ISYE-6740

Homework 5

1. a)

Given m training samples we know that our formulation for Adaboost is the following:

AdaBoost

- constructing D_t:
  - $D_1(i) = 1/m$
  - given D_t and h_t: $\epsilon_t = \frac{1}{m}\sum_{i=1}^{m} D_t(i)\mathbb{I}\{y_i \neq h_t(x_i)\}$

$$D_{t+1}(i) = \frac{D_t(i)}{Z_t} \times \begin{cases} e^{-\alpha_t} & \text{if } y_i = h_t(x_i) \\ e^{\alpha_t} & \text{if } y_i \neq h_t(x_i) \end{cases}$$

$$= \frac{D_t(i)}{Z_t} \exp(-\alpha_t\, y_i\, h_t(x_i))$$

where Z_t = normalization constant

$$\alpha_t = \tfrac{1}{2} \ln\left(\frac{1-\epsilon_t}{\epsilon_t}\right) > 0$$

- final classifier:
  - $H_{\text{final}}(x) = \text{sign}\left(\sum_t \alpha_t h_t(x)\right)$

Figure 1: Adaboost Classifier

Given our observations:

X1 = (-1, 0, +), X2 = (-0.5, 0.5, +), X3 = (0, 1, -), X4 = (0.5, 1, -)

X5 = (1, 0, +), X6 = (1,-1, +), X7 = (0, -1, -), X8 = (0, 0, -)

We know based on the formula that our first iteration will equal $\frac{1}{8}$ for all observations, then epsilon, alpha and Z we will derive with the provided formulas:

Iter (t)	ϵ_t	α_t	Z_t	$D_t(1)$	$D_t(2)$	$D_t(3)$	$D_t(4)$	$D_t(5)$	$D_t(6)$	$D_t(7)$	$D_t(8)$
1	$\frac{1}{4}$	0.55	$\frac{\sqrt{3}}{2}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$
2	$\frac{1}{6}$	0.81	$\frac{\sqrt{5}}{3}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{12}$	$\frac{1}{12}$

3	$\frac{1}{10}$	1.15	0.6	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{20}$	$\frac{1}{20}$	$\frac{3}{20}$	$\frac{3}{20}$	$\frac{1}{20}$	$\frac{1}{20}$

The misclassified points as categorized as by h_t are as follows:

$h_1(x) = 1$ If $x_1 < -\frac{1}{4}$, -1

$h_2(x) = 1$ If $x_1 > \frac{3}{4}$, -1

$h_3(x) = 1$ If $x_1 < \frac{3}{4}$, -1

Therefore, with our misclassified points and our values for α, we know that our final classifier can be denoted with:

$$h(x) = \text{sign}(0.55\text{sign}\left(-x_1 - \frac{1}{4} + \right) + 0.81\text{sign}\left(x_1 - \frac{3}{4}\right) + 1.15\text{sign}\left(-x_2 + \frac{3}{4}\right)$$

We can now draw our misclassified points as the decision hump in the figure as follows:

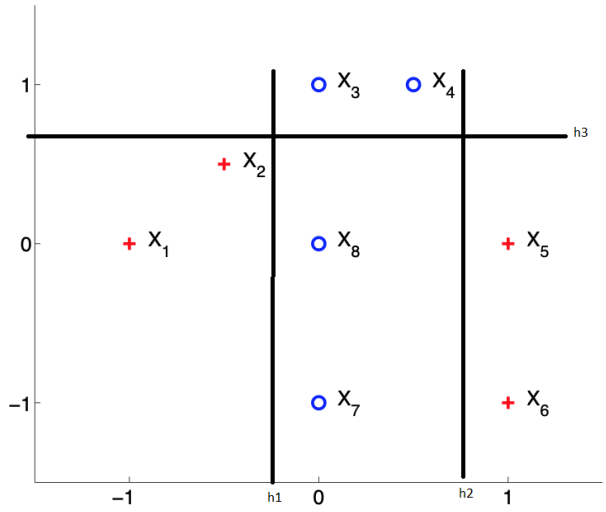


Figure 2: Decision Humps AdaBoost

b) The training error for Adaboost converges to 0 because the training error has to go down exponentially to ensure that the weight error of the classifier ϵ_t is better than $\epsilon_t < 0.5$

2.

a)

m = 4601 observations

n_{spam}= 1813

$n_{non-spam}$= 2788

$n_{features}$=57

b)

