CS/ECE/ME532 Activity 22

Estimated Time: 30 minutes for P1, 30 minutes for P2, 10 minutes for P3

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) 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? The index is 46 for the third peak, so x^i is 0.46. The shift on the x axis is what is determined by x^i.

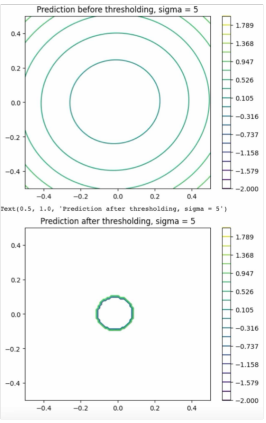
   b) Run the regression script for the following choices of regularization and kernel parameters:

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

   a smaller σ value makes the kernel fit line have more bumps and dips. Each kernel covers less area, so they affect each other's pull on the line less. Smaller kernels will rely on/use fewer training data. Larger σ gives a flatter regression line. Small σ will fit to training data more. σ determines the width of the peak, aka degree of the slope.

   (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. λ pulls the solution closer to the x-axis / 0. The bumps are similar to the same σ value, but lower than the first line.

   c) What principle could you apply to select appropriate values for λ and σ? Use cross validation



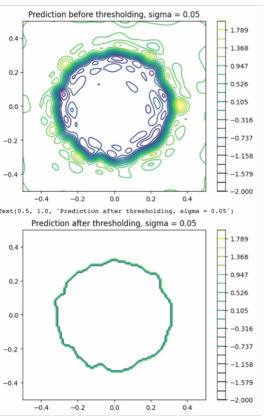
Kernel Classification. The kernel classification script performs classification using the squared error loss using the Gaussian kernel $K(\boldsymbol{u}, \boldsymbol{v}) = \exp\left(-||\boldsymbol{u} - \boldsymbol{v}||_2^2/(2\sigma^2)\right)$. The code is set up to use `N=500` training samples.

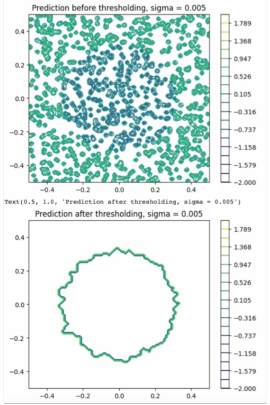
The code creates a contour plot of the predicted class, *before thresholding* (i.e, before applying the sign function).

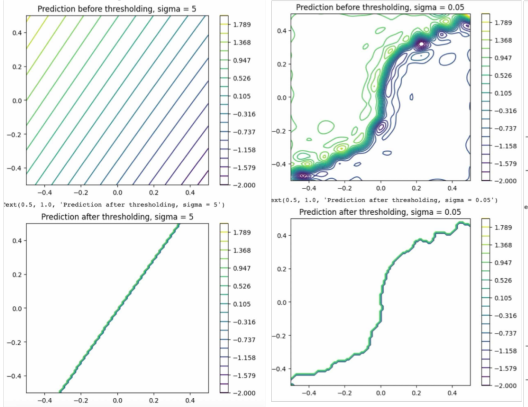
Run the code for the following values of the kernel parameter σ.

a) $\sigma = 5$

b) $\sigma = 0.05$







very small σ will overfit to the data. The prediction before thresholding for small σ has a contour around each individual data point. Then after thresholding, the prediction boundary is bumpy and jagged, so it tries to go around each point instead of just the general pattern.

c) $\sigma = 0.005$

Use the results to discuss the impact of the kernel parameter σ. Is there a downside to choosing a very small value for σ? Run additional values for σ if needed.

On the other hand, σ that is too large may underfit the training data and over generalize.

3. SVM. You use a kernel-based support vector machine for binary classification with labels $d^i = \{+1, -1\}$. Given training features and labels $(\boldsymbol{x}^i, d^i), i = 1, 2, \ldots, N$ you use a kernal $K(\boldsymbol{u}, \boldsymbol{v})$ and design the classifier weights $\boldsymbol{\alpha}$ as

$$\hat{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha}} \sum_{i=1}^{N} \left(1 - d^i \sum_{j=1}^{N} \alpha_j K(\boldsymbol{x}^i, \boldsymbol{x}^j)\right)_+ + \lambda \sum_{i=1}^{N} \sum_{j=1}^{n} \alpha_i \alpha_j K(\boldsymbol{x}^i, \boldsymbol{x}^j)$$

a) Assume the optimization problem has been solved to obtain the weights $\boldsymbol{\alpha}$. Express the classification procedure for a measured feature $\boldsymbol{x}$. $\hat{d} = \text{sign}\left(\sum_i \alpha_i K(x, x^i)\right)$

sign of a weighted sum of kernels

b) Suppose $N = 1000$ and $\alpha_i = 0, i = 1, 2, \ldots, 99, 102, 103, \ldots, 1000$. Identify the support vectors and write the classification procedure in terms of the support vectors.

Support vectors are the only features with a non zero weight, so x^{100} and x^{101} are support vectors.

$$\hat{d} = \text{sign}\left(\alpha_{100} K(x, x^{100}) + \alpha_{101} K(x, x^{101})\right)$$

1b).