Homework 5: SGD for Multiclass Linear SVM

Due: Thursday, April 8, 2021 at 11:59PM EST **Instructions:** Your answers to the questions below, including plots and mathematical work, should be submitted as a single PDF file. It's preferred that you write your answers using software that typesets mathematics (e.g.LaTeX, LyX, or MathJax via iPython), though if you need to you may scan handwritten work. You may find the minted package convenient for including source code in your LaTeX document. If you are using LyX, then the listings package tends to work better.

---

1 Derivation

Suppose our output space and our action space are given as follows: $\mathcal{Y} = \mathcal{A} = \{1, \ldots, k\}$. Given a non-negative class-sensitive loss function $\Delta : \mathcal{Y} \times \mathcal{A} \to [0, \infty)$ and a class-sensitive feature mapping $\Psi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^d$. Our prediction function $f : \mathcal{X} \to \mathcal{Y}$ is given by

$$f_w(x) = \arg\max_{y \in \mathcal{Y}} \langle w, \Psi(x, y) \rangle .$$

For training data $(x_1, y_1), \ldots, (x_n, y_n) \in \mathcal{X} \times \mathcal{Y}$, let $J(w)$ be the ℓ_2-regularized empirical risk function for the multiclass hinge loss. We can write this as

$$J(w) = \lambda \|w\|^2 + \frac{1}{n} \sum_{i=1}^{n} \max_{y \in \mathcal{Y}} \left[\Delta(y_i, y) + \langle w, \Psi(x_i, y) - \Psi(x_i, y_i) \rangle \right]$$

for some $\lambda > 0$.

1. Show that $J(w)$ is a convex function of w. You may use any of the rules about convex functions described in our notes on Convex Optimization, in previous assignments, or in the Boyd and Vandenberghe book, though you should cite the general facts you are using. [Hint: If $f_1, \ldots, f_m : \mathbb{R}^n \to \mathbb{R}$ are convex, then their pointwise maximum $f(x) = \max\{f_1(x), \ldots, f_m(x)\}$ is also convex.]

   Solution:

   Consider for now just the second summation term, and in it, only 1 term in the summation.

   $$\max_{y \in \mathcal{Y}} \left[\Delta\left(y_i, y\right) + \langle w, \Psi(x_i, y) - \Psi(x_i, y_i)\rangle\right]$$

   Using the hint, we need just to prove that, for all y,

   $$f(w) = \langle w, \Psi(x_i, y) - \Psi(x_i, y_i)\rangle + \Delta\left(y_i, y\right)$$

   is convex wrt w.

   $f(w)$ is a function of w on $\mathbb{R}^n \to \mathbb{R}$ of the form $Aw + b$, where A is $\Psi(x_i, y) - \Psi(x_i, y_i)$ in $\mathbb{R}^{1*n}$, where $n =$ num_outFeatures which is a vector independent of w. b is a scalar value, also independent of w. So, using "2.3.2 Affine functions", $f(w)$ is a affine function of w.

   Note: The multiplication in affine functions is matrix multiplication, not inner product. However, because Ψ is a vector it can be said to have dimensions 1xn, and so the two here are equivalent

   From "Section 3.1.1" of the book, "All affine functions are convex"

   So, for all y, $f(w)$ is convex

   $f_1, \ldots, f_m : \mathbb{R}^n \to \mathbb{R}$ are convex, then their pointwise maximum $f(x) = \max\{f_1(x), \ldots, f_m(x)\}$ is also convex (from "3.2.3 Pointwise maximum and supremum")

   So

   $$\max_{y \in \mathcal{Y}} \left[\Delta\left(y_i, y\right) + \langle w, \Psi(x_i, y) - \Psi(x_i, y_i)\rangle\right]$$

   is also convex

   All norms are convex and since λ is ≥ 0, so first term $\lambda\|w\|^2$ is convex.

   Finally Summation of convex functions (this includes l2 norm as well as the terms in summation of maximums) is also convex, and so $J(w)$ is convex □

2. Since $J(w)$ is convex, it has a subgradient at every point. Give an expression for a subgradient of $J(w)$. You may use any standard results about subgradients, including the result from an earlier homework about subgradients of the pointwise maxima of functions. (Hint: It may be helpful to refer to $\hat{y}_i = \arg\max_{y \in \mathcal{Y}} [\Delta(y_i, y) + \langle w, \Psi(x_i, y) - \Psi(x_i, y_i) \rangle]$.)

