1. *Kernel Support Vector Machines.* In class (and in one of your homeworkj questions), we saw that if we wished to fit a classifier based on a (potentially nonlinear) feature mapping $\varphi : \mathbf{R}^n \to \mathbf{R}^N$, the classification rule (i.e., prediction on an input x) is

$$\hat{y} = \mathbf{sign}(\theta^T \varphi(x)).$$

Let $K(x, z) = \varphi(x)^T \varphi(z)$ be the *kernel function* associated with feature mapping φ. Given a collection of pairs $(x_i, y_i) \in \mathbf{R}^n \times \{-1, 1\}$, we find a classifier θ by solving

$$\text{minimize} \quad \sum_{i=1}^m f(y_i \varphi(x_i)^T \theta) + \frac{\lambda}{2} \|\theta\|_2^2, \tag{1}$$

where $f : \mathbf{R} \to \mathbf{R}$ is a convex, non-increasing function and $\lambda \geq 0$ is a regularization parameter. Equivalent to problem (1) is to solve the dual problem

$$\text{maximize} \quad -\sum_{i=1}^m f^*(\alpha_i) - \frac{1}{2\lambda} \alpha^T \, \mathbf{diag}(y) G \, \mathbf{diag}(y) \alpha \tag{2}$$

with variable $\alpha \in \mathbf{R}^m$, where $G \in \mathbf{S}^m$ is the *Gram matrix*, whose entries are $G_{ij} = K(x_i, x_j)$. To recover the optimal $\theta^\star$ for problem (1) given an optimal dual variable $\alpha^\star$, we may set

$$\theta^\star = -\frac{1}{\lambda} \sum_{i=1}^m y_i \varphi(x_i) \alpha_i^\star = \sum_{i=1}^m \varphi(x_i) \nu_i^\star,$$

where $\nu^\star = -\frac{1}{\lambda} \mathbf{diag}(y) \alpha^\star$.

Now, given a $\theta \in \mathbf{R}^N$ taking the form $\theta = \sum_{i=1}^m \nu_i \varphi(x_i)$, we can define the prediction function $p_\theta : \mathbf{R}^n \to \mathbf{R}$ by

$$p_\theta(x) = \theta^T \varphi(x) = \sum_{i=1}^m \nu_i \varphi(x_i)^T \varphi(x) = \sum_{i=1}^m K(x_i, x) \nu_i,$$

where K is the kernel function for φ. To make a prediction on an input x, we simply take the sign $\hat{y} = \mathbf{sign}(p_\theta(x))$, and we view the magnitude of $p_\theta(x)$ as the "confidence" the classifier gives to its prediction.

Using the kernel $K(x, z) = (1 + x^T z)^6$ and objective $f(t) = (1 - t)_+ = \max\{0, 1 - t\}$, implement the dual problem (2) for the problem data in `kernel_svm_data.*`. Solve the resulting problem with regularization multipliers $\lambda = 10^{-3}, 10^{-2}, 1, 10$. For each λ, plot a contour plots of the resulting prediction function $p_\theta(x)$ as a function of $x \in [-1.5, 1.5]^2 \subset \mathbf{R}^2$. What does increasing/decreasing the regularization λ do?

A few notes on implementation: numerical instability issues mean that in your code, the Gram matrix G may be slightly indefinite; in this case, you should replace G with $G + \epsilon I$ for $\epsilon = 10^{-6}$ or some other small constant. Different solvers may experience instability, so if one solver does not work (e.g. SCS) try another. In our solution, we also plotted the datapoints in a scatterplot to give some intuition. Be sure to include your code and the four contour plots. *Hint.* You may use that $f^*(s) = s$ for $s \in [-1, 0]$ and $f^*(s) = +\infty$ otherwise.