CSED515 Machine Learning - Assignment 3

Dongwoo Kim

Due date: 23:59pm, April 19th, 2023

Remark

Assignment Submission. All students must submit their homework via PLMS. Submit your answer on PLMS in a single PDF file named with `your_student_id.pdf`. You can scan your hand-written answers or write your answer with a tablet (If you are ambitious enough, try to use LaTeX to write your answers). Any violation in the submission format may bring 10% penalty.

Late Homework Policy. We do not allow late submission. If you have any question regarding this, please send an email to the lecturer.

Honor Code. We strongly encourage students to form study groups. Students may discuss and work on homework problems in groups. However, each student must write down their solutions independently, i.e., each student must understand the solution well enough in order to reconstruct it by him/herself. Using code or solutions obtained from the web (GitHub/Google/ etc.) is considered an honor code violation. We check all the submissions for plagiarism. We take the honor code very seriously and expect students to do the same.

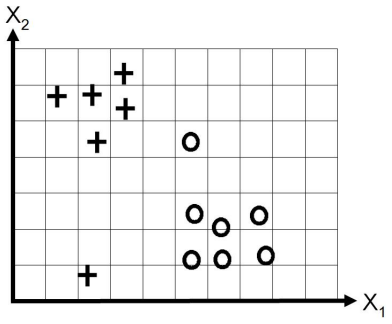
Importance Notice Note that the first assignment aims to identify your background on linear algebra, probability and statistics, and vector calculus. Show your work with your answers. You will get zero score if there is no proper explanation. You can use any external resources to solve the problems, however, please make sure that you have understood the details of what you have written. If you think you are not ready to solve these questions, please consider to take preliminary classes before taking this class. The list of preliminary courses are provided during the second lecture.

1. **Regularizing Separate Terms in 2D Logistic Regression** [20pt].

   (a) Consider the data in the following figure, where we fit the model

   $$p(y = 1 \mid \boldsymbol{x}, \boldsymbol{w}) = \sigma\left(w_0 + w_1 x_1 + w_2 x_2\right).$$

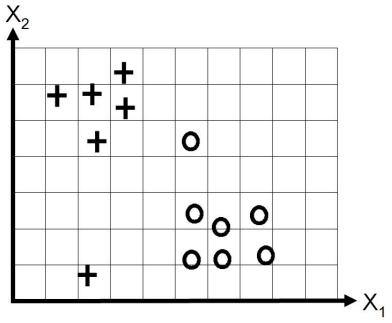
   Suppose we fit the model by maximum likelihood, i.e., we minimize $J(\boldsymbol{w}) = -\ell\left(w, \mathcal{D}_{\text{train}}\right)$ where $\ell\left(w, \mathcal{D}_{\text{train}}\right)$ is the log likelihood on the training set. Sketch a possible decision boundary corresponding to $\hat{\boldsymbol{w}}$ (a rough sketch is enough). How many classification errors does your method make on the training set?



   (b) Now suppose we regularize only the w_0 parameter, i.e., we minimize

   $$J_0(\boldsymbol{w}) = -\ell\left(\boldsymbol{w}, \mathcal{D}_{\text{train}}\right) + \lambda w_0^2$$

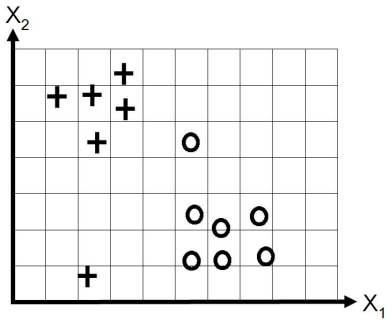
   Suppose λ is a very large number, so we regularize w_0 all the way to 0 , but all other parameters are unregularized. Sketch a possible decision boundary. How many classification errors does your method make on the training set?



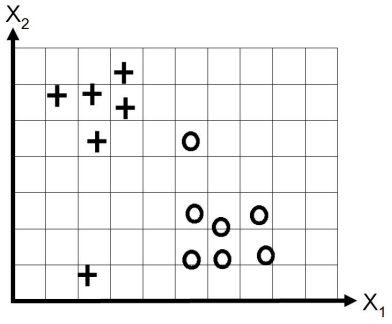
(c) Now suppose we heavily regularize only the w_1 parameter, i.e., we minimize

$$J_1(\boldsymbol{w}) = -\ell\left(\boldsymbol{w}, \mathcal{D}_{\text{train}}\right) + \lambda w_1^2$$

Sketch a possible decision boundary. How many classification errors does your method make on the training set?



(d) Now suppose we heavily regularize only the w_2 parameter. Sketch a possible decision boundary. How many classification errors does your method make on the training set?



