1. Reformulations of Multiclass Hinge Loss

Problem 1.2.1. Show that if $\Delta(y, y) = 0$ for all $y \in \mathcal{Y}$, then $\ell_2\left(h, (x_i, y_i)\right) = \ell_1(h, (x_i, y_i))$.

Solution Assume $\Delta(y, y) = 0$ for all $y \in \mathcal{Y}$, then

$$\begin{aligned}
\ell_2(h, (x_i, y_i)) &= \max_{y \in \mathcal{Y}} \left[\Delta\left(y_i, y\right) + h(x_i, y) - h(x_i, y_i)\right] \\
&= \max\left[\Delta\left(y_i, y_i\right) + h(x_i, y_i) - h(x_i, y_i), \max_{y \neq y_i}\left[\Delta\left(y_i, y\right) + h(x_i, y) - h(x_i, y_i)\right]\right] \\
&= \max\left[0, \max_{y \neq y_i}\left[\Delta\left(y_i, y\right) + h(x_i, y) - h(x_i, y_i)\right]\right] \\
&= \max_{y \neq y_i}\left[\max\left[0, \Delta\left(y_i, y\right) + h(x_i, y) - h(x_i, y_i)\right]\right] \\
&= \ell_1(h, (x_i, y_i))
\end{aligned}$$

Problem 1.2.2.a. Show that under the conditions above, $\ell_1(h, (x_i, y_i)) = \ell_2(h, (x_i, y_i)) = 0$.

Solution Since $\Delta(y, y) = 0$, $\ell_1 = \ell_2$.
Also, since $m_{i,y}(h) = h(x_i, y_i) - h(x_i, y) \geq \Delta(y_i, y)$

$$\Delta(y_i, y) - m_{i,y}(h) = \Delta(y_i, y) + h(x_i, y) - h(x_i, y_i) \leq 0 \qquad \forall y \neq y_i$$

Then it is clear that $\ell_1(h, (x_{i,}y_i)) = 0 = \ell_2(h, (x_{i,}y_i))$

Problem 1.2.2.b. Show that under the conditions above, we make the correct prediction on x_i. That is, $f(x_i) = \arg\max_{y \in \mathcal{Y}} h(x_i, y) = y_i$.

Solution Assume $f(x_i) = \arg\max_{y \in \mathcal{Y}} h(x_i, y) \neq y_i$.
Then $\exists\ y'$ such that $h(x_i, y') > h(x_i, y_i)$.
Then $h(x_i, y_i) - h(x_i, y) < 0$. But this contradicts the fact that

$$h(x_i, y_i) - h(x_i, y') \geq \Delta(y_i, y) > 0$$

Thus we conclude that $f(x_i) = \arg\max_{y \in \mathcal{Y}} h(x_i, y) = y_i$

2. SGD for Multiclass Linear SV

Problem 2.2. 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.

Solution

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

Problem 2.3. Give an expression for the stochastic subgradient based on the point (x_i, y_i).

Solution

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

Problem 2.4.

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

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

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

Problem 3. 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:

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

Solution Note that

$$\ell\left(h, (x, y)\right) = \max\left[\Delta\left(-1, y'\right)\right) + h(x, y') - h(x, -1), \Delta\left(1, y'\right)\right) + h(x, y') - h(x, 1)]$$

Either $y = y'$ or $y \neq y'$.
If $y = y'$, then $\ell(h, (x, y)) = 0$.
Otherwise

$$\begin{aligned} \ell(h, (x, y)) &= \Delta(y, y') + h(x, y') - h(x, y) \\ &= \begin{cases} 1 + g(x) & \text{if } y = -1 \\ 1 - g(x) & \text{if } y = 1 \end{cases} \\ &= 1 - yg(x) \end{aligned}$$

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

Gradient Boosting Machines

Problem 7.1. Consider the regression framework, where $\mathcal{Y} = \mathbf{R}$. Suppose our loss function is given by

$$\ell(\hat{y}, y) = \frac{1}{2}(\hat{y} - y)^2,$$

and at the beginning of the m'th round of gradient boosting, we have the function $f_{m-1}(x)$. Show that the h_m chosen as the next basis function is given by

$$h_m = \arg\min_{h \in \mathcal{F}} \sum_{i=1}^{n} \left[(y_i - f_{m-1}(x_i)) - h(x_i)\right]^2.$$

In other words, at each stage we find the base prediction function $h_m \in \mathcal{F}$ that is the best fit to the residuals from the previous stage.

Solution Note that

$$\begin{aligned}(g_m)_i &= \frac{\partial}{\partial f(x_i)} \sum_{j=1}^{n} \ell(y_j, f(x_j)) \\ &= \frac{\partial}{\partial f(x_i)} \frac{1}{2}(y_i - f(x_i))^2 \\ &= f(x_i) - y_i\end{aligned}$$

Thus $h_m = \arg\min_{h \in \mathcal{F}} \sum_{i=1}^{n} \left[(y_i - f_{m-1}(x_i)) - h(x_i)\right]^2$.

Problem 7.2. Now let's consider the classification framework, where $\mathcal{Y} = \{-1, 1\}$. In lecture, we noted that AdaBoost corresponds to forward stagewise additive modeling with the exponential loss, and that the exponential loss is not very robust to outliers (i.e. outliers can have a large effect on the final prediction function). Instead, let's consider the logistic loss

$$\ell(m) = \ln\left(1 + e^{-m}\right),$$

where $m = yf(x)$ is the margin. Similar to what we did in the ℓ_2-Boosting question, write an expression for h_m as an argmin over $\mathcal{F}$.

Solution Note that $\ell(y, f(x)) = \ln\left(1 + e^{-yf(x)}\right)$.
Then

$$\begin{aligned}(g_m)_i &= \frac{\partial}{\partial f(x_i)} \sum_{j=1}^{n} \ln\left(1 + e^{-y_j f(x_j)}\right) \\ &= \frac{\partial}{\partial f(x_i)} \ln\left(1 + e^{-y_i f(x_i)}\right) \\ &= \frac{-y_i e^{-y_i f(x_i)}}{1 + e^{-y_i f(x_i)}}\end{aligned}$$

Thus $h_m = \arg\min_{h \in \mathcal{F}} \sum_{i=1}^{n} \left[\frac{-y_i e^{-y_i f(x_i)}}{1+e^{-y_i f(x_i)}} - h(x_i)\right]^2$