Machine Learning and Computational Statistics **PRELIMINARY VERSION** Homework 5: Generalized Hinge Loss and Multiclass SVM

Zhuoru Lin

Due: Thursday, April 6, 2017, at 10pm (Submit via Gradescope)

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 Introduction

NOTE: THIS PROBLEM SET IS NOT YET COMPLETE. A SIGNIFICANT PROGRAMMING PORTION WILL SOON BE ADDED. STAY TUNED.

The goal of this problem set is to get more comfortable with the multiclass hinge loss and multiclass SVM. In several problems below, you are asked to justify that certain functions are convex. For these problems, you may use any of the rules about convex functions described in our notes on Convex Optimization (https://davidrosenberg.github.io/mlcourse/Notes/convex-optimization.pdf) or in the Boyd and Vandenberghe book. In particular, you will need to make frequent use of the following result: If $f_1, \ldots, f_m : \mathbf{R}^n \to \mathbf{R}$ are convex, then their pointwise maximum

$$f(x) = \max\{f_1(x), \ldots, f_m(x)\}$$

is also convex.

2 Convex Surrogate Loss Functions

It's common in machine learning that the loss functions we really care about lead to optimization problems that are not computationally tractable. The 0/1 loss for binary classification is one such example[1]. Since we have better machinery for minimizing convex functions, a standard approach is to find a **convex surrogate loss function.** A convex surrogate loss function is a convex function

[1] Interestingly, if our hypothesis space is linear classifiers and we are in the "realizable" case, which means that there is some hypothesis that achieves 0 loss (with the 0/1 loss), then we can efficiently find a good hypothesis using linear programming. This is not difficult to see: each data point gives a single linear constraint, and we are looking for a vector that satisfies the constraints for each data point.

that is an upper bound for the loss function of interest[2]. If we can make the upper bound small, then the loss we care about will also be small[3]. Below we will show that the multiclass hinge loss based on a class-sensitive loss Δ is a convex surrogate for the multiclass loss function Δ, when we have a linear hypothesis space. We'll start with a special case, that the hinge loss is a convex surrogate for the 0/1 loss.

[2] At this level of generality, you might be wondering: "A convex function of WHAT?". For binary classification, we usually are talking about a convex function of the margin. But to solve our machine learning optimization problems, we will eventually need our loss function to be a convex function of some $w \in \mathbf{R}^d$ that parameterizes our hypothesis space. It'll be clear in what follows what we're talking about.

[3] This is actually fairly weak motivation for a convex surrogate. Much better motivation comes from the more advanced theory of **classification calibrated** loss functions. See Bartlett et al's paper "Convexity, Classification, and Risk Bounds." http://www.eecs.berkeley.edu/~wainwrig/stat241b/bartlettetal.pdf

2.1 Hinge loss is a convex surrogate for 0/1 loss

1. Let $f : \mathcal{X} \to \mathbf{R}$ be a classification score function for binary classification.

   (a) For any example $(x, y) \in \mathcal{X} \times \{-1, 1\}$, show that

   $$1(y \neq \operatorname{sign}(f(x)) \leq \max\{0, 1 - yf(x)\},$$

   where $\operatorname{sign}(x) = \begin{cases} 1 & x > 0 \\ 0 & x = 0 \\ -1 & x < 0 \end{cases}$.

   Answers: If $y = sign(x)$ then we automatically have $1(y \neq sign(f(x))) = 0 \leq max(0, 1 - yf(x))$.

   If $y \neq sign(x)$, then $yf(x) >< 0$ and $1 = 1(y \neq sign(f(x))) \leq 1 - yf(x) = max(0, 1 - yf(x))$.

   (b) Show that the hinge loss $\max\{0, 1 - m\}$ is a convex function of the margin m.
   Answer: Since both 0 and 1-m are convex function of m. $max(0, 1 - m)$

   (c) Suppose our prediction score functions are given by $f_w(x) = w^T x$. The hinge loss of f_w on any example (x, y) is then $\max\left\{0, 1 - yw^T x\right\}$. Show that this is a convex function of w.
   Define $g(w) = 1 - yw^T x$ and $g'(w') = 1 - y - w'^T x$. Then we have:

   $$\begin{aligned} g(\theta w + (1-\theta)w') &= 1 - y(\theta w + (1-\theta)yw't)^T x \\ &= 1 - \theta y w^T x - (1-\theta)yw'^T x \\ &\leq (1 - \theta y w^T x) + (1 - (1-\theta)yw'^T x) \\ &= g(\theta w) + g((1-\theta)w'). \end{aligned} \tag{1}$$

   Hence $g(w)$ is a convex function, so is $max\{0, g(w)\}$.

