Assignment 6

Benjamin Jakubowski

April 11, 2016

2. Convex Surrogate Loss Functions

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

Our objective in this section is to show the hinge loss is a convex surrogate for 0/1 loss. Recall a convex surrogate loss function is a convex function that is an upper bound for the loss function of interest. To show hinge loss is a convex surrogate for the 0/1 loss, we first show it is an upper bound on the 0-1 loss, and then show it is convex.

2.1.a Show $1(y \neq \text{sign}(f(x)) \leq \max\{0, 1 - yf(x)\}$

We first show for any example $(x, y) \in \mathcal{X} \times \{-1, 1\}$, $1(y \neq \text{sign}(f(x)) \leq \max\{0, 1-yf(x)\}$. We do this through enumeration:

y	$f(x)$	$1(y \neq \text{sign}(f(x))$	$\max\{0, 1 - yf(x)\}$
-1	>0	1	$1 - yf(x) = 1 + f(x) > 1$ (since $f(x) > 0$)
-1	0	1	$1 + (-1)0 = 1 \geq 1$
-1	<0	0	≥ 0 (by definition)
1	>0	0	≥ 0 (by definition)
1	0	1	$1 - (1)0 = 1 \geq 1$
1	<0	1	$1 - (1)(f(x)) = 1 - f(x) > 1$ (since $f(x) < 0$)

2.1.b Show hinge loss is a convex function of the margin m

We next show the hinge loss $\max\{0, 1 - m\}$ is a convex function of the margin m. First, recall

$$f(x) = \max\{f_1(x), \cdots, f_n(x)\}$$

is convex if the f_i's are convex. Since $f_1(m) = 0$ and $f_2(m) = 1 - m$ are convex (simply note $f''(m) \geq 0$ for all m for both functions), $f(m) = \max\{0, 1 - m\}$ is convex.

2.1.c Hinge loss a convex function of w

Now let our prediction score function be $f_w(x) = w^T x$. Then the hinge loss of f_w on an example (x, y) is $\max\{0, 1 - yw^T yx\}$. This is a convex function of w, since:

- The hinge loss is a convex function of $m = f(x)$.
- $f_w(x)$ is an affine function
- The composition of an affine function with a convex function is convex.

Thus, we've shown the hinge loss is an upper bound on the 0-1 loss, and it is convex: thus it is a convex surrogate loss function for the 0-1 loss.

2.2 Multiclass Hinge Loss

Now consider the multiclass problem- let

- $\mathcal{Y} = \{1, \cdots, k\}$
- $\mathcal{H} = \{h : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}\}$
- $\mathcal{F} = \{f(x) = \arg\max_{y \in \mathcal{Y}} h(x, y) | h \in \mathcal{H}\}$

Note $\mathcal{Y}$ is our output space, $\mathcal{H}$ is our base hypothesis space, and $\mathcal{F}$ is our final multiclass hypothesis space. Additionally, note our action space $\mathcal{A} = \mathcal{Y}$.

Now suppose we have a multiclass loss function $\Delta : \mathcal{Y} \times \mathcal{A} \to \mathbb{R}$. Our goal is to find $f \in \mathcal{F}$ minimizing the empirical class-sensitive loss:

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

We can't use $0/1$ loss, since we know it is computation intractable- thus we're looking for a convex surrogate loss function. We show the generalized hinge loss is such a function.

2.2.1 Show $h(x, y) \leq h(x, f(x))$

Suppose we have chosen an $h \in \mathcal{H}$, from which we get $f(x) = \arg\max_{y \in \mathcal{Y}} h(x, y)$. We use contradiction to show that for any $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ we have

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

Suppose not. Then, for some $(x', y') \in \mathcal{X} \times \mathcal{Y}$,

$$h(x', y') > h(x', f(x'))$$

Then, let

$$f^*(x) = \begin{cases} f(x) & \text{for all } x \neq x' \\ y' & \text{for } x = x' \end{cases}$$

Since $h(x', f^*(x')) = h(x', y') > h(x', f(x'))$, f is not the arg max. This is a contradiction, so we conclude

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

for all $x, y \in \mathcal{X} \times \mathcal{Y}$.