2. **SVM** [20pt]. Consider a dataset with 2 points in $\mathbb{R}$ ($x_1 = 0, y_1 = -1$) and ($x_2 = \sqrt{2}, y_2 = 1$). Consider mapping each point to 3d using the feature vector $\phi(x) = [1, \sqrt{2}x, x^2]^\top$. The max margin classifier has the form

$$\begin{aligned} \min ||\boldsymbol{w}||^2 \text{ s.t.} \\ y_1(\boldsymbol{w}^\top \phi(x_1) + b) \geq 1 \\ y_2(\boldsymbol{w}^\top \phi(x_2) + b) \geq 1 \end{aligned}$$

(a) Write down a vector that is parallel to the optimal vector $\boldsymbol{w}$.

(b) What is the value of the margin that is achieved by this $\boldsymbol{w}$?

(c) Solve for $\boldsymbol{w}$ using the fact the margin is equal to $1/||\boldsymbol{w}||$.

(d) Solve for b using your value for $\boldsymbol{w}$.

3. **Kernel and Corresponding Features** [15pt]. We learned that SVM can be used to classify linearly inseparable data by transforming it to a higher dimensional space with a kernel $K(x, z) = \phi(x)^T \phi(z)$, where $\phi(x)$ is a feature mapping. Let K_1 and K_2 be $R^n \times R^n$ kernels, K_3 be a $R^d \times R^d$ kernel and $a, b, c \in R^+$ be a positive constant. $\phi_1 : R^n \to R^d, \phi_2 : R^n \to R^d$, and $\phi_3 : R^d \to R^d$ are feature mappings of K_1, K_2 and K_3 respectively. Explain how to use ϕ_1 and ϕ_2 to obtain the following kernels (You need to show the feature mapping of each kernel K using ϕ_1, ϕ_2, ϕ_3).

   (a) $K(x, z) = cK_1(x, z)$

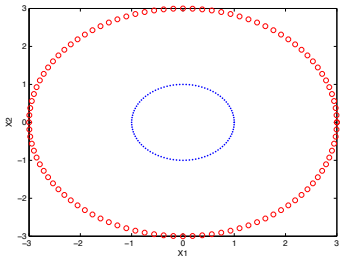
   (b) $K(x, z) = K_1(x, z) + K_3(x, z)$

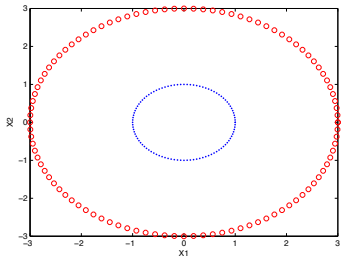
   (c) $K(x, z) = K_1(x, z)K_2(x, z)$

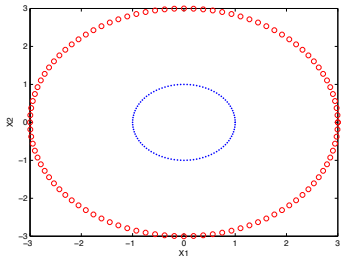
4. **Exponential Kernel** [10pt]. Prove that the exponential of kernel is again a kernel, i.e. $K(x, z) = \exp(K_1(x, z))$. Assume that we know the fact that addition and multiplication of kernels yield valid kernels. (Hint: use the Taylor expansion of exp)

5. **SVM with Kernel** [15pt]. You are given the following 3 plots, which illustrates a dataset with two classes. Draw the decision boundary when you train an SVM classifier with linear, polynomial (order 2) and RBF kernels respectively (a rough sketch is enough). Classes have equal number of instances.

   - Linear kernel:
   $$\kappa(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{x}^\top \boldsymbol{x}'$$
   - Polynomial kernel of degree p :
   $$\kappa(\boldsymbol{x}, \boldsymbol{x}') = \left(\boldsymbol{x}^\top \boldsymbol{x}' + 1\right)^p$$
   - RBF kernel:
   $$\kappa(\boldsymbol{x}, \boldsymbol{x}') = \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{x}'\|^2}{2\sigma^2}\right)$$







6. **SVM Back-propagation** [20pt]. During the lecture, we have discussed the SVM from loss minimization perspective with hinge loss. In this question, we will draw a computational graph of this approach and then use the back-propagation to update the parameter of the SVM. The objective function of SVM can be formulated as follows:

$$\mathcal{L} = \sum_{i=1}^{N} \max\left\{0, 1 - y_n\left(\langle \boldsymbol{w}, \boldsymbol{x}_n \rangle + b\right)\right\} + \lambda \|\boldsymbol{w}\|^2$$

Let's assume $\boldsymbol{x} = [x_1, x_2]^\top \in \mathbb{R}^2$ (therefore, $\boldsymbol{w} = [w_1, w_2]^\top \in \mathbb{R}^2$). Solve the following questions:

(a) Assume that we only have a single data point in our model ($N = 1$). Draw a computational graph of the objective function.

(b) Perform a forward propagation with $x_1 = -2, x_2 = 4, y = -1, w_1 = 3, w_2 = 2, b = 3, \lambda = 1$.

(c) Perform a backward propagation for the model parameters w_1, w_2, and b given the result of the forward propagation.

(d) Perform a single gradient descent step with learning rate 1 on the model parameters.