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 $\mathbf{w}$.

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

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

$$\begin{aligned}\mathbf{w}^{(t+1)} &\leftarrow \mathbf{w}^{(t)} - \eta \frac{\partial L(\mathbf{w})}{\partial \mathbf{w}} \\ &= \mathbf{w}^{(t)} - \eta \sum_n \left[\sigma\left(\mathbf{w}^T\mathbf{x}_n + b\right) - y_n\right] \mathbf{x}_n\end{aligned} \tag{4}$$

**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)**

Calculate the gradient,

$$\frac{\partial L(w)}{\partial w} = \sum_n \{[\sigma\,(wx_n) - y_n]\, x_n\} \quad \textbf{(3 points)} \tag{5}$$

Substitute the data points and then set w = 0,

$$\frac{\partial L(w)}{\partial w} = [\sigma(w) - 0] + [\sigma(w) - 1] + [\sigma(w) - 1] + [\sigma(w) - 1] = -1 \tag{6}$$

Therefore, according to the GD update rule, $w = 0 - 0.001 * (-1) = 0.001$ **(3 points)**

Predictions on training data:

$$\begin{aligned}\hat{y}_1 &= \mathbb{I}\left[\sigma\left(w^* x_1\right) > 0.5\right] = 1 \neq y_1 \\ \hat{y}_2 &= \mathbb{I}\left[\sigma\left(w^* x_2\right) > 0.5\right] = 1 = y_2 \\ \hat{y}_3 &= \mathbb{I}\left[\sigma\left(w^* x_3\right) > 0.5\right] = 1 = y_3 \\ \hat{y}_4 &= \mathbb{I}\left[\sigma\left(w^* x_4\right) > 0.5\right] = 1 = y_3\end{aligned} \tag{7}$$

The training accuracy is $\frac{3}{4}$. **(3 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)**

The test accuracy is $\frac{2}{3}$.

## Problem 2 Kernel methods (10 points)

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

$$k(\mathbf{x}, \mathbf{x}') = \begin{cases} 1, & \text{if } \mathbf{x} = \mathbf{x}' \\ 0, & \text{otherwise} \end{cases} \quad \forall \mathbf{x}, \mathbf{x}' \in \mathbb{R}^D. \tag{8}$$

**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)**

For any $N$ distinct points $\{\mathbf{x}_1, \cdots, \mathbf{x}_N\}$, the kernel matrix $\mathbf{K}$ is an $N$-by-$N$ identity matrix, which is positive (semi)definite. (This answers the question already under the simplifying assumption.)

In general, if there are $M \leq N$ distinct points $\{\mathbf{x}^{(1)}, \cdots, \mathbf{x}^{(M)}\}$ in this set, then the kernel matrix $\mathbf{K}$ can be written as $\sum_{j \in M} \mathbf{v}_j \mathbf{v}_j^T$ where $\mathbf{v}_j \in \{0, 1\}^N$ is such that $\mathbf{v}_{jn} = 1$ if $\mathbf{x}_n = \mathbf{x}^{(j)}$ and 0 otherwise. Clearly this matrix is positive semidefinite and by Mercer's theorem $k$ is a kernel function.

Alternatively, one can verify that the following mapping $\boldsymbol{\phi}$ certifies that $k$ is a kernel by definition: $\boldsymbol{\phi}(\boldsymbol{x})$ is a point in $\{0, 1\}^{\mathbb{R}}$ so that only the coordinate corresponding to $\boldsymbol{x}$ is 1.

**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)**

By the formula derived in the lecture, the coefficient $\boldsymbol{\alpha}$ is $\mathbf{K}^{-1}\mathbf{y} = \mathbf{y}$ since $\mathbf{K}$ is the identity matrix. Therefore, the prediction of the least square solution on any example $(\mathbf{x}_m, y_m)$ is

$$\sum_{n=1}^{N} \alpha_n k(\mathbf{x}_n, \mathbf{x}_m) = \alpha_m = y_m,$$

that is, it correctly predicts the outcome and thus has 0 square loss.

**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)**

The prediction is $\sum_{n=1}^{N} a_n k(\mathbf{x}_n, \mathbf{x})$, and since $\mathbf{x} \neq \mathbf{x}_n$ for all $n$, by the definition of the kernel, the prediction is $\sum_{n=1}^{N} a_n 0 = 0$.

## 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 ($\xi$) in the primal formulation to be 0 (and thus can be removed from the optimization). **(7 points)**

General primal formulation of SVM for separable data is:

$$\begin{aligned} \min_{\mathbf{w}, b} \quad & \frac{1}{2}\|\mathbf{w}\|_2^2 \\ \text{s.t.} \quad & y_n[\mathbf{w}^T \boldsymbol{\phi}(x_n) + b] \geq 1, \forall n \end{aligned}$$

Plugging in the specific dataset gives:

$$\begin{aligned} \min_{w_1, w_2, b} \quad & \frac{1}{2}(w_1^2 + w_2^2) && \textbf{(1 points)} \\ \text{s.t.} \quad & w_1 + b \geq 1 && \textbf{(1 points)} \\ & -w_2 - b \geq 1 && \textbf{(1 points)} \\ & -w_1 + b \geq 1 && \textbf{(1 points)} \end{aligned}$$

(**Remark:** Give 1 points for partial credit if only the general primal formulation is present.)
General dual formulation of SVM is:

$$\begin{aligned} \max_{\alpha} \quad & \sum_n \alpha_n - \frac{1}{2}\sum_{m,n} y_m y_n \alpha_m \alpha_n k(x_m, x_n) \\ \text{s.t.} \quad & \alpha_n \geq 0, \forall n \\ & \sum_n a_n y_n = 0 \end{aligned}$$

Plugging in the specific dataset gives:

$$\begin{aligned} \max_{\alpha_1, \alpha_2, \alpha_3 \geq 0} \quad & \alpha_1 + \alpha_2 + \alpha_3 - \frac{1}{2}\alpha_1^2 - \frac{1}{2}\alpha_2^2 - \frac{1}{2}\alpha_3^2 + \alpha_1\alpha_3 && \textbf{(1 points)} \\ \text{s.t.} \quad & \alpha_2 = \alpha_1 + \alpha_3 && \textbf{(1 points)} \\ & \alpha_i \geq 0 && \textbf{(1 points)} \end{aligned}$$

(**Remark:** Give 1 points for partial credit if only the general dual formulation is present.)

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

Eliminating the dependence on $\alpha_2$ using the constraint $\alpha_1 + \alpha_3 = \alpha_2$, we arrive at the objective

$$\max_{\alpha_1,\alpha_2 \geq 0} \quad 2\alpha_1 - \alpha_1^2 + 2\alpha_2 - \alpha_2^2. \qquad \textbf{(2 points)}$$

Clearly, we can maximize over $\alpha_1$ and $\alpha_2$ separately, which gives $\alpha_1^* = \alpha_3^* = 1$ and thus $\alpha_2^* = 2$ **(2 points)**.

The primal solution can be found by

$$(w_1^*, w_2^*)^T = \sum_{n=1}^{3} y_n \alpha_n^* \boldsymbol{\phi}(x_n) = [0, -2]^T \qquad \textbf{(2 points)}$$

$$b^* = y_1 - \mathbf{w}^{*T} \boldsymbol{\phi}(x_1) = 1 \qquad \text{(using any example works in this case)} \qquad \textbf{(2 points)}$$