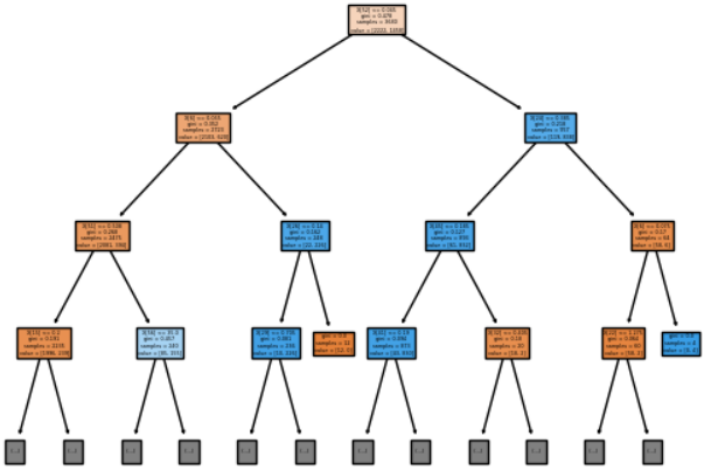


Figure 3: Decision Tree for CART

c)

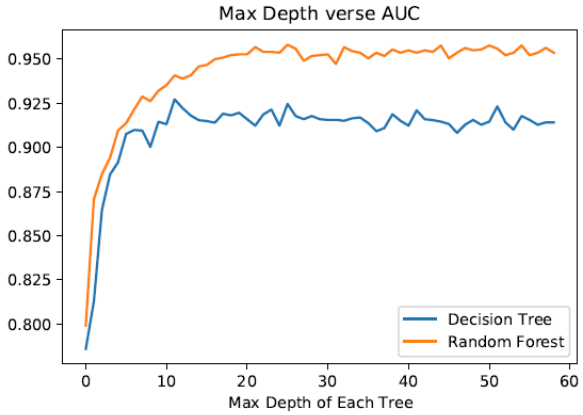


Figure 4: AUC Curve for RF over Decision Tree

3.

a)
Our model's linear regression line, when rounding is approximately:
y = 7.4 + 0.02x
The MSE in our case is 10.5

b)
Not completed

c)
Provided is the cross validation curve showing our score of alpha over the MSE (Mean Squared Error):

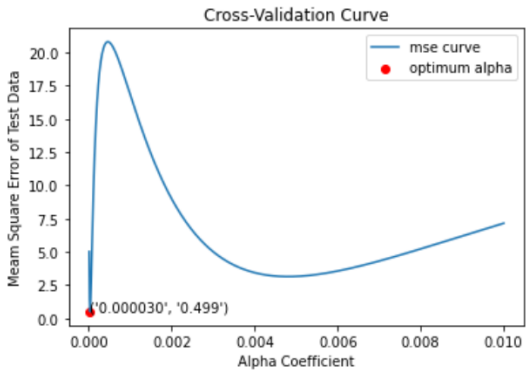


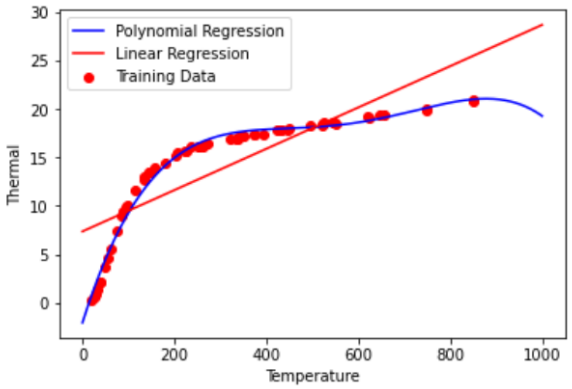
Figure 5: Cross Validation Curve

d)

When looking at the coefficient at the 400th degree, we are able to calculate our Linear and Polynomial regression values as follows:

Linear = 15.9
PolyNomial = 16.9

We can see that the Polynomial regression has an extremely accurate results, however we can very likely conclude that we will experience overfitting in this case as the model is very tightly fit against our data, and likely wouldn't react well to any of our test observations.



4.

