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

April 11, 2017

1 Introduction

2 Convex Surrogate Loss Functions

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

(a) For any example $(x, y) \in X \times \{-1, 1\}$, show that $1(y \neq sign(f(x)) \leq max\{0, 1 - yf(x)\}$.

ANSWER

If $y \neq sign(f(x))$, $yf(x) \leq 0$, and $1 - yf(x) \geq 1$ therefore, the inequality holds,

If $y = sign(f(x))$, $lhs = 0$ and $rhs \geq 0$ therefore, the inequality holds,

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

ANSWER

$f_1(x) = 0$, $f_2(x) = 1 - m$ are convex, so according to the result given their pointwise maximum $f(x) = max\{0, 1 - m\}$ is also convex.

(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 $max\{0, 1 - yw^T x\}$. Show that this is a convex function of w.

ANSWER

$f_w(x)$ is an affine function and is a convex function of w. Similarly, $1 - yw^T x$ is also a convex function as it is affine, so the hinge loss $max\{0, 1 - yw^T x\}$ is also a convex function as it is the pointwise maximum of 2 convex functions.

2.2 Multiclass Hinge Loss

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

ANSWER

For any $x \in X$ and $y \in Y$,

$h(x, f(x)) = max_{y \in \mathcal{Y}}(h(x, y))$

So, by definition $h(x, f(x)) \geq (h(x, y))$

(2) Justify the following two inequalities:

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

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

$$\ell(h, (x, y)) = \max_{y_0 \in \mathcal{Y}}[\Delta(y, y_0)) + h(x, y_0) - h(x, y)]$$

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

$$\ell(h, (x, y)) \ge \Delta(y, f(x)),$$

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 $\ell(f, (x, y))$.]

ANSWER

Using the solution from the previous part,
$h(x, f(x)) \ge (h(x, y))$
$h(x, f(x)) - h(x, y) \ge 0$

$\therefore \Delta(y, f(x)) + h(x, f(x)) - h(x, y) \ge \Delta(y, f(x))$

In the second inequality we are replacing f(x) with y' which would maximize the expression, so it can be written as,
$\Delta(y, f(x)) + h(x, f(x)) - h(x, y) \le \max_{f \in \mathcal{F}}[\Delta(y, f(x)) + h(x, f(x)) - h(x, y)]$

$\le \max_{y' \in \mathcal{Y}}[\Delta(y, y')) + h(x, y') - h(x, y)]$

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

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

ANSWER

$\ell(h_w, (x_i, y_i)) = \max_{y \in \mathcal{Y}} [\Delta(y_i, y) + h(x_i, y) - h(x_i, y_i)]$

Now since $h_w(x,y) = \langle w, \Psi(x,y)$
$= \max_{y \in \mathcal{Y}} [\Delta(y_i, y) + \langle w, \Psi(x_i, y)\rangle - \langle w, \Psi(x_i, y_i)\rangle]$

Using the linearity property of inner product,
$= \max_{y \in \mathcal{Y}} [\Delta(y_i, y) + \langle w, \Psi(x_i, y) - \Psi(x_i, y_i)\rangle]$

(4) We will now show that the generalized hinge loss $\ell(h_w, (x_i, y_i))$ 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 $\max_{y \in \mathcal{Y}}[\Delta(y_i, y) + \langle w, \Psi(x_i, y) - \Psi(x_i, y_i)\rangle]$ is a convex function of w.

ANSWER

(a)

Since both $\Delta(y_i, y)$and $\Psi(x_i, y) - \Psi(x_i, y_i)$ are constant with respect to w, it would be affine.

(b)

Using the results from the previous part,
$\forall y \in \mathcal{Y}, \Delta(y_i, y) + \langle w, \Psi(x_i, y) - \Psi(x_i, y_i\rangle$ is affine and convex.

$\therefore \max_{y \in \mathcal{Y}}[\Delta(y_i, y) + \langle w, \Psi(x_i, y) - \Psi(x_i, y_i)\rangle]$ is convex.

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

ANSWER

Since,
$\ell(h_w, (x_i, y_i)) \geq \Delta(y, f(x))$

We proved in the last part that $\ell(h_w, (x_i, y_i))$ is convex,

$\therefore$ $\ell(h_w, (x_i, y_i))$ is the convex surrogate for $\Delta(y, f(x))$

3 SGD for Multiclass SVM

3.1 Question 1

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

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

We will now show that 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 $w \mapsto \max_{y \in \mathcal{Y}} [\Delta(y_i, y)) + \langle w, \Psi(x_i, y) - \Psi(x_i, y_i) \rangle]$ is a convex function.

(a) $\frac{1}{n} \sum_{i=1}^{n} \max_{y \in \mathcal{Y}} [\Delta(y_i, y)) + \langle w, \Psi(x_i, y) - \Psi(x_i, y_i) \rangle]$ 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) Let $f(w) = \max_{y \in \mathcal{Y}} [\Delta(y_i, y)) + \langle w, \Psi(x_i, y) - \Psi(x_i, y_i) \rangle]$

We know that $f(w)$ is convex, so all the functions in the summation are convex. And the sum of convex functions is also convex.

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

(b)

$\|w\|^2 = w^T w$
$\nabla \|w\|^2 = 2$ and therefore $\|w\|^2$ is a convex function of w.

