DS-GA 1003: Homework 6
Generalized Hinge Loss and Multiclass SVM

Due on Monday, April 11, 2016

Professor David Ronsenberg

See complete code at: *git@github.com:cryanzpj/1003.git*

Yuhao Zhao
Yz3085

2. Convex Surrogate Loss Function

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

– 2.1.1.

i) if $y \neq sign(f(x)), yf(x) \leq 0, 1 - yf(x) \geq 1 = I(y \neq sign(f(x)))$.
Therefore, $I(y \neq sign(f(x))) \leq max\{0, 1 - yf(x)\}$
ii) if $y = sign(f(x)), I(y \neq sign(f(x))) = 0, max\{0, 1 - yf(x)\} \geq 0$.
Therefore, $I(y \neq sign(f(x))) \leq max\{0, 1 - yf(x)\}$

– 2.1.2.

Since $f_1(m) = 0$ is convex , $f_2(m) = 1 - m$ is an affine function and thus convex.
Since the point-wise maximum of convex functions is also convex, $max\{f_1, f_2\} = max\{0, 1 - m\}$ is a convex function of the margin m.

– 2.1.3.

$f_1(m) = 0$ is convex, we need to show $f_2 = 1 - yw^T x$ is convex.
It's sufficient to show f_2 is affine. For $\forall w_1, w_2$ and $\alpha \in [0, 1]$:

$$\begin{aligned}
\alpha f_2(w_1) + (1-\alpha) f_2(w_2) &= \alpha(1 - yw_1^T x) + (1-\alpha)(1 - yw_2^T x) && (1)\\
&= 1 - y(\alpha w_1^T + (1-\alpha) w_2^T)x && (2)\\
&= f(\alpha w_1 + (1-\alpha) w_2 && (3)
\end{aligned}$$

Therefore f_2 is affine w.r.t w and thus convex, and $max\{f_1, f_2\} = max\{0, 1 - yw^T x\}$ is a convex function of w.

• 2.2 Multiclass Hinge Loss.

– 2.2.1.

Since $f(x) = \underset{y \in \mathcal{Y}}{\operatorname{argmax}}\ h(x, y)$, f(x) is the y that max h, by definition:

$$h(x, f(x)) \leq h(x, y) \qquad \text{for } \forall x \in \mathcal{X}, y \in \mathcal{Y} \tag{4}$$

– 2.2.2.

Since from 2.2.1, $h(x, f(x)) - h(x, y) \leq 0$:

$$\begin{aligned}
\Delta(y, f(x)) &\leq \Delta(y, f(x)) + h(x, f(x)) - h(x, y) && (5)\\
&\leq \max_{y' \in \mathcal{Y}} \Delta(y, y') + h(x, y') - h(x, y) && (6)\\
&= \ell(h, (x, y)) && (7)
\end{aligned}$$

– 2.2.3.

For $\mathcal{H} = \{h_w(w, \Psi(x, y)) | w \in R^d\}$:

$$\begin{aligned}
\ell(h_w, (x_i, y_i)) &= \max_{y \in \mathcal{Y}} \Delta(y_i, y) + h_w(x_i, y) - h_w(x_i, y_i) && (8)\\
&= \max_{y \in \mathcal{Y}} \Delta(y_i, y) + \langle w, \Psi(x_i, y)\rangle - \langle w, \Psi(x_i, y_i)\rangle && (9)\\
&= \max_{y \in \mathcal{Y}} \Delta(y_i, y) + \langle w, \Psi(x_i, y) - \Psi(x_i, y_i)\rangle && (10)
\end{aligned}$$

– **2.2.4.**

Let $f(w) = \Delta(y_i, y) + \langle w, \Psi(x_i, y) - \Psi(x_i, y_i)\rangle$, for $\forall w_1, w_2 \in R^d$ and $\alpha \in [0, 1]$:

$$\begin{aligned}
\alpha f(w_1) + (1-\alpha) f(w_2) =& \alpha(\Delta(y_i, y) + \langle w_1, \Psi(x_i, y) - \Psi(x_i, y_i)\rangle && (11)\\
&+ (1-\alpha)(\Delta(y_i, y) + \langle w_2, \Psi(x_i, y) - \Psi(x_i, y_i)\rangle && (12)\\
&= \Delta(y_i, y) + \langle \alpha w_1, \Psi(x_i, y) - \Psi(x_i, y_i)\rangle + \langle (1-\alpha) w_2, \Psi(x_i, y) - \Psi(x_i, y_i)\rangle && (13)\\
&= \Delta(y_i, y) + \langle \alpha w_1 + (1-\alpha) w_2, \Psi(x_i, y) - \Psi(x_i, y_i)\rangle && (14)\\
&= f(\alpha w_1 + (1-\alpha) w_2) && (15)
\end{aligned}$$

