CS 475 Machine Learning: Homework 1
Supervised Classifiers 1
Due: Thursday September 26, 2019, 11:59pm
100 Points Total Version 1.0

Asef Islam (aislam5)

Instructions

We have provided this LaTeX document for turning this homework. We give you one or more boxes to answer each question. The question to answer for each box will be noted in the title of the box.

Other than your name, do not type anything outside the boxes. Leave the rest of the document unchanged.

For written answers, replace the **\TextRequired** (**Place Answer Here**) command with your answer. For the following example *Question 0.1* , you would place your answer where **\TextRequired** (**Place Answer Here**) is located,

Place Answer Here

Do not change the height or title of the box. If your text goes beyond the box boundary, it will be cut off. We have given sufficient space for each answer, so please condense your answer if it overflows. The height of the box is an upper bound on the amount of text required to answer the question - many answers can be answered in a fraction of the space. Do not add text outside of the boxes. We will not read it.

For True/False or Multiple Choice questions, place your answers within the defined table. To mark the box(es) corresponding to your answers, replace **\Unchecked** (☐) commands with the **\Checked** (☒) command. Do not make any other changes to the table. For example, in *Question 0.2* ,

☒	Logistic Regression
☐	Perceptron

For answers that require a single equation, we will provide a specific type of box, such as in the following example *Question 0.3* . Please type the equation where **\EquationRequired** (**Type Equation Here**) without adding any $ signs or **\equation** commands. Do not put any additional text in this field.

$\mathbf{w} =$ **Type Equation Here**

For answers that require multiple equations, such as a derivation, place all equations within the specified box. You may include text short explanations if you wish (as shown in *Question 0.4*). You can put the equations in any format you like (e.g. within $ or $$, the **\equation** environment, the **\align** environment) as long as they stay within the box.

$x + 2$ x is a real number

the following equation uses the variable y

$y + 3$

Do not change any formatting in this document, or we may be unable to grade your work. This includes, but is not limited to, the height of textboxes, font sizes, and the spacing of text and tables. Additionally, do not add text outside of the answer boxes. Entering your answers are the only changes allowed.
We strongly recommend you review your answers in the generated PDF to ensure they appear correct. We will grade what appears in the answer boxes in the submitted PDF, NOT the original latex file.

1) Regularization (26 points)

In class, we discussed adding a regularization penalty term to our objective function. This gives us an optimization problem of the following general form

$$\underset{\mathbf{w} \in \mathbb{R}^D}{\operatorname{argmin}} \ \ell(\mathbf{w}) + \lambda \cdot \Omega(\mathbf{w}) \tag{1}$$

where ℓ is our usual loss function, $\lambda \geq 0$, and

$$\Omega_q(\mathbf{w}) \stackrel{\text{def}}{=} \sum_{j=1}^{M} |w_j|^q \tag{2}$$

where $q \geq 0$ is a hyper-parameter that can be experimented with. In class, we discussed $q = 2$ and $q = 1$ as they are the most common in practice. This is a nice family of regularization functions because $\Omega_q(\mathbf{w})$ is convex in $\mathbf{w}$ for $q \geq 1$.

(1) (5 points) Show that $\Omega_q(\mathbf{w})$ is convex in $\mathbf{w}$ for $q = 1$ and $q = 2$.[1]

