Optimal σ for Maximum Kernel Separation

Chieh Wu

Electrical and Computer Engineering Dept., Northeastern University, Boston, MA

Abstract

Although the Gaussian kernel is the most common kernel choice for kernel methods, its σ value is a hyperparameter that must be tuned for each dataset. This work proposes to set the σ value based on the maximum kernel separation. The source code is made publicly available on `https://github.com/endsley/opt_gaussian_-`.

1 Method

Let $X \in \mathbb{R}^{n \times d}$ be a dataset of n samples with d features and let $Y \in \mathbb{R}^{n \times k}$ be the corresponding one-hot encoded labels where k denotes the number of classes. Given $\kappa_X(\cdot,\cdot)$ and $\kappa_Y(\cdot,\cdot)$ as two kernel functions that applies respectively to X and Y to construct kernel matrices $K_X \in \mathbb{R}^{n \times n}$ and $K_Y \in \mathbb{R}^{n \times n}$. Also let $\mathcal{S}$ and $\bar{\mathcal{S}}$ be sets of all pairs of samples of (x_i, x_j) from a dataset X that belongs to the same and different classes respectively, then the average kernel value for all (x_i, x_j) pairs with the same class is

$$d_{\mathcal{S}} = \frac{1}{|\mathcal{S}|} \sum_{i,j \in \mathcal{S}} e^{-\frac{||x_i - x_j||^2}{2\sigma^2}} \tag{1}$$

and the average kernel value for all (x_i, x_j) pairs between different classes is

$$d_{\bar{\mathcal{S}}} = \frac{1}{|\bar{\mathcal{S}}|} \sum_{i,j \in \bar{\mathcal{S}}} e^{-\frac{||x_i - x_j||^2}{2\sigma^2}}. \tag{2}$$

We propose to find the σ that maximizes the difference between $d_{\mathcal{S}}$ and $d_{\bar{\mathcal{S}}}$ or

$$\max_{\sigma} \quad \frac{1}{|\mathcal{S}|} \sum_{i,j \in \mathcal{S}} e^{-\frac{||x_i - x_j||^2}{2\sigma^2}} - \frac{1}{|\bar{\mathcal{S}}|} \sum_{i,j \in \bar{\mathcal{S}}} e^{-\frac{||x_i - x_j||^2}{2\sigma^2}}. \tag{3}$$

It turns that that is expression can be computed efficiently. Let $g = \frac{1}{|\mathcal{S}|}$ and $\bar{g} = \frac{1}{|\bar{\mathcal{S}}|}$, and let $\mathbf{1}_{n \times n} \in \mathbb{R}^{n \times n}$ be a matrix of 1s, then we can define Q as

$$Q = -gK_Y + \bar{g}(\mathbf{1}_{n \times n} - K_Y). \tag{4}$$

Or Q can be written more compactly as

$$Q = \bar{g}\mathbf{1}_{n \times n} - (g + \bar{g})K_Y. \tag{5}$$

Given Q, Eq. (3) becomes

$$\min_{\sigma} \quad \mathrm{Tr}(K_X Q). \tag{6}$$

Since this is a convex objective, it can be solved with BFGS.

Below, we plot out the average within cluster kernel and the between cluster kernel values as we vary σ. From the plot, we can see that the maximum separation is discovered via BFGS.

Preprint. Under review.

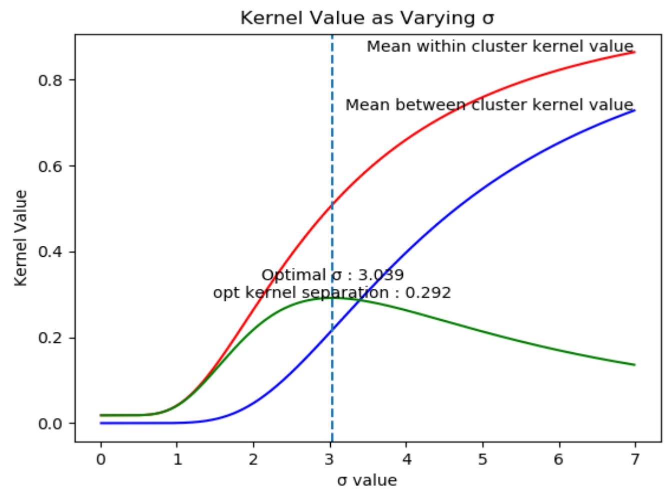

Figure 1: Maximum Kernel separation.

Relation to HSIC. From Eq. (6), we can see that the σ that causes maximum kernel separation is directly related to HSIC. Given that the HSIC objective is normally written as

$$\min_{\sigma} \quad \mathrm{Tr}(K_X H K_Y H), \tag{7}$$

by setting $Q = HK_Y H$, we can see how the two formulations are equivalent. We also notice that the $Q_{i,j}$ element is positive/negative for (x_i, x_j) pairs that are with/between classes respectively. Therefore, the argument for the global optimum should be equivalent for both objectives. Below, we show a figure of HSIC values as we vary σ. Notice how the optimal σ is almost equivalent to the solution from maximum kernel separation.

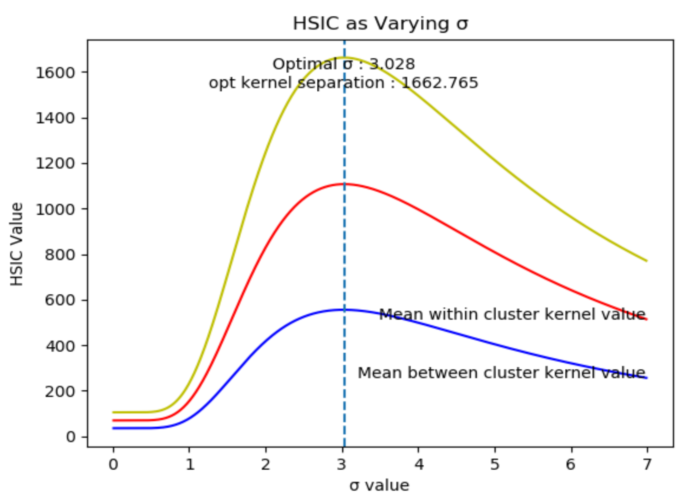

Figure 2: Maximal HSIC.