DS-GA 1003: Machine Learning and Computational Statistics
Homework 5 - Extra: Boosting

1 AdaBoost (Optional)

Introduction

Given training set $D = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, where y_i's are either $+1$ or -1, suppose we have a weak learner G_t at time t and we will perform AdaBoost T times. Initialize observation weights uniformly by setting $W^1 = (w_1^1, \ldots, w_n^1)$ and $w_i = 1/n$ for $i = 1, 2, \ldots, n$. For $t = 1, 2, \ldots, n$:

1. Fit the weak learner at time t to weighted samples: G_t that depends on (D, W^t)
2. Compute the weighted misclassifications: $\mathrm{err}_t = \sum_D w_i^t \mathbb{1}_{\{G_t(x_i) \neq y_i\}} / \sum_i w_i^t$
3. Compute the contribution coefficient for the weak learner: $\alpha_t = \frac{1}{2} \log(\frac{1}{\mathrm{err}_t} - 1)$
4. Update the weights: $w_i^{t+1} = w_i^t \exp(-\alpha_t y_i G_t(x_i))$

After T steps, the cumulative contributions of weak learners is $G(x) = \mathrm{sign}(\sum_{t=1}^T \alpha_t G_t(x))$ as the final output. We will prove that with a reasonable weak learner the error of the output decreases exponentially fast with the number of iterations.

Exponential bound on the training loss

More precisely, we will show that the training error $L(G, D) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{G(x_i) \neq y_i\}} \leq \exp(-\gamma^2 T)$ where the error of the weak learner is less than $1/2 - \gamma$ for some $\gamma > 0$. To start, let's denote two cumulative variables: the output at time t as $f_t = \sum_{s \leq t} \alpha_s G_s$ and $Z_t = \frac{1}{n} \sum_{i=1}^n \exp(-y_i f_t(x_i))$.

1. For any function g into $\{-1, +1\}$, show that $\mathbb{1}_{\{g(x) \neq y\}} < \exp\left(-y g(x)\right)$.

   SN: When $g(x) = y$ we have $0 < e^{-1}$, when $g(x) \neq y$ we have $1 < e$.
2. Use this to show $L(G, D) < Z_T$

   SN: $L(G, D) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{G(x_i) \neq y_i\}} < \frac{1}{n} \sum_{i=1}^n \exp(-y_i G(x_i)) = \frac{1}{n} \sum_{i=1}^n \exp(-y_i f_T(x_i)) = Z_T$
3. Show that $w_i^{t+1} = \exp(-y_i f_t(x_i))$

   SN: Using an inductive argument $w_i^{t+1} = \exp(-y_i \sum \alpha_t G_t(x_i)) = \exp(-y_i f_t(x_i))$.

4. Use part 3 to show $\frac{Z_{t+1}}{Z_t} = 2\sqrt{\mathrm{err}_{t+1}(1-\mathrm{err}_{t+1})}$ (Hint: use the definition of weight updates and separate the sum on where G_t is equal to 1 and -1.)

   SN:

$$\begin{aligned}
\frac{Z_{t+1}}{Z_t} &= \frac{\sum \exp(-y_i f_{t+1}(x_i))}{\sum \exp(-y_i f_t(x_i))} &(1)\\
&= \frac{\sum \exp(-y_i f_t(x_i)) \exp(-y_i \alpha_{t+1} G_{t+1}(x_i))}{\sum \exp(-y_i f_t(x_i))} &(2)\\
&= \frac{\sum w_i^{t+1} \exp(-y_i \alpha_{t+1} G_{t+1}(x_i))}{\sum w_i^{t+1}} &(3)\\
&= \exp(-\alpha_{t+1})(1-\mathrm{err}_{t+1}) + \exp(\alpha_{t+1})\mathrm{err}_{t+1} &(4)\\
&= \frac{1}{\sqrt{1/\mathrm{err}_{t+1}-1}}(1-\mathrm{err}_{t+1}) + \sqrt{1/\mathrm{err}_{t+1}-1}(\mathrm{err}_{t+1}) &(5)\\
&= 2\sqrt{\mathrm{err}_{t+1}(1-\mathrm{err}_{t+1})} &(6)
\end{aligned}$$