A function $f(x)$ is convex if $\lambda f(x) + (1-\lambda) f(y) \geq f(\lambda x + (1-\lambda) y)$ for any $\lambda \in [0, 1]$ and any x, y in the domain of the f.
For q = 1, we seek to prove that $f = \lambda\Omega_1(\mathbf{x}) + (1 - \lambda)\Omega_1(\mathbf{y}) - \Omega_1(\lambda\mathbf{x} + (1-\lambda)\mathbf{y}) \geq 0$. Then $f = \lambda\sum_{j=1}^{M} |x_j| + (1-\lambda)\sum_{j=1}^{M} |y_j| - \sum_{j=1}^{M} |\lambda x_j + (1-\lambda) y_j| = \sum_{j=1}^{M} (|x_j| + |y_j| - |\lambda x_j + (1-\lambda) y_j|)$. By the triangle inequality $|\lambda x_j| + |(1-\lambda) y_j| \geq |\lambda x_j + (1-\lambda) y_j|$ and thus $f \geq 0$ and $\Omega_1(\mathbf{w})$ is convex.
For q = 2, we seek to prove that $f = \lambda\Omega_2(\mathbf{x}) + (1 - \lambda)\Omega_2(\mathbf{y}) - \Omega_2(\lambda\mathbf{x} + (1-\lambda)\mathbf{y}) \geq 0$.Then $f = \lambda\sum_{j=1}^{M} |x_j|^2 + (1-\lambda)\sum_{j=1}^{M} |y_j|^2 - \sum_{j=1}^{M} |\lambda x_j + (1-\lambda) y_j|^2 = \lambda\sum_{j=1}^{M} x_j^2 + (1-\lambda)\sum_{j=1}^{M} y_j^2 - \sum_{j=1}^{M} (\lambda x_j + (1-\lambda) y_j)^2 = \lambda\sum_{j=1}^{M} x_j^2 + (1-\lambda)\sum_{j=1}^{M} y_j^2 - \sum_{j=1}^{M} (\lambda^2 x_j^2 + 2\lambda(1-\lambda) x_j y_j + (1-\lambda)^2 y_j^2) = \sum_{j=1}^{M} [(\lambda - \lambda^2) x_j^2 + (\lambda - \lambda^2) y_j^2 - 2\lambda(1-\lambda) x_j y_j] = \lambda(1-\lambda)\sum_{j=1}^{M} (x_j^2 - 2 x_j y_j - y_j^2) = \lambda(1-\lambda)(x_j - y_j)^2 \geq 0$ because $\lambda, 1-\lambda$, and $(x_j - y_j)^2$ are each always ≥ 0. Hence $f \geq 0$ as desired and $\Omega_2(\mathbf{w})$ is convex.

(2) (5 points) If the loss function $\ell(\mathbf{w})$ is convex in $\mathbf{w}$, is $\ell(\mathbf{w}) + \lambda \cdot \Omega(\mathbf{w})$ necessarily convex? Why or why not?

If $\ell(\mathbf{w})$ and $\Omega(\mathbf{w})$ are convex then $p = \mu\ell(\mathbf{x}) + (1-\mu)\ell(\mathbf{y}) - \ell(\mu\mathbf{x} + (1-\mu)\mathbf{y}) \geq 0$ and $q = \mu\Omega(\mathbf{x}) + (1-\mu)\Omega(\mathbf{y}) - \Omega(\mu\mathbf{x} + (1-\mu)\mathbf{y}) \geq 0$ for any $\mu \in [0, 1]$ and any pair of $\mathbf{x}, \mathbf{y}$ in the domains of each function. Then if $f(\mathbf{w}) = \ell(\mathbf{w}) + \lambda\Omega(\mathbf{w})$, $\mu f(\mathbf{x}) + (1-\mu) f(\mathbf{y}) - f(\mu\mathbf{x} + (1-\mu)\mathbf{y}) = \mu\ell(\mathbf{x}) + (1-\mu)\ell(\mathbf{y}) - \ell(\mu\mathbf{x} + (1-\mu)\mathbf{y}) + \lambda(\mu\Omega(\mathbf{x}) + (1-\mu)\Omega(\mathbf{y}) - \Omega(\mu\mathbf{x} + (1-\mu)\mathbf{y})) = p + \lambda q \geq 0$ because p, λ, and q are each ≥ 0 and thus f satisfies the definition of convexity.

[1] For a challenge, but no extra points, show that $\Omega_q(\mathbf{w})$ is convex for all $q \geq 1$.

(3) (5 points) In words, what does Ω_0 compute? Why can't we use it in gradient-based optimization?

$\Omega_0 = \sum_{j=1}^{M} |w_j|^0$ and $|w_j|^0 = 1$ if $w_j \neq 0$ and $= 0$ if $w_j = 0$, thus this function is not continuous and thus not differentiable. If we can not take the derivative we can not use it in gradient-based optimization.

(4) (6 points) Consider the following modification to the optimization problem in equation (1): set $q = 2$ but add $\lambda \geq 0$ to the set of variables being optimized.

$$\underset{\mathbf{w} \in \mathbb{R}^D, \lambda \geq 0}{\operatorname{argmin}} \quad \ell(\mathbf{w}) + \lambda \cdot \Omega_2(\mathbf{w}) \tag{3}$$

Why will the optimal value of λ be *zero*?

$\lambda \cdot \Omega_2(\mathbf{w})$ is ≥ 0 because $\lambda \geq 0$ and $\Omega_2(\mathbf{w}) \geq 0$ since it is a sum of squares. Thus, minimization of $\ell(\mathbf{w}) + \lambda \cdot \Omega_2(\mathbf{w})$ will occur when $\lambda \cdot \Omega_2(\mathbf{w}) = 0$. This can be guaranteed to happen when $\lambda = 0$ and thus the optimal value of λ will always be determined to be 0.