2.2.2 Justifying inequalities to derive the generalized hinge loss

We next demonstrate two inequalities in order to derive the generalized hinge loss. First, from above, we have

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

Adding $\Delta(y, f(x))$ to both sides gives

$$\Delta(y, f(x)) + h(x, y) \leq \Delta(y, f(x)) + h(x, f(x))$$

Subtracting off $h(x, y)$ yields the first desired inequality:

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

Now, recall $f(x) \in \mathcal{Y} = \mathcal{A}$. Thus,

$$[\Delta(y, f(x)) + h(x, f(x)) - h(x, y)] \in \{\Delta(y, y') + h(x, y') - h(x, y) | y' \in Y\}$$

Hence, we have

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

Now, for future reference, note we define the RHS as the generalized hinge loss:

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

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

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

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

2.2.3 Generalized hinge loss in $h_w(x, y)$

Now, let

- $\Psi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^d$ be a class-sensitive feature mapping
- $\mathcal{H} = \{h_w(x, y) = \langle w, \Psi(x, y) \rangle | w \in \mathbb{R}^d\}$

We show now rewrite the generalized hinge loss for $h_w(x, y)$ on example (x_i, y_i):

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

Note the first two steps are just substitution, and the last step follows from the linearity of inner products.
Next, we proceed to show the generalized hinge loss $\ell(h_w, (x_i, y_i))$ is a convex function of w in three steps

2.2.4A $\Delta(y_i, y) + \langle w, \Psi(x_i, y) - \Psi(x_i, y_i) \rangle$ AFFINE FUNCTION OF w

Consider $\Delta(y_i, y) + \langle w, \Psi(x_i, y) - \Psi(x_i, y_i) \rangle$. Note this is an affine function of w by definition, since it's just an inner product with w plus a scalar.

2.2.4B $\max_{y \in \mathcal{Y}} [\Delta(y_i, y) + \langle w, \Psi(x_i, y) - \Psi(x_i, y_i) \rangle]$ CONVEX FUNCTION OF w

Next, note $\max_{y \in \mathcal{Y}} [\Delta(y_i, y) + \langle w, \Psi(x_i, y) - \Psi(x_i, y_i) \rangle]$ is a convex function of w since its the pointwise maximum of convex (affine) functions.

2.2.5 $\ell(h_w, (x_i, y_i))$ IS A CONVEX SURROGATE FOR $\Delta(y_i, f_w(x_i))$

Finally we conclude that $\ell(h_w, (x_i, y_i))$ is a convex surrogate for $\Delta(y_i, f_w(x_i))$ since

- $\ell(h_w, (x_i, y_i))$ is convex
- $\ell(h_w, (x_i, y_i)) \geq \Delta(y_i, f_w(x_i)$

3. Hinge Loss is a Special Case of Generalized Hinge Loss

We now show that hinge loss is a special case of generalized hinge loss.
Let

- $\mathcal{Y} = \{-1, 1\}$
- $\Delta(y, \hat{y}) = 1(y \neq \hat{y})$
- Our compatibility function be:

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

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

We now show for this choice of h, multiclass hinge loss reduce to hinge loss. We do so by considering two cases, $y = 1$ and $y = -1$, each with two subcases, $y' = 1$ and $y' = -1$.

- **y = -1**
  - **y' = -1**:

$$\begin{aligned}\Delta(y, y') + h(x, y') - h(x, y) &= \Delta(-1, -1) + h(x, -1) - h(x, -1) \\ &= 0 + (-g(x)/2) - (-g(x)/2) \\ &= 0\end{aligned}$$

  - **y' = 1**:

$$\begin{aligned}\Delta(y, y') + h(x, y') - h(x, y) &= \Delta(-1, 1) + h(x, 1) - h(x, -1) \\ &= 1 + (g(x)/2) - (-g(x)/2) \\ &= 1 + 2(g(x)/2) = 1 + g(x) = 1 - yg(x)\end{aligned}$$

