S&DS 365 Homework 5 Solutions

Yale University, Department of Statistics

April 20, 2021

Problem 1: Deriving adaboost from forward stage wise classification

a)

Let $w_i = e^{-y_i G(x_i)}$. Then we have

$$
\begin{aligned}
&\text{argmin}_{\alpha>0,\mathcal{F}} \sum_i e^{-y_i(G(x_i)+\alpha F(x_i)} \\
=&\text{argmin}_{\alpha>0,\mathcal{F}} \sum_i e^{-y_i(G(x_i)} e^{-y_i \alpha F(x_i)} \\
=&\text{argmin}_{\alpha>0,\mathcal{F}} \sum_i w_i e^{-y_i \alpha F(x_i)}
\end{aligned}
$$

now, $F(x_i)$ can take two values, either -1 or $+1$, and each y_i is also either -1 or $+1$. Therefore if $F(x_i) = y_i$ we get $e^{-\alpha}$ and if $F(x_i) \neq y_i$ we get e^{α}. So really our terms can take two different values depending on if the predicted $F(x_i)$ agrees with y_i or not.

If subtract a constant from our argmin, this does not change the minimizing term so we have write

$$
\text{argmin}_{\alpha>0,\mathcal{F}} \sum_i w_i (e^{\alpha F(x_i)} - e^{-\alpha})
$$

now these terms either take the value $e^{\alpha} - e^{\alpha}$ or 0. We can then scale the terms without changing the argmin and the terms will now be either $0, 1$ depending on if $F(x_{i)} = y_i$.

$$
\begin{aligned}
&\text{argmin}_{\alpha>0,\mathcal{F}} \sum_i w_i \frac{(e^{\alpha F(x_i)} - e^{-\alpha})}{e^{\alpha} - e^{-\alpha}} \\
=&\text{argmin}_F \sum_i w_i 1_{F(x_i)=y_i}
\end{aligned}
$$

which is independent of α.

b)

Now we fix our choice of G_m we optimize the α

$$
\begin{aligned}
&\text{argmin}_{\alpha>0} \sum_i e^{-y_i(G(x_i)+\alpha G_m(x_i)} \\
=&\text{argmin}_{\alpha} \sum_i w_i e^{-\alpha y_i G_m(x_i)}
\end{aligned}
$$

again, $G_m(x_i)$ takes values -1,+1 as does y_i so if the agree we get $e^{-\alpha}$ and if they disagree we get $e^{-\alpha}$. If we define

$$\epsilon_m = \frac{\sum_i w_i 1_{y_i \neq G_m(x_i)}}{\sum_i w_i}$$

which represents the relative weights of the misclassified terms. We can divide our argmin by a constant without affecting the result and then break the sum into two parts and we have

$$\begin{aligned}
&\text{argmin}_\alpha \frac{\sum_i w_i e^{-\alpha y_i G_m(x_i)}}{\sum_i w_i} \\
&= \text{argmin}_\alpha \frac{\sum_{i|y_i \neq G_m(x_i)} w_i e^{-\alpha y_i G_m(x_i)} + \sum_{i|y_i = G_m(x_i)} w_i e^{-\alpha y_i G_m(x_i)}}{\sum_i w_i} \\
&= \text{argmin}_\alpha e^{\alpha} \epsilon_m + \epsilon^{-\alpha}(1 - \epsilon_m)
\end{aligned}$$

we then take the derivative and optimize α

$$\begin{aligned}
e^{\alpha}\epsilon_m - e^{-\alpha}(1 - \epsilon_m) &= 0 \\
e^{2\alpha} &= \frac{1 - \epsilon_m}{\epsilon_m} \\
\alpha &= \frac{1}{2} \ln \left(\frac{1 - \epsilon_m}{\epsilon_m} \right)
\end{aligned}$$

BONUS

We also must show $\alpha > 0$, (students don't need to show this but worth noting).

First, assume $\epsilon_m \leq 1 - \epsilon_m$ (we got more right than wrong in classifying). Suppose not, then let $\tilde{G} = -G$, that is we flip all our classifications since we got more wrong than right anyways.