(c)

Using the last 2 parts,

Both $\lambda_k \|w_k\|^2$ and $\frac{1}{n}\sum_{i=1}^{n} \max_{y \in \mathcal{Y}} [\Delta(y_i, y)) + \langle w, \Psi(x_i, y) - \Psi(x_i, y_i) \rangle]$ are convex,

Therefore, $J(w)$ is a convex function of w.

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

ANSWER

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

$\partial J(w) = 2\lambda w + \partial [\frac{1}{n} \sum_{i=1}^{n} \Delta(y_i, \hat{y}) + \langle w, \Psi(x_i, \hat{y}) \rangle - \langle w, \Psi(x_i, y_i) \rangle)]$

Using the linearity property of inner products,
$= 2\lambda w + \frac{1}{n} \sum_{i=1}^{n} \partial \langle w, (\Psi(x_i, \hat{y}) - \Psi(x_i, y_i)) \rangle$

$\partial J(w) = 2\lambda w + \frac{1}{n} \sum_{i=1}^{n} (\Psi(x_i, \hat{y}) - \Psi(x_i, y_i))$

3.3 Question 3

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

ANSWER

At point (x_i, y_i) the subgradient can be written as,
g = $2\lambda w + \frac{1}{n}\sum_{i=1}^{n}(\Psi(x_i, \hat{y}) - \Psi(x_i, y_i))$

For updating w, it can be expressed as
$w_t = w_{t-1} - \eta_t \nabla_w J(w_{t-1})$

$= w_{t-1} - \eta_t g_{t-1}$

3.4 Question 4

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

ANSWER

The minibatch subgradient can be written as,
g = $2\lambda w + \frac{1}{m}\sum_{j=i}^{i+m-1}(\Psi(x_j, \hat{y}) - \Psi(x_j, y_j))$

For the update on w, it can be expressed as,

$w_{t-1} - \eta_t(2\lambda w_{t-1} + \frac{1}{n}\sum_{j=1}^{m}\sum_{i=1}^{n}(\Psi(x_{i+j-1}, \hat{y}) - \Psi(x_{i+j-1}, y_{i+j-1})))$

4 Another Formulation of Generalized Hinge Loss

4.1 Question 1

Show that $\ell(h,(x_i,y_i)) = \max_{y'\in\mathcal{Y}}[\Delta(y_i,y')) - m_{i,y'}(h)]$.

ANSWER

The generalized hinge loss is

$$\ell(h,(x_i,y_i)) = \max_{y'\in\mathcal{Y}}[\Delta(y_i,y') + h(x_i,y') - h(x_i,y_i)]$$

$$= \max_{y'\in\mathcal{Y}}[\Delta(y_i,y')) - m_{i,y'}(h)]$$

4.2 Question 2

Suppose $\Delta(y, y_0) \geq 0$ for all $y, y_0 \in \mathcal{Y}$. 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 presened above are equivalent, in the sense that

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

(Hint: This is easy by piecing together other results we have already attained regarding the relationship between ℓ and Δ.)

ANSWER

$\Delta(y_i, y) - m_{i,y'} \geq \Delta(y_i, y) \geq 0$

Since the term in the maximum is non-negative as shown above,

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

4.3 Question 3

In the context of the generalized hinge loss, $\Delta(y, y_0)$ is like the "target margin" between the score for true class y and the score for class y_0. Suppose that our prediction function f gets the correct class on x_i. That is, $f(x_i) = arg \max_{y_0 \in \mathcal{Y}} h(x_i, y_0) = y_i$. 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 $y \neq y_i$. Show that $\ell(h, (x_i, y_i)) = 0$ if we assume that $\Delta(y, y) = 0$ for all $y \in \mathcal{Y}$.

ANSWER

Using the results from the previous problem

$\ell(h, (x_i, y_i)) = \max_{y \in \mathcal{Y}}(\Delta(y_i, y) - m_{i,y'}(h)))$

Since $m_{i,y}(h) = h(x_i, y_i) - h(x_i, y) \geq \Delta(y_i, y)$,
the above expression for loss would be maximum when $y = y_i$, in all the other cases it would be negative.

$\therefore \ell(h, (x_i, y_i)) = \Delta(y_i, y_i)$

$= 0$

5 Hinge Loss is a Special Case of Generalized Hinge Loss

Let $Y = \{-1, 1\}$. 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

$$h(x, 1) = g(x)/2$$

$$h(x, -1) = -g(x)/2.$$

Show that for this choice of h, the multiclass hinge loss reduces to hinge loss: $\ell(h, (x, y)) = \max_{y_0 \in \mathcal{Y}} [\Delta(y, y_0)) + h(x, y_0) - h(x, y)] = max\{0, 1 - yg(x)\}$

ANSWER

If $y = y_0$,

$\ell(h, (x, y)) = \Delta(y_0, y_0) + h(x, y_0) - h(x, y_0)$

$= 0$

If $y \neq y_0$,

$\ell(h, (x, y)) = \Delta(y, y_0)) + h(x, y_0) - h(x, y)$

=$1(y \neq y_0) + 1/2(-g(x) - g(x)$ (case y = 1) or $1/2(g(x) + g(x))$ (case y=-1)

= $1 + (-g(x))$(case y = 1) or $g(x)(casey = -1)$

$= 1 - yg(x)$

$\therefore, \ell(h(x, y)) = \max\{0, 1 - yg(x)\}$