1) Let $< \mathbf{x}^T = (f_1, f_2), t >$ represent the data point $\mathbf{x}$ located at feature coordinates (f_1, f_2) with the corresponding label $t \in \{-1, 1\}$. Assume the following data is available $< \mathbf{x}_1^T = (0,0), -1 >, < \mathbf{x}_2^T = (0,1), 1 >, < \mathbf{x}_3^T = (1,0), 1 >$. We want to use a linear (non-kernel) Support Vector Machine classifier to specify the discriminant function in the form of $y(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b$, where $\mathbf{w} = (w_1, w_2)^T$. Let the a_1, a_2, a_3 denote Lagrangian multipliers for slackness constraints of $\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{x}_3$, respectively.

1. Plot the data points and derive the decision boundary by inspecting the data. What can be said about the Lagrange multipliers?
2. Write up the primal Lagrangian, apply the optimality condition and express the $\mathbf{w}$ in terms of data points. Now substitute $\mathbf{w}$ in the discriminant function and derive its kernel statement. Explain why these models are said to be 'sparse'?

2) Suppose that $\mathbf{x}$ and $\mathbf{y}$ are 2 dimensional feature vectors, and $k_1(\mathbf{x}, \mathbf{y})$ and $k_2(\mathbf{x}, \mathbf{y})$ are both valid kernels. Show that the following kernels are valid:

1. $k_1(\mathbf{x}, \mathbf{y}) + k_2(\mathbf{x}, \mathbf{y})$
2. $\mathbf{x}^T \left(\lambda_i \mathbf{e}_i \mathbf{e}_i^T \right) \mathbf{y}$, where $\mathbf{e}_i = (e_{i1}, e_{i2})^T$ and $\lambda_i \geq 0$.
3. $\mathbf{x}^T \mathbf{A} \mathbf{y}$, where $\mathbf{A} = \sum_i \lambda_i \mathbf{e}_i \mathbf{e}_i^T$ and $\lambda_i \geq 0$ (symmetric semi-positive definite matrix).
4. $(\mathbf{x}^T \mathbf{y})^n$, where n is a positive integer.
5. $a_1(\mathbf{x}^T \mathbf{y}) + \cdots + a_n(\mathbf{x}^T \mathbf{y})^n$, where a_n's are positive real number.
6. $e^{\mathbf{x}^T \mathbf{y}}$