CS/ECE/ME532 Period 24 Activity

Estimated Time: 40 minutes for P1

1. **Kernel regression.** Kernel regression predicts a value d corresponding to value x as $\hat{d}(x) = \sum_{i=1}^{N} \alpha_i K(x, x^i)$ where the measured data is $(d^i, x^i), i = 1, 2, \ldots N$ and $K(u, v)$ is the kernel function. We will assume Gaussian kernels, $K(u, v) = \exp\left(-(u - v)^2/(2\sigma^2)\right)$. Scripts are provided to help you explore properties of kernel regression with respect to the kernel parameter σ and ridge regression parameter λ.

   a) The ridge regression criterion for a classifier in a high-dimensional feature space may be written as

   $$\min_{\boldsymbol{w}} \sum_{i=1}^{N} \left(d^i - \boldsymbol{\phi}^T(\boldsymbol{x}^i)\boldsymbol{w}\right)^2 + \lambda \boldsymbol{w}^T\boldsymbol{w}$$

   Using $\boldsymbol{w} = \sum_{j=1}^{N} \alpha_j \boldsymbol{\phi}(\boldsymbol{x}^j)$, defining $\boldsymbol{\alpha} = \begin{bmatrix} \alpha_1 & \alpha_2 & \ldots & \alpha_N \end{bmatrix}^T$ and $[\boldsymbol{K}]_{i,j} = \boldsymbol{\phi}^T(\boldsymbol{x}^i)\boldsymbol{\phi}(\boldsymbol{x}^j)$ allows us to rewrite the ridge regression criterion in terms of $\boldsymbol{\alpha}$ as

   $$\min_{\boldsymbol{\alpha}} (\boldsymbol{d} - \boldsymbol{K}\boldsymbol{\alpha})^T(\boldsymbol{d} - \boldsymbol{K}\boldsymbol{\alpha}) + \lambda \boldsymbol{\alpha}^T\boldsymbol{K}\boldsymbol{\alpha}$$

   where $\boldsymbol{d} = \begin{bmatrix} d^1 & d^2 & \cdots & d^N \end{bmatrix}^T$. Show that the $\boldsymbol{\alpha}$ that solves the minimization is

   $$\boldsymbol{\alpha} = (\boldsymbol{K} + \lambda \mathrm{I})^{-1}\boldsymbol{d}$$

   You may do this by either rewriting the cost function in terms of the "perfect" square

   $$\begin{aligned}\min_{\boldsymbol{\alpha}} \left(\boldsymbol{\alpha} - (\boldsymbol{K}^T\boldsymbol{K} + \lambda \mathrm{K})^{-1}\boldsymbol{K}^T\boldsymbol{d}\right)^T (\boldsymbol{K}^T\boldsymbol{K} + \lambda \mathrm{K}) \left(\boldsymbol{\alpha} - (\boldsymbol{K}^T\boldsymbol{K} + \lambda \mathrm{K})^{-1}\boldsymbol{K}^T\boldsymbol{d}\right)& \\ + \boldsymbol{d}^T\boldsymbol{d} - \boldsymbol{d}^T\boldsymbol{K}(\boldsymbol{K}^T\boldsymbol{K} + \lambda \mathrm{K})^{-1}\boldsymbol{K}^T\boldsymbol{d}&\end{aligned}$$

   or by differentiation with respect to $\boldsymbol{\alpha}$. In either case you will need to use the fact that $\boldsymbol{K} = \boldsymbol{K}^T$.

   b) Run the regression script with $\sigma = 0.04$ and $\lambda = 0.01$. Figure 1 displays several of the kernels $K(x, x^i)$. What is the value x^i associated with the kernel having the third peak from the left? What property of the kernel is determined by x^i? What property is determined by σ?

   c) Run the regression script for the following choices of regularization and kernel parameters:

      i. $\lambda = 0.01, \sigma = 0.04$
      ii. $\lambda = 0.01, \sigma = 0.2$
      iii. $\lambda = 0.01, \sigma = 1$

iv. $\lambda = 1, \sigma = 0.04$

v. $\lambda = 1, \sigma = 0.2$

(Note that you need to rerun the entire script each time to ensure the random number generator is reset and you obtain identical data.) You may choose additional cases if it helps you understand the nature of the solution. Discuss how λ and σ affect the characteristics of the kernel regression to the measured data, and support your conclusions with rationale and plots.

d) What principle could you apply to select appropriate values for λ and σ?