2.2 Generalized Hinge Loss

Consider the multiclass output space $\mathcal{Y} = \{1, \ldots, k\}$. Suppose we have a base hypothesis space $\mathcal{H} = \{h : \mathcal{X} \times \mathcal{Y} \to \mathbf{R}\}$ from which we select a compatibility score function. Then our final multiclass hypothesis space is $\mathcal{F} = \left\{f(x) = \arg\max_{y \in \mathcal{Y}} h(x, y) \mid h \in \mathcal{H}\right\}$. Since functions in $\mathcal{F}$ map into $\mathcal{Y}$, our action space $\mathcal{A}$ and output space $\mathcal{Y}$ are the same. Nevertheless, we will write our class-sensitive loss function as $\Delta : \mathcal{Y} \times \mathcal{A} \to \mathbf{R}$, even though $\mathcal{Y} = \mathcal{A}$. We do this to indicate that the true class goes in the first slot of the function, while the prediction (i.e. the action) goes in the second slot. This is important because we do not assume that $\Delta(y, y') = \Delta(y', y)$. It would not be unusual to have this asymmetry in practice. For example, false alarms may be much less costly than no alarm when indeed something is going wrong.

Our ultimate goal would be to find $f \in \mathcal{F}$ minimizing the empirical cost-sensitive loss:

$$\min_{f \in \mathcal{F}} \sum_{i=1}^{n} \Delta\left(y_i, f(x_i)\right).$$

Since binary classification with 0/1 loss is both intractable and a special case of this formulation, we know that this more general formulation must also be computationally intractable. Thus we are looking for a convex surrogate loss function.

1. Suppose we have chosen an $h \in \mathcal{H}$, from which we get the decision function $f(x) = \arg\max_{y \in \mathcal{Y}} h(x, y)$. Justify that for any $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, we have

$$h(x, y) \leq h(x, f(x)).$$

Answer: f(x) is define to be the $y \in \mathcal{Y}$ such that h(x,y) is at maximun. We automatically have $h(x, y) \leq h(x, f(x))$ for all x.

2. Justify the following two inequalities:

$$\begin{aligned}\Delta\left(y, f(x)\right) &\leq \Delta\left(y, f(x)\right) + h(x, f(x)) - h(x, y) \\ &\leq \max_{y' \in \mathcal{Y}} \left[\Delta\left(y, y'\right)) + h(x, y') - h(x, y)\right]\end{aligned}$$

The RHS of the last expression is called the **generalized hinge loss:**

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

We have shown that for any $x \in \mathcal{X}, y \in \mathcal{Y}, h \in \mathcal{H}$ we have

$$\ell\left(h, (x, y)\right) \geq \Delta(y, f(x)),$$

where, as usual, $f(x) = \arg\max_{y \in \mathcal{Y}} h(x, y)$. [You should think about why we cannot write the generalized hinge loss as $\ell\left(f, (x, y)\right)$.]

Answer: Since $f(x) = argmax_{y \in \mathcal{Y}} h(x, y)$, we must have $h(x, f(x)) \leq h(x, y)$. Therefore: $\Delta\left(y, f(x)\right) \leq \Delta\left(y, f(x)\right) + h(x, f(x)) - h(x, y)$. Since $f(x) \in \mathcal{A}$, we automatically have the second inequality.

3. We now introduce a specific base hypothesis space $\mathcal{H}$ of linear functions. Consider a class-sensitive feature mapping $\Psi : \mathcal{X} \times \mathcal{Y} \to \mathbf{R}^d$, and $\mathcal{H} = \left\{ h_w\left(x, y\right) = \langle w, \Psi(x, y) \rangle \mid w \in \mathbf{R}^d \right\}$. Show that we can write the generalized hinge loss for $h_w(x, y)$ on example (x_i, y_i) as

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

Answer: By definition of generalized hinge loss from part 2, we have:

$$\begin{aligned} \ell\left(h_w, (x_i, y_i)\right) \qquad &= \max_{y \in \mathcal{Y}} \left[\Delta(y_i, y) + h(x_i, y) - h(x_i, y_i) \right] \\ &= \max_{y \in \mathcal{Y}} \left[\Delta\left(y_i, y\right)\right) + \langle w, \Psi(x_i, y) - \Psi(x_i, y_i) \rangle \right]. \end{aligned}$$

4. We will now show that the generalized hinge loss $\ell\left(h_w, (x_i, y_i)\right)$ is a convex function of w. Justify each of the following steps.

