CSCI 567, Spring 2021
Sirisha Rambhatla

Homework #3

Released: Feb 19, 2021
Due: 11:59 pm, Mar 5, 2021
Total scores: 40 points

Instructions

Notes on notation:

- Unless stated otherwise, scalars are denoted by small letter in normal font, vectors are denoted by small letters in bold font and matrices are denoted by capital letters in bold font.
- $\|.\|$ means L2-norm unless specified otherwise i.e. $\|.\| = \|.\|_2$

Problem 1 Logistic Regression (15 points)

Recall that the logistic regression model is defined as:

$$p(y = 1|\mathbf{x}) = \sigma\left(\mathbf{w}^T\mathbf{x} + b\right) \tag{1}$$

$$\sigma(z) = \frac{1}{1 + e^{-z}} \tag{2}$$

Given a training set $\mathcal{D} = \{(\mathbf{x}_n, y_n)\}_{n=1}^N$, where $\mathbf{x}_n \in \mathbb{R}^{K\times 1}$ and $y_n \in \{0, 1\}$, we will minimize the cross-entropy error function to solve $\mathbf{w}$.

$$\begin{aligned}\min_{\mathbf{w},b} L(\mathbf{w}, b) &= \min_{\mathbf{w},b} -\sum_n \{y_n \log\left[p\left(y_n = 1|\mathbf{x}_n\right)\right] + (1 - y_n)\log\left[p\left(y_n = 0|\mathbf{x}_n\right)\right]\} \\ &= \min_{\mathbf{w},b} -\sum_n \left\{y_n \log\left[\sigma\left(\mathbf{w}^T\mathbf{x}_n + b\right)\right] + (1 - y_n)\log\left[1 - \sigma\left(\mathbf{w}^T\mathbf{x}_n + b\right)\right]\right\}\end{aligned} \tag{3}$$

1.1 Please derive the update rule for w using Gradient Descent (GD) method. **(3 points)**

1.2 Suppose we have four training samples $(x_1, y_1) = (1, 0)$, $(x_2, y_2) = (1, 1)$, $(x_3, y_3) = (1, 1)$ and $(x_4, y_4) = (1, 1)$. Suppose our logistic regression model is $p(y = 1|x) = \sigma(wx)$ We initialize this model with $w = 0$ and use learning rate = 0.001. When using GD to optimize this model, after one batch iteration, what's the training accuracy? **(9 points)**

1.3 Based on the model we get in problem 1.2, if we have a test dataset containing three samples: $(x_1, y_1) = (-1, 0)$, $(x_2, y_2) = (1, 1)$, $(x_3, y_3) = (1, 0)$, what is the test accuracy? **(3 points)**

Problem 2 Kernel methods (10 points)

In class, we studied kernel functions and their properties. Consider the following function:

$$k(\mathbf{x}, \mathbf{x}') = \begin{cases} 1, & \text{if } \mathbf{x} = \mathbf{x}' \\ 0, & \text{otherwise} \end{cases} \quad \forall \mathbf{x}, \mathbf{x}' \in \mathbb{R}^D. \tag{4}$$

2.1 Prove that this is a valid kernel. You can apply the Mercer's theorem mentioned in the lecture and for simplicity assume that the N points $\{\mathbf{x}_1, \cdots, \mathbf{x}_N\}$ you pick are distinct (i.e., $\mathbf{x}_i \neq \mathbf{x}_j$ if $i \neq j$). (If you want, you can also prove without making this assumption with Mercer's theorem or even prove by using the definition of kernel functions.) **(3 points)**

2.2 Suppose now you are given a training set $\{(\mathbf{x}_n \in \mathbb{R}^D, y_n \in \mathbb{R})\}_{n=1}^N$ for a linear regression problem, where $\mathbf{x}_i \neq \mathbf{x}_j$ if $i \neq j$. Show that by using this kernel, the least square solution (with no regularization) will always lead to a total square loss of 0—meaning that all the training examples are *predicted accurately* by the least square solution. **(4 points)**

2.3 Although the least square solution has 0 loss on the training set, it in fact does not generalize to the test data at all (that is, this algorithm completely overfits the training data). Specifically, show that for any unseen data point $\mathbf{x}$, that is, $\mathbf{x} \notin \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$, the prediction of the least square solution is always 0. (If you have worked out the feature mapping in Q2.1, which was optional, you will understand better why this happens.) **(3 points)**

Problem 3 Support Vector Machine (15 points)

Consider the dataset consisting of points (x, y), where x is a real value, and $y \in \{-1, 1\}$ is the class label. There are only three points $(x_1, y_1) = (0, 1)$, $(x_2, y_2) = (\frac{\pi}{2}, -1)$, $(x_3, y_3) = (\pi, 1)$. Let the feature mapping $\boldsymbol{\phi}(x) = [\cos x, \sin x]^T$, corresponding to the kernel function $k(x, y) = \cos(x - y)$.

3.1 Write down the primal and dual formulations of SVM for this dataset in the transformed two-dimensional feature space based on $\boldsymbol{\phi}(\cdot)$. Note that we assume the data points are separable and set the hyperparameter C to be $+\infty$, which forces all slack variables (ξ) in the primal formulation to be 0 (and thus can be removed from the optimization). **(7 points)**

3.2 Next, solve the dual formulation. Based on that, derive the primal solution. **(8 points)**