Then we flipped everything so let

$$\tilde{\epsilon}_m = \frac{\sum_i w_i 1_{y_i \neq \tilde{G}_m(x_i)}}{\sum_i w_i}$$

we have $\tilde{\epsilon}_m = 1 - \epsilon$ and $\tilde{\epsilon}_m \leq \epsilon_m$. However, G is supposed to be the minimizer so it should have the smallest ϵ_m possible,

$$G \in \text{argmin}_m \sum_i w_i 1_{y_i \neq G_m(x_i)} \implies \epsilon_m \leq \frac{\sum_i w_i 1_{y_i \neq F_m(x_i)}}{\sum_i w_i}$$

for any choice of F. This poses a contradiction, $\tilde{G}$ cannot be better than G thus it must be that $\epsilon_m \leq 1 - \epsilon_m$ and therefore

$$\alpha_m = \frac{1}{2} \ln \frac{1 - \epsilon_m}{\epsilon_m} \geq 0$$

c)

Parts a) and b) are the same as in the textbook. However parts c and d are different. We have:

$$\tilde{w}_i \leftarrow \tilde{w}_i \text{exp}\left(-\frac{\alpha_m}{2} y_i G_m(x_i) \right)$$

so $y_i G_m(x_i)$ takes values +1,-1 depending on if the numbers are the same. This is the same as

$$-y_i G_m(x_i) = 2 * 1\{y_i \neq G_m(x_i)\} - 1$$

$$\begin{aligned} &\tilde{w}_i \exp\left(-\frac{\alpha_m}{2}(2 * 1\{y_i \neq G_m(x_i)\} - 1)\right) \\ =&\tilde{w}_i \exp\left(\alpha_m 1\{y_i \neq G_m(x_i)\})\exp(\frac{\alpha_m}{2})\right) \end{aligned}$$

but each weight is then scaled by $e^{\frac{\alpha_m}{2}}$ so when we renormalise this cancels out and weights are the same.

Problem 2: Regularization

a)

Taking derivative,

$$\begin{aligned} \frac{X^T}{n}(X\theta - y) + 2\lambda\theta = 0 \\ \hat{\theta} = \left(\frac{(X^T X)^{-1}}{n} + 2\lambda I\right)^{-1} \frac{X^T}{n} y \end{aligned}$$

note if $\lambda = 0$ this would be usual LS estimator.

b)

If $X = I$ we would have

$$\begin{aligned} \hat{\theta} &= \left((\frac{1}{n} + 2\lambda)I\right)^{-1} \frac{y}{n} \\ &= \frac{\frac{y}{n}}{\frac{1}{n} + 2\lambda} = \frac{y}{1 + 2n\lambda} \end{aligned}$$

c)

Define

$$\text{sign}(x) = \begin{cases} 1 & x > 0 \\ 0 & x = 0 \\ -1 & x = 0 \end{cases}$$

then gradient descent is

$$\theta_{k+1} = \theta_k - \eta_k(\frac{X^T}{n}(X\theta_k - y) + \lambda \text{sign}(\theta_k))$$

d)

Now define

$$\text{sign}(x) = \begin{cases} 1 & x > 0 \\ [-1, 1] & x = 0 \\ -1 & x = 0 \end{cases}$$

that is we don't know exactly what we want to set the value at 0 to, but leave it as a place holder for now. We then have setting $X = I$,

$$\begin{aligned} \frac{1}{n}(\theta - y) + \lambda \text{sign}(\theta) = 0 \\ \hat{\theta} = y - n\lambda \text{sign}(\hat{\theta}) \end{aligned}$$

looking at coordinate i we have

$$\hat{\theta}_i = y - n\lambda \text{sign}(\hat{\theta}_i) \tag{1}$$

we then propose the following estimator

$$\hat{\theta}_i = \begin{cases} y_i - n\lambda & y_i \geq n\lambda \\ 0 & |y_i| < n\lambda \\ y_i + n\lambda & y_i \leq -n\lambda \end{cases}$$

We now show how this satisfies (1). If $y_i > n\lambda$ then $\hat{\theta}_i > 0$ and this $\text{sign}(\hat{\theta}_i) = 1$ and thus

$$y_i - n\lambda \text{sign}(\hat{\theta}_i) = y_i - n\lambda$$

if $y_i < -n\lambda$ we have $\hat{\theta}_i < 0$ and $\text{sign}(\hat{\theta}_i) = -1$ and we have

$$y_i - n\lambda \text{sign}(\hat{\theta}_i) = y_i + n\lambda$$

If $|y_i| \leq n\lambda$ then we try to set $\text{sign}(\hat{\theta}_i)$ to be some value between $[-1, 1]$ to make $y_i - n\lambda \text{sign}(\hat{\theta}_i) = 0$ and thus agree with $\hat{\theta}_i = 0$ in this case. We have

$$\begin{aligned} y_i - n\lambda \text{sign}(\hat{\theta}_i) = 0 \\ \text{sign}(\hat{\theta}_i) = \frac{y_i}{n\lambda} \in [1, 1] \end{aligned}$$

so our estimator satisfies the conditions.