   (a) The expression $\Delta(y_i, y) + \langle w, \Psi(x_i, y) - \Psi(x_i, y_i)\rangle$ is an affine function of w.

   (b) The expression $\max_{y \in \mathcal{Y}}\left[\Delta\left(y_i, y\right)\right) + \langle w, \Psi(x_i, y) - \Psi(x_i, y_i)\rangle]$ is a convex function of w.

Answer:
(a) $\Delta(y_i, y) + \langle w, \Psi(x_i, y) - \Psi(x_i, y_i)\rangle = \Delta(y_i, y) + w^T(\Psi(x_i, y) - \Psi(x_i, y_i))$ is an affine function of w.
(b) By part (a) $\Delta(y_i, y) + \langle w, \Psi(x_i, y) - \Psi(x_i, y_i)\rangle$ is an affine function of w, therefore is also a convex function of w. $\max_{y \in \mathcal{Y}}\left[\Delta\left(y_i, y\right)\right) + \langle w, \Psi(x_i, y) - \Psi(x_i, y_i)\rangle]$ is the pointwise maximun of a group of convex function, hence is also a convex function of w.

5. Conclude that $\ell\left(h_w, (x_i, y_i)\right)$ is a convex surrogate for $\Delta(y_i, f_w(x_i))$.

Answer: By part 2, we know that $\ell\left(h_w, (x_i, y_i)\right)$ is a surrogate for $\Delta(y_i, f_w(x_i))$. By part 4 we know $\ell\left(h_w, (x_i, y_i)\right)$ is convex. There fore $\ell\left(h_w, (x_i, y_i)\right)$ is a convex surrogate of $\Delta(y_i, f_w(x_i))$.

3 SGD for Multiclass SVM

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 $\Delta : \mathcal{Y} \times \mathcal{A} \to \mathbf{R}^{\geq 0}$ and a class-sensitive feature mapping $\Psi : \mathcal{X} \times \mathcal{Y} \to \mathbf{R}^d$. Our prediction function $f : \mathcal{X} \to \mathcal{Y}$ is given by

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

1. For a training set $(x_1, y_1), \ldots (x_n, y_n)$, let $J(w)$ be the ℓ_2-regularized empirical risk function for the multiclass hinge loss. We can write this as

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

We will now show that $J(w)$ is a convex function of w. Justify each of the following steps. As we've shown it in a previous problem, you may use the fact that $w \mapsto \max_{y \in \mathcal{Y}} \left[\Delta\left(y_i, y\right) + \langle w, \Psi(x_i, y) - \Psi(x_i, y_i) \rangle \right]$ is a convex function.

(a) $\frac{1}{n} \sum_{i=1}^{n} \max_{y \in \mathcal{Y}} \left[\Delta\left(y_i, y\right) + \langle w, \Psi(x_i, y) - \Psi(x_i, y_i) \rangle \right]$ is a convex function of w.

(b) $\|w\|^2$ is a convex function of w.

(c) $J(w)$ is a convex function of w.

Answer:
(a) Since $w \mapsto \max_{y \in \mathcal{Y}} \left[\Delta\left(y_i, y\right) + \langle w, \Psi(x_i, y) - \Psi(x_i, y_i) \rangle \right]$ is a convex function. The summation of convex functions is convex. Thus $\frac{1}{n} \sum_{i=1}^{n} \max_{y \in \mathcal{Y}} \left[\Delta\left(y_i, y\right) + \langle w, \Psi(x_i, y) - \Psi(x_i, y_i) \rangle \right]$ is a convex function of w.
(b) Since $\|w\|$ is a convex function of w. Then $\|w\|^2$ is a composition of convex function, which is convex.
(c) By part (a) and (b) we have $J(w)$ is a summation of convex function, which 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 $\hat{y}_i = \arg\max_{y \in \mathcal{Y}} \left[\Delta\left(y_i, y\right) + \langle w, \Psi(x_i, y) - \Psi(x_i, y_i)\rangle\right]$.)

