COMPSCI 671D: Homework 4

Due: February 28, 2019

Instructions: This homework is out of *100 points.* All the questions in this homework are mandatory. We expect mathematical rigour for required proofs and derivations. The final answers of the homework problems should be submitted as a single PDF file (we do not appreciate you using any other format for submissions). If we require you to also report your code, you must embed it in the PDF you are submitting. It will be okay if you also feel the need to submit a separate code file along with the PDF to support your answers. We recommend using well typed LaTeX solutions. Figures, plots, and tables should be well captioned and indexed. Please start answering a new problem on new page. You are welcome to look up the answers on the Internet in order to prove these, but you must type the proofs yourself. Do not ask your classmates to provide a link for you to the answers, and please do not share your link with others; each student must look it up individually because we think this is a good way to learn. If you have referred any web sources, research papers or textbooks you must cite them. Thank you!

1 Convexity I *(20 Points)*

Prove/disprove the following properties for convex function(s):

1. ***(3 points).*** The sum of two convex functions is also convex.
2. ***(9 points).*** Let $f(x) = h(g_1(x), g_2(x), .., g_n(x)))$. Then for each of the following cases, prove that f is convex:
   - ***(3 points).*** h is convex, $\forall i \in \{1, .., n\}$ g_i is a convex function and h is increasing in its i-th component.
   - ***(3 points).*** h is convex, $\forall i \in \{1, .., n\}$ g_i are affine functions.
   - ***(3 points).*** h is convex, $\forall i \in \{1, .., n\}$ g_i is a concave function and h is decreasing in its i-th component.
3. ***(8 points).*** The maximum of a convex function f over the polyhedron P is achieved at one of its vertices.

2 Convexity II *(30 Points)*

Let the primal optimization problem be: $\min_{x \in \mathbb{R}^2} f(x)$ s.t. $g(x) \leq 0$, then let $\mathcal{L}(x, \lambda) = \max_{\lambda \in \mathbb{R}} q(\lambda)$ be the Lagrangian dual problem where $q(\lambda) = \min_x (f(x) + \lambda g(x))$.

Consider $f(x) = x_1^2 + x_2^2$ and $g(x) = 1 - x_1 - x_2$.

1. ***(5 points).*** Plot the feasible region for the primal problem in x_1 and x_2 space. Let (x_1^*, x_2^*) be the point in (x_1, x_2) space which minimizes $f(x)$ s.t. $g(x) \leq 0$. Plot (x_1^*, x_2^*) on the plot which also contains the feasible region.

2. ***(10 points).*** Plot sets $R = \{(y, z) | y = g(\omega), z = f(\omega) \text{ for } \omega \in \mathbb{R}^2\}$ and F (set of feasible solution space for primal problem) in (y, z) space where $y = g(\omega)$ and $z = f(\omega)$.

3. ***(5 points).*** Let y^*, z^* and λ^* be the optimal values of $y = g(\omega)$, $z = f(\omega)$ and λ respectively which optimizes $\mathcal{L}(\omega, \lambda)$. Plot the optimal solution (y^*, z^*) on a plot. Also draw $z + \lambda^* y = q(\lambda^*)$ in $y = g(\omega)$ and $z = f(\omega)$ space.

4. ***(10 points).*** Show that $q(\lambda)$ is a concave function on S_q where $S_q = \{\lambda | q(\lambda) \in \mathbb{R}\}$.

3 Support Vector Machine *(30 Points)*

For linearly separable data, support vector machine tries to find two parallel hyperplanes parameterized by (w, b) such that the margin between two classes is maximized. If the class indicator $y_i = 1$, then $w^T x_i + b \geq 1$ and if the class indicator $y_i = -1$, then we have $w^T x_i + b \leq -1$. Thus, a perfectly learned separator satisfies $y_i(w^T x_i + b) \geq 1$. However, if the data is not fully separable then we allow for error margins $\epsilon_i > 0$. We update the primal problem for SVM as follows:

$$\min_w \left(\frac{1}{2} \|w\|_2^2 + C \cdot \sum_i \epsilon_i \right)$$

$$s.t. \;\; y_i(w^T x_i + b) \geq (1 - \epsilon_i) \;\; \forall i \in \{1, ..., n\}$$

$$\epsilon_i \geq 0 \;\; \forall i \in \{1, ..., n\}$$

1. ***(7 points).*** Show that the Lagrangian Dual Problem of the SVM primal problem can be given by a quadratic program with Lagrange multipliers $\alpha_i \;\; \forall i \in \{1, ..., n\}$.