5. Show that the function $g(a) = a(1-a)$ is monotonically increasing on $[0, 1/2]$. Show that $1-a \leq \exp(-a)$. And use the assumption on the weak learner to show that $\frac{Z_{t+1}}{Z_t} \leq \exp(-2\gamma^2)$

   SN: $g'(a) = 1 - 2a \geq 0$ on $[0, 1/2]$. Expand e^{-x} in Taylor series. Then,

$$\begin{aligned}
\frac{Z_{t+1}}{Z_t} &= 2\sqrt{\mathrm{err}_{t+1}(1-\mathrm{err}_{t+1})} &(7)\\
&\leq 2\sqrt{(1/2-\gamma)(1/2+\gamma)} &(8)\\
&= \sqrt{1-4\gamma^2} &(9)\\
&= \exp(-2\gamma^2) &(10)
\end{aligned}$$

6. Conclude the proof!

   SN: Since $Z_0 = 1$, write $Z_T = \frac{Z_T}{Z_{T-1}} \frac{Z_{T-1}}{Z_{T-2}} \ldots \frac{Z_1}{Z_0}$ which is the missing link.

2 Additive model

Introduction

The main function in AdaBoost, $G(x) = \mathrm{sign}(\sum_{t=1}^{T} \alpha_t G_t(x))$, is an additive expansion in a set of 'basis' functions, $f(x) = \sum_{t=1}^{T} \alpha_t G_t(x)$. The function f is similar to the way of representing a vector as a linear combination of the basis vectors in linear algebra: Given a set of basis elements, find the correct coefficients. Here we have $G_t(x)$'s as basis functions and α's as coefficients.

In the additive model, the algorithm starts by initializing $f_0(x) = 0$, and then for $t = 1, \ldots, T$ iterate over the following for some loss function L:

1. Compute $(\alpha_t, G_t) = \mathrm{argmin}_{\alpha, G} \sum_{i=1}^{n} L(y_i, f_{t-1}(x_i) + \alpha G(x_i))$
2. Find the expansion at time t: $f_t(x) = f_{t-1}(x) + \alpha_t G_t(x)$

In the next problem, show that using exponential loss will lead to AdaBoost.

Exponential loss and AdaBoost

Consider the loss function $L(y, f(x)) = \exp(-yf(x))$.

1. Write the first step of the additive model using the exponential loss function. Show that it can be written as:

$$(\alpha_t, G_t) = \text{argmin}_{\alpha,G} \sum_{i=1}^{n} w_i^t \exp(-\alpha y_i G(x_i)))$$

   SN: Rewrite line 1 of the intro

2. Show that for fixed positive alpha:

$$G_t = \text{argmin}_G \sum_{i=1}^{n} w_i^t \mathbb{1}_{\{G(x_i) \neq y_i\}}$$

   (Hint: split the sum in part 1 for $y_i = G(x_i)$ and otherwise.)

   SN: Use the part that depends on G in the following equality:

$$\sum_{i=1}^{n} w_i^t \exp(-\alpha y_i G(x_i)) = e^{-\alpha} \sum_{y_i = G(x_i)} w_i^t + e^{\alpha} \sum_{y_i \neq G(x_i)} w_i^t \tag{11}$$

$$= (e^{\alpha} - e^{-\alpha}) \sum_{i}^{n} w_i^t \mathbb{1}_{\{G(x_i) \neq y_i\}} + e^{-\alpha} \sum_{i}^{n} w_i^t \tag{12}$$

3. Plug this G_t back into the first equation and solve for α to obtain $\alpha_t = \frac{1}{2} \log \frac{1}{\text{err}_t} - 1$

   SN: Take derivative of equation 12 and set it to zero: $(e^{\alpha} + e^{-\alpha})\text{err}_t = e^{\alpha}$. And solve for α where error is the same as in the previous problem.

4. Show that the weight iterations are given by:

$$w_i^{t+1} = w_i^t \exp(-\alpha_t y_i G_t(x_i))$$

   And conclude the equivalence.

   SN: Use line 2 of the introduction