(5) (5 points) Since we can't seem to optimize λ with training data, how might we modify our experimental setup to enable choosing the parameters λ and q in a principled way?

We can search for the optimum values of λ and q experimentally using a somewhat brute-force approach. In other words, we can incrementally sweep through a range of possible values for both parameters and for each combination experimentally calculate the accuracy on test data and choose the pair of values that maximizes the accuracy.

2) Linear Regression (12 points)

Suppose you observe n data points, (x_1, y_1), ..., (x_n, y_n), where all x_i and all y_i are *scalars* (i.e., one-dimensional). Suppose further that each data point is paired with an *example weight*, $\alpha_i \geq 0$. These weights can be useful, for example, if some data points should have more (large α_i) or less (small α_i) influence on the loss. Suppose you choose the model $\hat{y} = w \cdot x$ and aim to minimize the α-weighted sum of squares error

$$\frac{1}{2}\sum_{i=1}^{n} \alpha_i (w \cdot x_i - y_i)^2 \tag{4}$$

Derive the closed-form solution for w showing each step. Is the solution necessarily a global minimum? Explain why or why not.

We seek to find w to minimize

$$\frac{1}{2}\sum_{i=1}^{n} \alpha_i (w \cdot x_i - y_i)^2$$

so we take the derivative with respect to w and set it to 0.

$$\frac{\partial}{\partial w}\frac{1}{2}\sum_{i=1}^{n} \alpha_i (w \cdot x_i - y_i)^2 = \sum_{i=1}^{n} (\alpha_i (w \cdot x_i - y_i) \cdot x_i) = \sum_{i=1}^{n} (\alpha_i \cdot w \cdot x_i^2 - \alpha_i \cdot y_i \cdot x_i) = 0$$

$$w \sum_{i=1}^{n} (\alpha_i \cdot x_i^2) = \sum_{i=1}^{n} (\alpha_i \cdot y_i \cdot x_i)$$

$$w^* = \frac{\sum_{i=1}^{n} \alpha_i \cdot y_i \cdot x_i}{\sum_{i=1}^{n} \alpha_i \cdot x_i^2}$$

Hence w^* is a local minimum for the sum of squares error as it produces a derivative of 0. If we can show that the sum of squares error is convex, then w^* is a global minimum. We can show convexity if the second derivative is ≥ 0

$$\frac{\partial^2}{\partial w^2}\frac{1}{2}\sum_{i=1}^{n} \alpha_i (w \cdot x_i - y_i)^2 = \frac{\partial}{\partial w}\sum_{i=1}^{n} (\alpha_i \cdot w \cdot x_i^2 - \alpha_i \cdot y_i \cdot x_i) = \sum_{i=1}^{n} \alpha_i \cdot x_i^2 \geq 0$$

because α_i and x_i^2 are both ≥ 0. Thus the sum of squares error is convex, and w^* is a global minimum.

3) Support vector machines (12 points)

In this question, we will ask you to extend the slack formulation of the support vector machine to allow for asymmetric costs for misclassification. Consider the following scenario, a doctor using a classifier to predict whether or not they should order more tests ($y = +1$) or triage the patient ($y = -1$) based on a preliminary set of tests they have already done (i.e., features). Clearly, we prefer to have more information that can be provided by additional tests, however, tests carry some risk and may be unnecessary. Mathematically, what we have is an asymmetry between *false-positives* and *false-negatives*.

Extend the slack formulation of the SVM from class to penalize the slack variables for *false-positives* and *false-negatives* differently. Rather than a single $C \geq 0$ coefficient, the new formulation should leverage two coefficients $C^{(+)}, C^{(-)} \geq 0$.

For reference, here is the slack formulation. Feel free to copy-paste and modify it.

$$\underset{w \in \mathbb{R}^D}{\text{minimize}} \ \frac{1}{2}||w||^2 + C\sum_{i=1}^{n} \xi_i \tag{5}$$

$$\text{subject to} \quad (w^\top x_i)y_i + \xi_i \geq 1, \quad i = 1 \dots n \tag{6}$$

$$\xi_i \geq 0, \quad i = 1 \dots n \tag{7}$$

$$\underset{w \in \mathbb{R}^D}{\text{minimize}} \ \frac{1}{2}||w||^2 + C^{(+)} \sum_{\{i|y_i=1\}}^{n} \xi_i + C^{(-)} \sum_{\{i|y_i=-1\}}^{n} \xi_i \tag{8}$$

$$\text{subject to} \quad (w^\top x_i)y_i + \xi_i \geq 1, \quad i = 1 \dots n \tag{9}$$

$$\xi_i \geq 0, \quad i = 1 \dots n \tag{10}$$