1 Separability

Assume the two sets of points are linearly separable, but their convex hulls intersect. By linearly separable, it means there is a $\boldsymbol{\omega}$ that for all points in $\{\boldsymbol{x_n}\}$ $\boldsymbol{\omega}^T \boldsymbol{x_n} + w_0 > 0$ and for all points in $\{\boldsymbol{x'_m}\}$ $\boldsymbol{\omega}^T \boldsymbol{x'_m} + w_0 < 0$.

Assume the two convex hulls intersects at point $\mathbf{z}$, which means $\mathbf{z} = \sum_n \alpha_n \boldsymbol{x_n} = \sum_n \beta_n \boldsymbol{y_n}$.

Since $\mathbf{z}$ belongs to convex hull $\{\boldsymbol{x_n}\}$, $\boldsymbol{\omega}^T \mathbf{z} = \sum_n \alpha_n \boldsymbol{\omega}^T \boldsymbol{x_n} > \sum_n \alpha_n (-w_0) = -w_0$.

Since $\mathbf{z}$ belongs to convex hull $\{\boldsymbol{x'_m}\}$, $\boldsymbol{\omega}^T \mathbf{z} = \sum_n \beta_n \boldsymbol{\omega}^T \boldsymbol{x'_m} < \sum_n \beta_n (-w_0) = -w_0$.

Contradiction.

So, if the two sets of points are linearly separable, their convex hulls do not intersect.

2 Logistic regression and gradient descent

(a)

$$\sigma'(a) = -\frac{1}{(1+e^{-a})^2} * e^{-a} = \sigma(a)^2 (\frac{1}{\sigma(a)} - 1) = \sigma(a)(1-\sigma(a))$$

(b)

$$\log\left[h_w\left(x^{(i)}\right)\right] = \log\frac{1}{1+e^{-wx^{(i)}}} = -\log\left(1+e^{-wx^{(i)}}\right)$$

$$\log\left[1-h_w\left(x^{(i)}\right)\right] = \log(1-\frac{1}{1+e^{-wx^{(i)}}}) = \log\left(e^{-wx^{(i)}}\right) - \log\left(1+e^{-wx^{(i)}}\right)$$

$$= -wx^{(i)} - \log\left(1+e^{-wx^{(i)}}\right)$$

$$\mathrm{J(w)} = L_W\left(\left\{x^{(i)}, y^{(i)}\right\}_{i=1}^{n}\right) = \sum_{i=1}^{n}\{-y^{(i)}\log\,(h_w\left(x^{(i)}\right) - \left(1 - y^{(i)}\right)\log\left[1 - h_w\left(x^{(i)}\right)\right]\}$$

$$= -\sum_{i=1}^{n}\left\{-y^{(i)}\log\left(1 + e^{-wx^{(i)}}\right)\right.$$

$$\left. + \left(1 - y^{(i)}\right)\left[-wx^{(i)} - \log\left(1 + e^{-wx^{(i)}}\right)\right]\right\}$$

$$= -\sum_{i=1}^{n}\left\{y^{(i)}wx^{(i)} - wx^{(i)} - \log\left(1 + e^{-wx^{(i)}}\right)\right\}$$

$$= -\sum_{i=1}^{n}\left\{y^{(i)}wx^{(i)} - \log\left(1 + e^{wx^{(i)}}\right)\right\}$$

$$\frac{\partial(y^{(i)}wx^{(i)})}{\partial w_j} = -y^{(i)}x_j^{(i)}$$

$$\frac{\partial\left(\log\left(1 + e^{wx^{(i)}}\right)\right)}{\partial w_j} = \frac{x_j^{(i)}e^{wx^{(i)}}}{1 + e^{wx^{(i)}}} = x_j^{(i)}h_w\left(x^{(i)}\right)$$

$$\frac{\partial \mathrm{J(w)}}{\partial w_j} = -\sum_{i=1}^{n} -y^{(i)}x_j^{(i)} - x_j^{(i)}h_w\left(x^{(i)}\right) = \sum_{i=1}^{n}(h_w\left(x^{(i)}\right) - y^{(i)})x_j^{(i)}$$

(c)

$$\frac{\partial^2 \mathrm{J(w)}}{\partial {w_j}^2} = \frac{\partial\sum_{i=1}^{n}(h_w\left(x^{(i)}\right) - y^{(i)})x_j^{(i)}}{\partial w_j} = \sum_{i=1}^{n}\frac{\partial h_w\left(x^{(i)}\right)}{\partial w_j}x_j^{(i)}$$

$$\frac{\partial h_w\left(x^{(i)}\right)}{\partial w_j} = h_w\left(x^{(i)}\right)\left(1 - h_w\left(x^{(i)}\right)\right) > 0$$

Since the second derivative of the cross entropy loss of logistic regression is greater than 0, the cross entropy loss of logistic regression is convex.