   Solution:

   Let
   $$f(w) = \max_{y \in \mathcal{Y}} \Delta(y_i, y) + \langle w, \Psi(x_i, y) - \Psi(x_i, y_i) \rangle$$

   $f(w)$ is not continuous as it uses max. So, $f(w)$ is not differentiable. But since $f(w)$ is convex, it has a subgradient. Let $g(w)$ be a subgradient of $f(w)$

   Let $f_1(w), f_2(w), \ldots f_k(w)$ be k functions for all the k values of y

   Let $f_m(w)$ be the maximum of all these functions. So, if $g(w) \in \partial f_m(w)$, then $g(w) \in \partial f(w)$ (from HW3 Q1)

   It is given that
   $$\hat{y}_i = \arg\max_{y \in \mathcal{Y}} [\Delta(y_i, y) + \langle w, \Psi(x_i, y) - \Psi(x_i, y_i) \rangle]$$

   So,
   $$g(w) = \frac{\partial f_m(w)}{\partial w} = \frac{\partial \left(\Delta(y_i, \hat{y}_i) + \langle w, \Psi(x_i, \hat{y}_i) - \Psi(x_i, y_i) \rangle \right)}{\partial w}$$

   $$g(w) = \Psi(x_i, \hat{y}_i) - \Psi(x_i, y_i)$$

   Now, we find the subgradient of $J(w)$:
   $$J(w) = \lambda \|w\|^2 + f(w)$$

   The sub gradient is:
   $$2\lambda w + \Psi(x_i, \hat{y}_i) - \Psi(x_i, y_i)$$

   □

3. Give an expression for the stochastic subgradient based on the point (x_i, y_i).

   Solution:

$$2\lambda w + \Psi(x_i, \hat{y}_i) - \Psi(x_i, y_i)$$

   where

$$\hat{y}_i = \arg\max_{y \in \mathcal{Y}} \left[\Delta\left(y_i, y\right) + \langle w, \Psi(x_i, y) - \Psi(x_i, y_i)\rangle\right]$$

   So, the sub gradient can also be written as:

$$2\lambda w + \Psi(x_i, \arg\max_{y \in \mathcal{Y}} \left[\Delta\left(y_i, y\right) + \langle w, \Psi(x_i, y) - \Psi(x_i, y_i)\rangle\right]) - \Psi(x_i, y_i)$$

□

4. Give an expression for a minibatch subgradient, based on the points $(x_i, y_i), \ldots, (x_{i+m-1}, y_{i+m-1})$.

   Solution:

$$2\lambda w + \frac{1}{m} \sum_{j=i}^{i+m-1} \Psi(x_i, \hat{y_i}) - \Psi(x_i, y_i)$$

   which is the same as:

$$2\lambda w + \frac{1}{m} \sum_{j=i}^{i+m-1} \Psi(x_i, \arg\max_{y \in \mathcal{Y}} [\Delta(y_i, y) + \langle w, \Psi(x_i, y) - \Psi(x_i, y_i) \rangle]) - \Psi(x_i, y_i)$$

□

(Optional) Hinge Loss is a Special Case of Generalized Hinge Loss

Let $\mathcal{Y} = \{-1, 1\}$. Let $\Delta(y, \hat{y}) = \mathbb{1}y \neq \hat{y}$. If $g(x)$ is the score function in our binary classification setting, then define our compatibility function as

$$\begin{aligned} h(x, 1) &= g(x)/2 \\ h(x, -1) &= -g(x)/2. \end{aligned}$$

Show that for this choice of h, the multiclass hinge loss reduces to hinge loss:

$$\ell\left(h, (x, y)\right) = \max_{y' \in \mathcal{Y}} \left[\Delta\left(y, y'\right)) + h(x, y') - h(x, y)\right] = \max\left\{0, 1 - yg(x)\right\}$$

2 Implementation

In this problem we will work on a simple three-class classification example. The data is generated and plotted for you in the skeleton code.

One-vs-All (also known as One-vs-Rest)

First we will implement one-vs-all multiclass classification. Our approach will assume we have a binary base classifier that returns a score, and we will predict the class that has the highest score.

5. Complete the methods `fit`, `decision_function` and `predict` from `OneVsAllClassifier` in the skeleton code. Following the `OneVsAllClassifier` code is a cell that extracts the results of the fit and plots the decision region. You can have a look at it first to make sure you understand how the class will be used.

   Solution:

□

6. Include the results of the test cell in your submission.

   Solution:

□

Multiclass SVM

In this question, we will implement stochastic subgradient descent for the linear multiclass SVM, as described in class and in this problem set. We will use the class-sensitive feature mapping approach with the "multivector construction", as described in the multiclass lecture.

7. Complete the function `featureMap` in the skeleton code.

   Solution:

□

8. Complete the function `sgd`.

   Solution:

□

9. Complete the methods `subgradient`, `decision_function` and `predict` from the class `MulticlassSVM`.

   Solution:

□

10. Following the multiclass SVM implementation, we have included another block of test code. Make sure to include the results from these tests in your assignment, along with your code.

    Solution:

□