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

- **y = 1**
  - **y' = -1**:

$$\begin{aligned}\Delta(y, y') + h(x, y') - h(x, y) &= \Delta(1, -1) + h(x, -1) - h(x, 1) \\ &= 1 + (-g(x)/2) - (g(x)/2) \\ &= 1 - 2(g(x)/2) = 1 - g(x) = 1 - yg(x)\end{aligned}$$

  - **y' = 1**:

$$\begin{aligned}\Delta(y, y') + h(x, y') - h(x, y) &= \Delta(1, 1) + h(x, 1) - h(x, 1) \\ &= 0 + (g(x)/2) - (g(x)/2) \\ &= 0\end{aligned}$$

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

4. Another formulation of Generalized Hinge Loss

In this section we investigate whether the margin loss defined in class is is just an instance of the generalized hinge loss. Recall we defined the margin of the compatibility score function 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:

$$\max_y[(\Delta(y_i, y) - m_{i,y}(h))_+]$$

4.1 Rewriting general hinge loss using m

First, note simple substitution yields:

$$\begin{aligned}\ell(h,(x_i,y_i)) &= \max_{y'\in\mathcal{Y}}[\Delta(y_i,y') + h(x_i,y') - h(x_i,y_i)] \\ &= \max_{y'\in\mathcal{Y}}[\Delta(y_i,y') - (h(x_i,y_i) - h(x_i,y'))] \\ &= \max_{y'\in\mathcal{Y}}[\Delta(y_i,y') - m_{i,y'}(h)]\end{aligned}$$

4.2 Show multiclass hinge loss and generalized hinge loss are equivalent

Suppose $\Delta(y,y') \geq 0$ for all $y, y' \in \mathcal{Y}$.

From 4.1, we know

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

Now, from 2.2, we know $\ell(h,(x_i,y_i)) \geq \Delta(y_i,f(x))$, where $f(x) = \arg\max_{y\in\mathcal{Y}} h(x,y)$. Since $\Delta(y,y') \geq 0$ for all $y, y' \in \mathcal{Y}$, and $f(x) \in \mathcal{Y}$, we know $\Delta(y_i,f(x)) \geq 0$. Thus,

$$\ell(h,(x_i,y_i)) \geq \Delta(y_i,f(x)) \geq 0$$

Thus

$$\ell(h,(x_i,y_i)) = \max_{y'\in\mathcal{Y}}[\Delta(y_i,y') - m_{i,y'}(h)] \geq 0$$

Now, noting that the $(\cdot)_+$ function is

$$(x)_+ = \begin{cases} x & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases}$$

it is clear that (given the maximum is positive),

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

4.3 Show $\ell(h,(x_i,y_i)) = 0$ if $\Delta(y,y) = 0$ for all $y \in \mathcal{Y}$

First, we assume

1. $f(x_i) = \arg\max_{y'\in\mathcal{Y}} h(x_i,y') = y_i$ (i.e. we predict correctly).
2. $m_{i,y}(h) = h(x_i,y_i) - h(x_i,y) \geq \Delta(y_i,y)$ for all $y \neq y_i$.
3. $\Delta(y,y) = 0$ for all $y \in \mathcal{Y}$.

Now, recall (from 4.3)

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

Now we consider two cases: $y = y_i$ and $y \neq y_i$.

- **Case 1**: If $y = y_i$, then

$$\begin{aligned} & \Delta(y_i,y) = \Delta(y_i,y_i) = 0 \text{ (by assumption 3)} \\ & m_{i,y}(h) = h(x_i,y_i) - h(x_i,y) = h(x_i,y_i) - h(x_i,y_i) = 0 \\ \implies \quad & (\Delta(y_i,y) - m_{i,y}(h))_+ = 0 \end{aligned}$$

- **Case 2**: If $y \neq y_i$, then

$$\begin{aligned} & (\Delta(y_i,y) - m_{i,y}(h)) \leq 0 \text{ (by assumption 2)} \\ \implies \quad & (\Delta(y_i,y) - m_{i,y}(h))_+ = 0 \end{aligned}$$

Thus,

$$\ell(h,(x_i,y_i)) = \max_{y' \in \mathcal{Y}} \Big[\big(\Delta(y_i,y') - m_{i,y'}(h)\big)_+ \Big] = \max\{0,\cdots,0\} = 0$$