2. ***(7 points).*** Show that for the primal and dual optimal solutions, the Lagrange multiplier α_i is non-zeros for samples where $y_i(W^T x_i + b) \leq 1$.

Now consider the diabetic retinopathy dataset (provided in the assignment zip file) which contains features extracted from the Messidor image set to predict whether an image contains signs of diabetic retinopathy or not. All features represent either a detected lesion, a descriptive feature of a anatomical part or an image-level descriptor. We will use Python 3.x for this part of the problem.

3. Read the data from the CSV into a numpy array (use appropriate datatype). Split the dataset into a training set and a testing set with ratio $|training| : |testing| = 3 : 2$.

4. ***(4 points).*** Use scikit-learn's SVM model with "linear" kernel, $C = 0$ and default values for other parameters to fit the training set for predicting the "class_label" given in the dataset. Report training accuracy, testing accuracy and the learned weights w^*.

5. ***(4 points).*** Plot C vs the sum of training errors values $\sum_i \epsilon_i$ for each $C \in [0, 1000]$. What do you observe? Comment on your observations.

6. ***(4 points).*** Now, plot C vs the testing accuracy. Which value of C would have been optimal for this dataset? Comment on your findings.

7. ***(4 points).*** Attempt anyone of the following:

   (a) ***(4 points).*** Let's choose the value of C to be the optimal one you found in the previous part. Let's change the kernel to "poly" and play with the *degree* of the polynomial. For what value of the *degree* of the polynomial kernel do you find achieves the best testing accuracy? Comment on your findings. (Plot *degree* vs testing accuracy, and include *degree* vs training accuracy on the same plot.)

OR

(b) ***(4 points).*** Let's choose the value of C to be the optimal one you found in the previous part. Let's change the kernel to "rbf" and play with the *gamma* parameter. For what value of the *gamma* do you find achieves the best testing accuracy? Comment on your findings. (Plot *gamma* vs testing accuracy, and include *gamma* vs training accuracy on the same plot.)

4 Kernels *(20 Points)*

1. ***(5 points).*** Consider l_2 regularized logistic regression with loss function

$$\mathcal{L}(f, \theta_0) = \sum_{i=1}^{n} \ln(1 + \exp(-y_i(f(\mathbf{x}_i) + \theta_0)) + \lambda \|f\|_{\mathcal{H}}^2$$

where $y_i \in \{-1, 1\}$, $f(\mathbf{x}) = \theta^\top \mathbf{x}$, and $\|f\|_{\mathcal{H}}^2 = \theta^\top \theta$. Define the reproducing kernel Hilbert space $\mathcal{H}$. Using the representer theorem, what do you know about the optimal solution?

2. ***(15 points).*** Suppose we have two kernels k_1 and k_2, then prove/disprove that following are valid kernels:

   (a) ***(3 points).*** $k(x, z) = \alpha k_1(x, z) + \beta k_2(x, z)$ for $\alpha, \beta \geq 0$

   (b) ***(3 points).*** $k(x, z) = k_1(x, z) k_2(x, z)$

   (c) ***(3 points).*** $k(x, z) = f(x) f(z)$ for $f : \mathcal{X} \to \mathbb{R}$

   (d) ***(3 points).*** $k(x, z) = f(k_1(x, z))$ for f a polynomial with positive cofficients

   (e) ***(3 points).*** $k(x, z) = e^{-\alpha ||x - z||_2^2}$