty over the graph.

A large class of kernel $K = \sum_{i=1}^{N_v} r^{-1}(\gamma_i)\, \phi_i \phi_i^T$ based on this