3 Boosting

(a)

When $G(x_i) \neq y_i$, $y_i f(x_i) < 0$. And then $\exp(-y_i f(x_i)) \geq 1$. So we have:

$$training\ error\ rate = \frac{1}{N}\sum_{i=1}^{N} I(G(x_i) \neq y_i) \leq \frac{1}{N}\sum_{i} \exp(-y_i f(x_i))$$

Now we need to prove that: $\frac{1}{N}\sum_i \exp(-y_i f(x_i)) = \prod_m Z_m$

Since we have

$$w_{m+1,\mathrm{i}} = \frac{w_{m\mathrm{i}}}{Z_m}\exp(-\alpha_m y_i G_m(x_i))$$

and

$$Z_m = \sum_{i=1}^{N} w_{m\mathrm{i}}\exp(-\alpha_m y_i G_m(x_i))$$

we can get

$$w_{m\mathrm{i}}\exp\left(-\alpha_m y_i G_m(x_i)\right) = Z_m w_{m+1,\mathrm{i}}$$

$$\frac{1}{N}\sum_i \exp\left(-y_i f(x_i)\right) = \frac{1}{N}\sum_i \exp\left(-\sum_{m=1}^{M}\alpha_m y_i G_m(x_i)\right)$$

$$= \sum_i w_{1\mathrm{i}} \prod_{m=1}^{M}\exp\left(-\alpha_m y_i G_m(x_i)\right) = Z_1 \sum_i w_{2\mathrm{i}} \prod_{m=2}^{M}\exp(-\alpha_m y_i G_m(x_i))$$

$$= Z_1 Z_2 \sum_i w_{3\mathrm{i}} \prod_{m=3}^{M}\exp(-\alpha_m y_i G_m(x_i)) = \cdots$$

$$= Z_1 Z_2 \ldots Z_{M-1} \sum_i w_{M\mathrm{i}}\exp\left(-\alpha_M y_i G_M(x_i)\right) = \prod_{m=1}^{M} Z_m$$

$$Z_m = \sum_{i=1}^{N} w_{m\mathrm{i}}\exp\left(-\alpha_m y_i G_m(x_i)\right) = \sum_{y_i = G_m(x_i)} w_{m\mathrm{i}} e^{-\alpha_m} + \sum_{y_i \neq G_m(x_i)} w_{m\mathrm{i}} e^{\alpha_m}$$

$$= (1-e_m)e^{-\alpha_m} + e_m e^{\alpha_m} = 2\sqrt{e_m(1-e_m)}$$

Since $e_m = \frac{1}{2} - \gamma_m$

$$Z_m = 2\sqrt{e_m(1-e_m)} = \sqrt{1-4{\gamma_m}^2}$$

Since $1 + \mathrm{x} \leq e^{x}$ for all real x

$$Z_m = \sqrt{1 - 4{\gamma_m}^2} \leq e^{-2{\gamma_m}^2}$$

$$\prod_{m=1}^{M} Z_m \leq \prod_{m=1}^{M} e^{-2{\gamma_m}^2} = \exp\left(-2\sum_{m=1}^{M} {\gamma_m}^2\right)$$

By weak learning assumption, $\gamma_m \geq \gamma$ for all m

$$\text{training error rate} \leq \exp\left(-2\gamma^2 M\right)$$

When M approaches to infinity, the training error rate approaches to 0.

(b)

Given training data $\mathrm{D} = \{(x_i, y_i)\}_{i=1}^{n}$ and maximum number of iterations T

Initialize weights: $d_{1,i} = \frac{w_i}{\sum_{i=1}^{n} w_i}$

for t=1, ⋯, T do

train a weak classifier: $h_t = argmin_h \sum_{i=1}^{n} d_{t,i} \times I(h(x_i) \neq y_i)$

compute its weighted error: $\epsilon_t = \sum_{i=1}^{n} d_{t,i} \times I(h(x_i) \neq y_i)$

compute coefficient: $\alpha_t = \frac{1}{2} ln \frac{1-\epsilon_t}{\epsilon_t}$

update weights: $d_{t+1,i} = \begin{cases} \frac{d_{t,i} \times e^{-\alpha_t}}{Z_t},\ if\ x_i\ is\ correctly\ classified, y_i = h_t(x_i) \\ \frac{d_{t,i} \times e^{\alpha_t}}{Z_t},\ if\ x_i\ is\ misclassified, y_i \neq y_i \end{cases}$

where Z_t is normalization constant for the discrete distribution, $Z_t =$

$\sum_{i=1}^{n} d_{t+1,i} = 1$

end for