Therefore, $f(w)$ is an affine function of w, thus convex.
let $f_j(w) = \Delta(y_i, y'_j) + \langle w, \Psi(x_i, y'_j) - \Psi(x_i, y_i)\rangle$, which $y'_j = j, j \in \{1, 2, ..., k\}$, f_j is convex:

$$\max_{y \in \mathcal{Y}} \Delta(y_i, y) + \langle w, \Psi(x_i, y) - \Psi(x_i, y_i)\rangle = max\{f_1, f_2, ..., f_k\} \qquad (16)$$

LHS is the point-wise maximum of k convex convex functions, thus is convex.

– **2.3.5.**

In 2.2.4 we showed that $\ell(h_w, (x_i, y_i))$ is convex, and in 2.2.2/2.2.3 we showed $\ell(h_w, (x_i, y_i))$ is an upper bound of $\Delta(y_i, f_w(x_i))$, which is our loss function of interest. Therefore, by definition, $\ell(h_w, (x_i, y_i))$ is a convex surrogate for $\Delta(y_i, f_w(x_i))$

3 Hinge Loss is a Special Case of Generalized Hinge Loss

$\ell(h, (x, y)) = \max_{y' \in \mathcal{Y}} \Delta(y, y') + h(x, y') - h(x, y)$, where $\Delta(y, \hat{y}) = I(y \neq \hat{y})$.
Since $h(x, y)$ in our case is constant, and $yin\{1, -1\}$:

$$\begin{aligned}
\ell(h, (x, y)) &= \max_{y' \in \mathcal{Y}} \Delta(y, y') + h(x, y') - h(x, y) && (17)\\
&= \max_{y' \in \mathcal{Y}} \Delta(y, y') + h(x, y') && (18)\\
&= max\{I(y \neq 1) + h(x, 1), I(y \neq -1) + h(x, -1)\} && (19)\\
&= max\{I(y \neq 1) + \frac{g(x)}{2}, I(y \neq -1) - \frac{g(x)}{2}\} && (20)
\end{aligned}$$

If $y = 1$, eqn(20) becomes:

$$max\{\frac{g(x)}{2}, 1 - \frac{g(x)}{2}\} = max\{0, 1 - g(x)\} = max\{0, 1 - yg(x)\} \qquad (21)$$

if $y = -1$, eqn(20) becomes:

$$max\{1 + \frac{g(x)}{2}, -\frac{g(x)}{2}\} = max\{1 + g(x), 0\} = max\{1 - yg(x), 0\} \qquad (22)$$

Therefore, in binary case:

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

4 Another Formulation of Generalized Hinge Loss

– **4.1.**

$$\ell(h,(x_i,y_i)) = \max_{y'\in\mathcal{Y}} \Delta(y_i,y') + h(x_i,y') - h(x_i,y_i) \tag{23}$$
$$= \max_{y'\in\mathcal{Y}} \Delta(y_i,y') - (h(x_i,y_i) - h(x_i,y')) \tag{24}$$
$$= \max_{y'\in\mathcal{Y}} \Delta(y_i,y') - m_{i,y'}(h) \tag{25}$$

– **4.2.**

From 2.2.2 we know that $\ell(h,(x,y)) \geq \Delta(y,f(x))$. If we assume $\Delta(y,y') \geq 0$ for $\forall y,y' \in \mathcal{Y}$:

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

Thus, $(\Delta(y_i,y) - m_{i,y}(h),0)_+ = max\{\Delta(y_i,y) - m_{i,y}(h),0\} = \Delta(y_i,y) - m_{i,y}(h)$
Therefore:

$$\max_{y\in\mathcal{Y}} (\Delta(y_i,y) - m_{i,y}(h),0)_+ = \max_{y\in\mathcal{Y}} (\Delta(y_i,y) - m_{i,y}(h))$$

