Homework #3

CSE 446/546: Machine Learning
Zuo Wang

Question 1

(a.)

We should decrease σ because decreasing the bandwith allows us to have a more localized and complex model. Mathmatically speaking, decreasing σ makes the kernel value smaller (exponential of a larger negative value) therefore the penalty term is smaller, resulting in a more complex model.

(b.)

False. Most neural networks actually have non-convex loss functions.

(c.)

False. If we initialize all weights to zero when training a deep neural network, the gradient of all the weights will be the same during backpropagation, resulting in all neurons learning the same features and updating their weights identically.

(d.)

True. Linear activation functions would result in a network that can only learn linear decision boundaries.

(e.)

True. The backward pass step requires computing the gradients of the loss with respect to the parameters, which involves matrix multiplications, and derivating activation function. And that is more time-consuming compared to the forward pass step.

(f.)

False. While neural network is consider one of the best models in most cases, it is not always the best choice for every circumstance. Generally speaking, neural networks are more complex, require a longer runtime, are better for larger dataset, and tend to overfit. In some cases, simple models with regularization might just perform better.

Question 2

(a.)

(b.)

From the support vectors and their corresponding labels, we know that:

$$\begin{bmatrix}1\\2\end{bmatrix} w + b = -1,\ \begin{bmatrix}0\\0.5\end{bmatrix} w + b = 1,\ \begin{bmatrix}2\\1\end{bmatrix} w + b = 1$$

Looking at the last two equations, we can construct the matrix:

$$\begin{aligned} w &= \left[\begin{array}{cc|c} 0 & 0.5 & 1-b \\ 2 & 1 & 1-b \end{array}\right] \\ &= \left[\begin{array}{cc|c} 2 & 1 & (1-b) \\ 0 & 1 & 2(1-b) \end{array}\right] \\ &= \left[\begin{array}{cc|c} 1 & 0 & -\frac{1}{2}(1-b) \\ 0 & 1 & 2(1-b) \end{array}\right] \\ &= \begin{bmatrix} -\frac{1}{2}(1-b) & 2(1-b) \end{bmatrix} \end{aligned}$$

And we plug w back into $\begin{bmatrix}1\\2\end{bmatrix} w + b = -1$:

$$\begin{aligned} \begin{bmatrix}1\\2\end{bmatrix} \cdot \begin{bmatrix} -\frac{1}{2}(1-b) & 2(1-b) \end{bmatrix} + b &= -1 \\ -\frac{1}{2}(1-b) + 4(1-b) + b &= -1 \\ \frac{7}{2}(1-b) + b &= -1 \\ -\frac{5}{2}b &= -\frac{9}{2} \\ b &= \frac{9}{5} \end{aligned}$$

$$\boxed{b = \frac{9}{5} \quad w = \begin{bmatrix} \frac{2}{5} & -\frac{8}{5} \end{bmatrix}}$$

(c.)

Question 3

$$\begin{aligned}
\phi(x) \cdot \phi(x') &= \sum_{i=0}^{\infty} \left(\frac{1}{\sqrt{i!}} e^{\frac{-x^2}{2}} x^i \right) \cdot \left(\frac{1}{\sqrt{i!}} e^{\frac{-x'^2}{2}} (x')^i \right) \\
&= \sum_{i=0}^{\infty} \frac{1}{i!} e^{\frac{-(x^2+x'^2)}{2}} x^i (x')^i \\
&= e^{\frac{-(x^2+x'^2)}{2}} \sum_{i=0}^{\infty} \frac{1}{i!} x^i (x')^i
\end{aligned}$$

Because the Taylor expansion of $e^z = \sum_{n=0}^{\infty} \frac{1}{n!} z^n$, the summation above just becomes $e^{xx'}$, therefore:

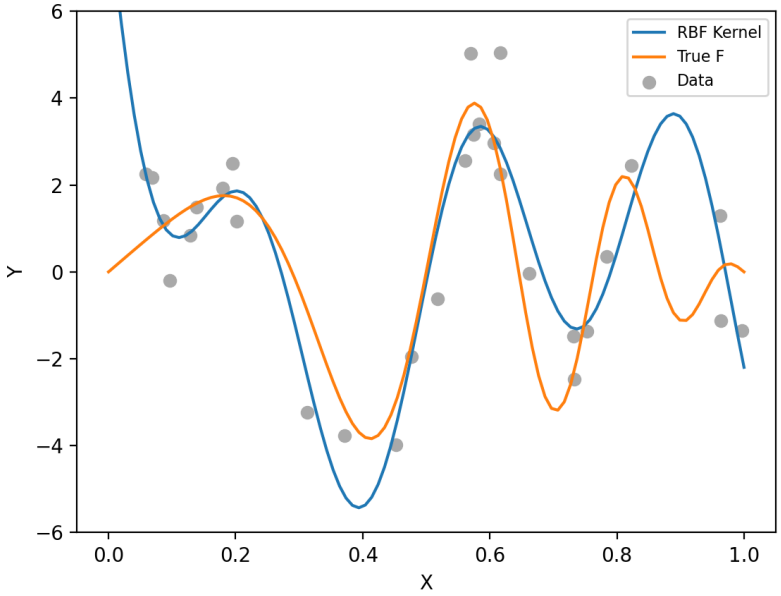
$$\begin{aligned}
\phi(x) \cdot \phi(x') &= e^{\frac{-(x^2+x'^2)}{2}} \cdot e^{xx'} \\
&= e^{\frac{-x^2-x'^2+2xx'}{2}} \\
&= e^{\frac{-(x-x')^2}{2}}
\end{aligned}$$

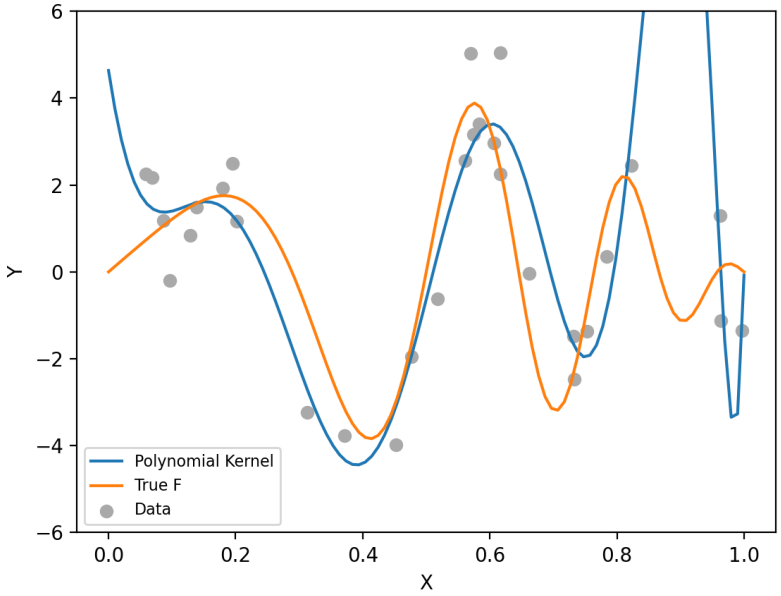
Question 4

(a.)

For the RBF kernel, $\lambda_{optimal} = 10^{-3},\ \gamma_{optimal} = 10.5416$
For the polynomial kernel, $\lambda_{optimal} = 10^{-3},\ d_{optimal} = 20.5102$

(b.)





Question 5

Question 6