Answer:
Let $\hat{y}_i = \arg\max_{y \in \mathcal{Y}} \left[\Delta\left(y_i, y\right) + \langle w, \Psi(x_i, y) - \Psi(x_i, y_i)\rangle\right]$.
Claim: $g(w) = 2\lambda w^T + \sum\limits_{i=1}^{n} \left[(\Psi(x, y) - \Psi(x, y_i))\right]$ is a subgradient of $J(w)$.
Proof:

$$
\begin{aligned}
J(w+v) &= \lambda\|w+v\|^2 + \frac{1}{n}\sum_{i=1}^{n} \max_{y \in \mathcal{Y}} \left[\Delta\left(y_i, y\right) + \langle w+v, \Psi(x_i, y) - \Psi(x_i, y_i)\rangle\right] \\
&= \lambda\|w\|^2 + \lambda\|v\|^2 + 2\lambda w^T v + \frac{1}{n}\sum_{i=1}^{n} \left[\Delta\left(y_i, \hat{y}_i\right) + \langle w, \Psi(x_i, \hat{y}_i) - \Psi(x_i, y_i)\rangle\right] \\
&+ \frac{1}{n}\left[\sum_{i=1}^{n} \langle v, \Psi(x_i, \hat{y}_i) - \Psi(x_i, y_i)\rangle\right] \\
&= J(w) + gv + \|v\|^2 \\
&\geq J(w) + gv
\end{aligned}
$$

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

Answer: $2\lambda w^T + [(\Psi(x,y) - \Psi(x,y_i))]$

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

Answer: $2\lambda w^T + \sum_{i=1}^{m-1} [(\Psi(x, y) - \Psi(x, y_i))]$

4 [OPTIONAL] Another Formulation of Generalized Hinge Loss

In lecture we defined the **margin** of the compatibility score function h on the ith example (x_i, y_i) for class y as

$$m_{i,y}(h) = h(x_i, y_i) - h(x_i, y),$$

and the loss on an individual example (x_i, y_i) to be:

$$\max_{y} \left([\Delta(y_i, y) - m_{i,y}(h)]_+ \right).$$

Here we investigate whether this is just an instance of the generalized hinge loss $\ell\left(h, (x, y)\right)$ defined above.

1. Show that $\ell\left(h, (x_i, y_i)\right) = \max_{y' \in \mathcal{Y}} \left[\Delta\left(y_i, y'\right) - m_{i,y'}(h)\right]$. (In other words, it looks just like loss above, but without the positive part.)
2. Suppose $\Delta\left(y, y'\right) \geq 0$ for all $y, y' \in \mathcal{Y}$. Show that for any example (x_i, y_i) and any score function h, the multiclass hinge loss we gave in lecture and the generalized hinge loss presented above are equivalent, in the sense that

   $$\max_{y \in \mathcal{Y}} \left([\Delta(y_i, y) - m_{i,y}(h)]_+ \right) = \max_{y \in \mathcal{Y}} \left(\Delta(y_i, y) - m_{i,y}(h) \right).$$

   (Hint: This is easy by piecing together other results we have already attained regarding the relationship between ℓ and Δ.)
3. In the context of the generalized hinge loss, $\Delta(y, y')$ is like the "target margin" between the score for true class y and the score for class y'. Suppose that our prediction function f gets the correct class on x_i. That is, $f(x_i) = \arg\max_{y' \in \mathcal{Y}} h(x_i, y') = y_i$. Furthermore, assume that all of our target margins are reached or exceeded. That is

   $$m_{i,y}(h) = h(x_i, y_i) - h(x_i, y) \geq \Delta(y_i, y),$$

   for all $y \neq y_i$. It seems like in this case, we should have 0 loss. This is almost the case. Show that $\ell\left(h, (x_i, y_i)\right) = 0$ if we assume that $\Delta\left(y, y\right) = 0$ for all $y \in \mathcal{Y}$.

5 [OPTIONAL] Hinge Loss is a Special Case of Generalized Hinge Loss

Let $\mathcal{Y} = \{-1, 1\}$. Let $\Delta(y, \hat{y}) = 1(y \neq \hat{y})$. 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:

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