$[x_{i_1}^2 \;\; x_{i_2}^2 \;\; \sqrt{2}x_{i_1}x_{i_2} \;\; \sqrt{2}x_{i_1} \;\; \sqrt{2}x_{i_2} \;\; 1]\begin{bmatrix} x_{j_1}^2 \\ x_{j_2}^2 \\ \sqrt{2}x_{j_1}x_{j_2} \\ \sqrt{2}x_{j_1} \\ \sqrt{2}x_{j_2} \\ 1 \end{bmatrix} = x_{i_1}^2x_{j_1}^2 + x_{i_2}^2x_{j_2}^2 + 2x_{i_1}x_{i_2}x_{j_1}x_{j_2} + 2x_{i_1}x_{j_1} + 2x_{i_2}x_{j_2} + 1$

$(x_{i_1}x_{j_1}+x_{i_2}x_{j_2}+1)(x_{i_1}x_{j_1}+x_{i_2}x_{j_2}+1)$
$= x_{i_1}^2x_{j_1}^2 + x_{i_1}x_{i_2}x_{j_1}x_{j_2} + x_{i_1}x_{j_1} + x_{i_2}x_{j_2}x_{i_1}x_{j_1} + x_{i_2}^2x_{j_2}^2 + x_{i_2}x_{j_2} + x_{i_1}x_{j_1} + x_{i_2}x_{j_2} + 1$
$= x_{i_1}^2x_{j_1}^2 + x_{i_2}^2x_{j_2}^2 + 2x_{i_1}x_{i_2}x_{j_1}x_{j_2} + 2x_{i_1}x_{j_1} + 2x_{i_2}x_{j_2} + 1$

CS/ECE/ME532 Activity 21

Estimated Time: 20 minutes for P1, 20 minutes for P2, 10 minutes for P3, 15 minutes for P4.

1. Consider performing regression using all quadratic and lower order functions of a 2-dimensional observation $\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$

$$\hat{y} = x_1^2 w_1 + x_2^2 w_2 + \sqrt{2}x_1x_2w_3 + \sqrt{2}x_1w_4 + \sqrt{2}x_2w_5 + w_6$$

a) Show that $\hat{y} = \boldsymbol{\phi}^T(\boldsymbol{x})\boldsymbol{w}$ and find $\boldsymbol{\phi}, \boldsymbol{w}$.

$\phi = [x_1^2 \;\; x_2^2 \;\; \sqrt{2}x_1x_2 \;\; \sqrt{2}x_1 \;\; \sqrt{2}x_2 \;\; 1]^T$
$w = [w_1 \;\; w_2 \;\; w_3 \;\; w_4 \;\; w_5 \;\; w_6]^T$

b) Show that the "kernel" $\boldsymbol{\phi}^T(\boldsymbol{x}_i)\boldsymbol{\phi}(\boldsymbol{x}_j)$ is identical to $(\boldsymbol{x}_i^T\boldsymbol{x}_j + 1)^2$.

$\hookrightarrow [x_{i_1} \; x_{i_2}]\begin{bmatrix} x_{j_1} \\ x_{j_2} \end{bmatrix} = x_{i_1}x_{j_1} + x_{i_2}x_{j_2} + 1$

c) The number of multiplications may be used as a crude measure of computational complexity. Compare the number of multiplications required to compute $\boldsymbol{\phi}^T(\boldsymbol{x}_i)\boldsymbol{\phi}(\boldsymbol{x}_j)$ (ignoring the $\sqrt{2}$ terms) to that required to compute $(\boldsymbol{x}_i^T\boldsymbol{x}_j + 1)^2$.

$\hookrightarrow$ 6 multiplications

$\hookrightarrow$ 3 multiplications, 2 for $x_i^Tx_j$, 1 for square

2. You are given N observations $y_i, \boldsymbol{x}_i, i = 1, 2, \ldots, N$ and solve the ridge-regression problem

$$\arg\min_{\boldsymbol{w}} ||\boldsymbol{y} - \boldsymbol{\Phi w}||_2^2 + \lambda||\boldsymbol{w}||_2^2$$

where $\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}$ and $\boldsymbol{\Phi} = \begin{bmatrix} \boldsymbol{\phi}^T(\boldsymbol{x}_1) \\ \boldsymbol{\phi}^T(\boldsymbol{x}_2) \\ \vdots \\ \boldsymbol{\phi}^T(\boldsymbol{x}_N) \end{bmatrix}$. You know the solution may be expressed in standard form as

$$\hat{\boldsymbol{w}} = (\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \lambda\boldsymbol{I})^{-1}\boldsymbol{\Phi}^T\boldsymbol{y}$$

$\Phi^T\Phi\Phi^T + \lambda\Phi^T$

$\Phi^T(\Phi\Phi^T + \lambda I) = (\Phi^T\Phi + \lambda I)\Phi^T$

a) Factor $\boldsymbol{\Phi}^T$ from the left and the right of $\boldsymbol{\Phi}^T\boldsymbol{\Phi}\boldsymbol{\Phi}^T + \lambda\boldsymbol{\Phi}^T$ to show that

$$(\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \lambda\boldsymbol{I})^{-1}\boldsymbol{\Phi}^T = \boldsymbol{\Phi}^T(\boldsymbol{\Phi}\boldsymbol{\Phi}^T + \lambda\boldsymbol{I})^{-1}$$

$\Phi^T(\Phi\Phi^T + \lambda I)^{-1} = (\Phi^T\Phi + \lambda I)^{-1}\Phi^T$

$\Phi^T = \Phi^T$

Hint: we did this a previous activity and you used the result in the breast cancer classification assignment.

