Problems 2

1. Ridge regression. (4 pts)

We encourage the parameters to be small, thus resulting in reducing the variance, by using a zero-mean Gaussian prior

$$p(\theta) = \prod_j N(\theta_j|0, \tau^2)$$

where τ controls the strength of the prior. The corresponding maximum-a-posteriori (MAP) estimation problem becomes

$$\operatorname*{argmax}_\theta p(\theta|D) = \operatorname*{argmax}_\theta \frac{p(D|\theta)p(\theta)}{p(D)} = \operatorname*{argmax}_\theta p(D|\theta)p(\theta),$$

where $D = \{(x^{(i)}, y^{(i)}), \ldots, (x^{(m)}, y^{(m)})\}$ and $p(D|\theta) = \prod_{i=1}^m p(y^{(i)}|x^{(i)}; \theta) = \prod_{i=1}^m N(\theta^\top x^{(i)}, \sigma^2)$.

(a) Show that it is equivalent to minimizing the following

$$\operatorname*{argmin}_\theta \sum_{i=1}^m (y^{(i)} - \theta^\top x^{(i)})^2 + \lambda ||\theta||_2^2,$$

where $\lambda \in \mathbb{R}, \tau \in \mathbb{R}$, and $\lambda = \frac{\sigma^2}{\tau^2}$. (1 pts)

(b) Show that the solution is given by

$$\hat{\theta}_{\text{ridge}} = (X^\top X + \lambda I)^{-1} X^\top y,$$

where $X \in \mathbb{R}^{m \times d}, \theta \in \mathbb{R}^d$ and $y \in \mathbb{R}^m$. (1 pts)

(c) Show that ordinary least squares regression on an augmented data can achieve the ridge regression estimates. We augment the data matrix X with n additional rows $\sqrt{\lambda} I$ and augment y with n zeros, so the augmented data is $\hat{X} \in \mathbb{R}^{(m+n) \times d}, \hat{y} \in \mathbb{R}^{m+n}$. (1 pts)

(d) We want to use kernels to the ridge regression in order to use high-dimensional features mapping $\phi(x^{(i)})$. Given the new input x_{new}, the prediction can be done by $\theta^\top \phi(x_{\text{new}})$. Show that we can make predictions without ever explicitly compute $\phi(x^{(i)})$. (1 pts)

2. [Coding problem] SVM (Support Vector Machine). You are given 3 datasets (decompress the attached dataset file `hw2_data.zip`). `hw2_2_X_(1-3).csv` is comma separated files, and 2 dimentional data and `hw2_2_y_(1-3).csv` is the corresponding labels, either $+1$ or -1. You are going to apply SVM to these dataset and you can use any existing SVM libraries, e.g. sklearn. (6 pts)

(a) Plot `hw2_2_X_1.csv` and `hw2_2_y_1.csv` (use different color for different classes. Apply linear SVM and draw decision boundary (solid line) and margins (dashed line) in the plot. Provide all support vectors and circle them in the plot. (2 pts)

(b) Plot `hw2_2_X_2.csv` and `hw2_2_y_2.csv` (use different color for different classes. Apply kernel SVM (using RBF kernel) and draw decision boundary (solid line) and margins (dashed line) in the plot. Provide all support vectors and circle them in the plot. (2 pts)

(c) Plot `hw2_2_X_3.csv` and `hw2_2_y_3.csv` (use different color for different classes. Apply linear SVM with soft margin and draw decision boundary (solid line) and margins (dashed line) in the plot. Provide all support vectors and circle them in the plot. Try at least 3 different hyperparameters C. (2 pts)