5. SGD for Multiclass SVM

5.1 Showing $J(w)$ is a convex function of w

Now suppose we have

- Outcome and action space: $\mathcal{Y} = \mathcal{A} = \{1,\cdots,k\}$
- Class-sensitive loss function $\Delta : \mathcal{Y} \times \mathcal{A} \to \mathbb{R}^{\geq 0}$
- Class-sensitive feature mapping $\Psi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^d$
- Prediction function $f : \mathcal{X} \to \mathcal{Y}$ given by

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

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:

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

We proceed to show $J(w)$ is convex in three steps:

5.1.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

$\frac{1}{n}\sum_{i=1}^{n}\max_{y\in\mathcal{Y}}[\Delta(y_i,y)+\langle w,\Psi(x_i,y)-\Psi(x_i,y_i)\rangle]$ is a convex function of w since it is a non-negative weighted sum of convex functions (see 2.2.4.a)

5.1.B $||w||^2$ IS A CONVEX FUNCTION OF w

$||w||^2$ is a convex function of w since[1]

- $g(x)^p$ is convex in x for $p \geq 1$ (here $p = 2$) if $g(x)$ is convex and non-negative, and
- Norms are convex (and non-negative)

5.1.C $J(w)$ IS A CONVEX FUNCTION OF w

Finally, by $\lambda > 0$ we again have a non-negative weighted sum of convex functions, so $J(w)$ is a convex function of w.

5.2 SUBGRADIENT OF $J(w)$

Since $J(w)$ is convex, it has a subgradient at every point. We find an expression for a subgradient of $J(w)$.
First, let $\hat{y}_i = \arg\max_{y\in\mathcal{Y}}[\Delta(y_i,y)+\langle w,\Psi(x_i,y)-\Psi(x_i,y_i)\rangle]$
Then, from homework 3 question 2.1, we know if

$$g \in \partial[\Delta(y_i,\hat{y}_i)+\langle w,\Psi(x_i,\hat{y}_i)-\Psi(x_i,y_i)\rangle]$$

then

$$g \in \partial[\ell(h_w,(x_i,y_i))]$$

where, as previously,

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

Next, note

$$(\Psi(x_i,\hat{y}_i)-\Psi(x_i,y_i)) \in \partial[\Delta(y_i,\hat{y}_i)+\langle w,\Psi(x_i,\hat{y}_i)-\Psi(x_i,y_i)\rangle]$$

Before proceeding, recall if

$$f = f_1 + \cdots + f_m$$
$$\partial f(x) = \partial f_1(x) + \cdots + \partial f_m(x)$$

This implies

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

[1] See `https://davidrosenberg.github.io/mlcourse/Notes/convex-optimization.pdf`

5.3 stochastic subgradient of $J(w)$ based on the point (x_i, y_i)

First, note

$$\begin{aligned} J(w) &= \lambda\Omega(w) + \frac{1}{n}\sum_{i=1}^{n} \ell(h_w, (x_i, y_i)) \\ &= \frac{1}{n}\sum_{i=1}^{n} g_i(w) \end{aligned}$$

where

$$g_i(w) = \lambda\Omega(w) + \ell(h_w, (x_i, y_i))$$

Then our stochastic subgradient based on the point (x_i, y_i) is just

$$2\lambda w + (\Psi(x_i, \hat{y}_i) - \Psi(x_i, y_i)) \in \partial[\lambda\Omega(w) + \ell(h_w, (x_i, y_i))] = \partial g_i(w)$$

5.4 Minibatch subgradient of $J(w)$ based on the point (x_i, y_i)

Now, we're finding a minibatch subgradient based on the points $(x_i, y_i), \cdots, (x_{i+m-1}, y_{i+m-1})$. Thus, our subgradient is

$$\left[2\lambda w + \frac{1}{m}\sum_{k=0}^{m-1} (\Psi(x_{i+k}, \hat{y}_{i+k}) - \Psi(x_{i+k}, y_{i+k}))\right] \in \partial J_{minibatch}(w)$$