Based on our lecture slides, we know that the bias and variance terms can be computed as follows:

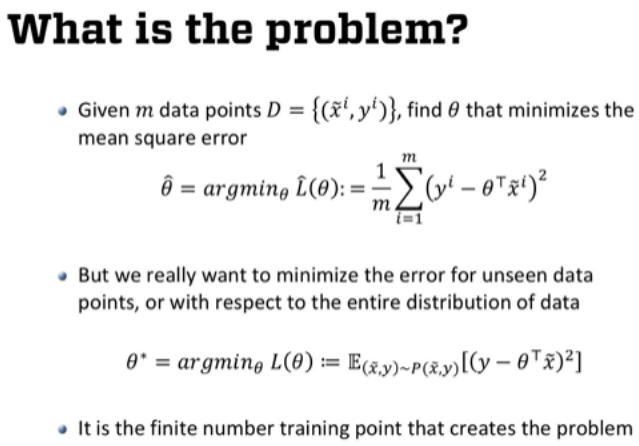


Figure 6: Bias & Variance Tradeoff

Our goal will be to find a derivative for A with respect to f in order to minimize our loss function

- The expected squared loss is

$$L(\hat{f}) := \mathbb{E}_D \mathbb{E}_{(x,y)} \left[\left(y - \hat{f}(x) \right)^2 \right]$$
$$= \mathbb{E}_D \left[\underbrace{\int\int \left(y - \hat{f}(x) \right)^2 p(x,y) dx dy}_{A} \right]$$

- Our goal is to choose $\hat{f}(x)$ that minimize $L(\hat{f})$. Calculus of variations

$$\frac{\partial A}{\partial f(x)} = 2 \int (y - f(x)) p(x,y) dy = 0$$
$$\Leftrightarrow \int f(x) p(x,y) dy = \int y p(x,y) dy$$
$$\Leftrightarrow h(x) := \int \frac{y p(x,y)}{p(x)} dy = \int y p(y|x) dt = \mathbb{E}_{y|x}[y] = \mathbb{E}[y|x]$$

Figure 7: Gradient for F loss Function

Therefore, we can start by taking the derivative of the objective function, which in this case is our coefficient β as follows:

$$\frac{\partial \mathrm{L}}{\partial \beta} = \; -2\mathrm{X}^{\mathrm{T}}(\mathrm{y} - \mathrm{X}\beta) + \; 2\lambda\beta = 0 \tag{1}$$

$$\hat{\beta} = \; -2\mathrm{X}^{\mathrm{T}}(\mathrm{y} - \mathrm{X}\beta) + \; 2\lambda\beta = 0 \tag{2}$$

When we simplify this further we know that beta can be reduced to the following:

$$\hat{\beta} = \; \left(\lambda \mathrm{I} \; + \; \mathrm{X}^{\mathrm{T}}\mathrm{X}\right)^{-1} \mathrm{X}^{\mathrm{T}}\mathrm{y} \tag{3}$$

$$\hat{\beta} = \; \left(\lambda \mathrm{I} \; + \; \mathrm{X}^{\mathrm{T}}\mathrm{X}\right)^{-1} \mathrm{X}^{\mathrm{T}}\beta^{*} + \left(\lambda \mathrm{I} \; + \; \mathrm{X}^{\mathrm{T}}\mathrm{X}\right)^{-1} \mathrm{X}^{\mathrm{T}}\epsilon \tag{4}$$

This means our error equation can now be simplified to the following form:

$$\mathrm{E}\left[\hat{\beta}\right] = \; \left(\lambda \mathrm{I} \; + \; \mathrm{X}^{\mathrm{T}}\mathrm{X}\right)^{-1} \mathrm{X}^{\mathrm{T}}\beta^{*} \tag{5}$$

Now finally our variance will be calculated as follows:

$$\mathrm{Var}\left(\hat{\beta}\right) = \; \mathrm{Var}\left(\lambda \mathrm{I} \; + \; \mathrm{X}^{\mathrm{T}}\mathrm{X}\right)^{-1} \mathrm{X}^{\mathrm{T}}\beta^{*} + \left(\lambda \mathrm{I} \; + \; \mathrm{X}^{\mathrm{T}}\mathrm{X}\right)^{-1} \mathrm{X}^{\mathrm{T}}\epsilon \tag{6}$$

$$= \; \left(\lambda \mathrm{I} \; + \; \mathrm{X}^{\mathrm{T}}\mathrm{X}\right)^{-1} \mathrm{X}^{\mathrm{T}}\mathrm{X}\left(\left(\lambda \mathrm{I} \; + \; \mathrm{X}^{\mathrm{T}}\mathrm{X}\right)^{-1}\right)^{-1} * \sigma^{2} \tag{7}$$

b-d) Not completed due to time constraint.