– **4.3.**

We assume that $m_{i,y}(h) = h(x_i,y_i) - h(x_i,y) \geq \Delta(y_i,y), \forall y \neq y_i$ and $\Delta(y,y) = 0$
For $\forall y_j \neq y_i$:

$$\Delta(y_i,y_j) - m_{i,y_j}(h) \leq 0 \tag{26}$$
$$(\Delta(y_i,y_j) - m_{i,y_j}(h))_+ = 0 \tag{27}$$

For $y = y_i$:

$$\Delta(y_i,y) - m_{i,y}(h) = 0 - 0 = 0 \tag{28}$$

Therefore:

$$\ell(h,(x,y)) = \max_{y'\in\mathcal{Y}} (\Delta(y_i,y') - m_{i,y'}(h))_+ = 0 \tag{29}$$

5. SGD for Multiclass SVM

– **5.1.**

a) In 2.2.4 we should 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 funtion.
Then it's obvious that $\frac{1}{n}\sum_{i=1}^{n} \max_{y\in\mathcal{Y}} \Delta(y_i,y) + \langle w, \Psi(x_i,y) - \Psi(x_i,y_i)\rangle$ is a convex function of w since it's a constant times the sum of n convex functions.

b) Let $f(w) = ||w||^2$, for $\forall w_1, w_2 \in R^d, \alpha \in [0,1]$

$$f(\alpha w_1 + (1-\alpha)w_2) = ||\alpha w_1 + (1-\alpha)w_2||^2 \leq \alpha^2||w_1||^2 + (1-\alpha)^2||w_2||^2 \tag{30}$$

Since $\alpha, (1-\alpha) \leq 1$ we have $\alpha^2 \leq \alpha$ and $(1-\alpha)^2 \leq 1-\alpha$

$$\text{RHS of enq (30)} \leq \alpha||w_1||^2 + (1-\alpha)||w_2||^2 = \alpha f(w_1) + (1-\alpha)f(w_2)$$

Therefore, $||w||^2$ is convex function of w.

c) $J(w) = \lambda||w||^2 + \frac{1}{n}\sum_{i=1}^{n} \max_{y\in\mathcal{Y}} \Delta(y_i,y) + \langle w, \Psi(x_i,y) - \Psi(x_i,y_i)\rangle$ is convex since it's a sum of convex functions.

– 5.2.

From homework 3 Q2.1:

Suppose $f_1, ..., f_m$ are functions and f(x) = $\max_{i=1,...,m} f_i(x)$, let k be the index that $f_k(x) = f(x)$, if we choose $g \in \partial f_k(x), g \in \partial f(x)$

Let $\hat{y}_i = \operatorname*{argmax}_{y \in \mathcal{Y}} \Delta(y_i, y) + \langle w, \Psi(x_i, y) - \Psi(x_i, y_i) \rangle$, if we choose $g_i \in \partial \Delta(y_i, \hat{y}_i) + \langle w, \Psi(x_i, \hat{y}_i) - \Psi(x_i, y_i) \rangle$, we also have $g_i \in \partial \max_{y \in \mathcal{Y}} \Delta(y_i, y) + \langle w, \Psi(x_i, y) - \Psi(x_i, y_i) \rangle$
Therefore we can choose $g_i = \Psi(x_i, \hat{y}_i) - \Psi(x_i, y_i)$
One subgradient for $J(w)$ is :

$$2\lambda w + \frac{1}{n} \sum_{i=1}^{n} \Psi(x_i, \hat{y}_i) - \Psi(x_i, y_i) \tag{31}$$

– 5.3.

The stochastic subgradient based on (x_i, y_i) is:

$$2\lambda w + \Psi(x_i, \hat{y}_i) - \Psi(x_i, y_i) \tag{32}$$

– 5.4.

The minibatch subgradient based on m data points $(x_i, y_i), ..., (x_{i+m-1}, y_{i+m-1})$ is:

$$2\lambda w + \frac{1}{m} \sum_{j=0}^{m-1} \Psi(x_{i+j}, \hat{y}_{i+j}) - \Psi(x_{i+j}, y_{i+j}) \tag{33}$$