b) Use the result of the previous part to show that

$$\hat{\boldsymbol{w}} = \boldsymbol{\Phi}^T(\boldsymbol{\Phi}\boldsymbol{\Phi}^T + \lambda\boldsymbol{I})^{-1}\boldsymbol{y}$$

$\hat{w} = \underline{(\Phi^T\Phi + \lambda I)^{-1}}\,\Phi^T y$

$(\Phi^T\Phi + \lambda I)^{-1}\Phi^T = \underline{\Phi^T(\Phi\Phi^T + \lambda I)^{-1}}$

$\hat{w} = \Phi^T(\Phi\Phi^T + \lambda I)^{-1} y$

c) Let the kernel matrix $\boldsymbol{K} = \boldsymbol{\Phi}\boldsymbol{\Phi}^T$. Express the i,j element of $\boldsymbol{K}$, $[\boldsymbol{K}]_{i,j}$ using $\boldsymbol{\phi}(\boldsymbol{x})$.

$[K]_{i,j} = \phi^T(x_i)\phi(x_j)$

$[K]_{i,j} = (x_i^T x_j + 1)^2$

d) Assume $\boldsymbol{\phi}(\boldsymbol{x})$ is defined as in Problem 1 and find $[\boldsymbol{K}]_{i,j}$ as a function of $\boldsymbol{x}_i^T\boldsymbol{x}_j$.

e) Recall from Problem 1 that $\hat{y}(\boldsymbol{x}) = \boldsymbol{\phi}^T(\boldsymbol{x})\hat{\boldsymbol{w}}$. Thus, $\hat{y}(\boldsymbol{x}) = \boldsymbol{\phi}^T(\boldsymbol{x})\boldsymbol{\Phi}^T(\boldsymbol{\Phi}\boldsymbol{\Phi}^T + \lambda \boldsymbol{I})^{-1}\boldsymbol{y}$. Show that

$$\hat{y}(\boldsymbol{x}) = \sum_{j=1}^{N} K(\boldsymbol{x}, \boldsymbol{x}_j)\alpha_j \qquad \alpha = (\Phi\Phi^T + \lambda I)^{-1} y$$

where $K(\boldsymbol{x}, \boldsymbol{x}_j) = (\boldsymbol{x}^T\boldsymbol{x}_j + 1)^2$. $\phi^T(x)\phi(x_j)$

$$\hat{y}(x) = \sum_{j=1}^{N} (x^T x_j + 1)^2 \alpha_j = \sum_{j=1}^{N} \overbrace{(x^T x_j + 1)^2}^{\phi^T(x)\phi(x_j)} (\Phi\Phi^T + \lambda I)^{-1} y = \phi^T(x)\,\Phi^T (\Phi\Phi^T + \lambda I)^{-1} y$$

3. Suppose $\boldsymbol{\phi}(\boldsymbol{x}) = \boldsymbol{x}$. Use the results of the previous problem.

a) Find the expression for the corresponding kernel $K(\boldsymbol{x}, \boldsymbol{x}_j)$. $= x^T x_j$

b) Express $\hat{y}(\boldsymbol{x})$ in terms of α_j and your expression for $K(\boldsymbol{x}, \boldsymbol{x}_j)$. How does each training sample influence the prediction $\hat{y}(\boldsymbol{x})$ at some new value $\boldsymbol{x}$?

$\hat{y}(x) = \sum_{j=1}^{N} \alpha_j (x^T x_j)$ x_j is used as part of each multiplication that sums to $\hat{y}(x)$, so each training sample is used in the prediction.

4. The results we developed in this exercise so far show that regression can be expressed entirely in terms of the kernel function $K(\boldsymbol{x}, \boldsymbol{x}_j)$:

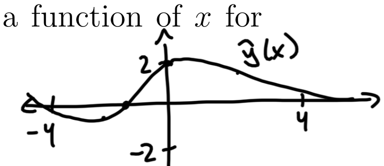
$$\hat{y}(\boldsymbol{x}) = \sum_{j=1}^{n} K(\boldsymbol{x}, \boldsymbol{x}_j)\alpha_j$$

where α_j is a function of the kernel matrix $\boldsymbol{K}$, regularization parameter λ, and data $\boldsymbol{y}$. This form allows us to perform regression when the high dimensional feature vector $\boldsymbol{\phi}(\boldsymbol{x})$ is not easily defined, but $K(\boldsymbol{x}, \boldsymbol{x}_j) = \boldsymbol{\phi}^T(\boldsymbol{x})\boldsymbol{\phi}(\boldsymbol{x}_j)$ is easily defined. One such case is the Gaussian kernel,

$$K(\boldsymbol{x}, \boldsymbol{x}_j) = \exp\left\{-\frac{||\boldsymbol{x} - \boldsymbol{x}_j||_2^2}{2\sigma}\right\}$$

For simplicity this problem assumes $\boldsymbol{x}$ is one dimensional, that is $\hat{y}(x)$ describes a graph of a function of one variable.

a) Suppose $x_1 = -2, x_2 = 0$, and $x_3 = 2$. Sketch $K(x, x_j)$ as a function of x for $j = 1, 2, 3$ assuming $\sigma = 1$.



b) Now sketch $\hat{y}(x)$ assuming $\alpha_1 = -1, \alpha_2 = 2$, and $\alpha_3 = 1$.

c) Fill in the blanks. The expression $\hat{y}(x) = \sum_{j=1}^{n} K(x, x_j)\alpha_j$ interpolates a value y corresponding to x as a <u>weighted</u> sum of <u>n</u> functions centered on the <u>y-axis</u>.

a).