ML HW 5

ID - 208935916

(1)

The root:

$\arg\max_{i\in\{1,2,3\}} G(S,i)$ where $G(S,i) = I(Y;X_i) = H[Y] - H[Y|X_i]$.

Lets see the following calculations:

- $H[Y] = \mathbb{P}_S[Y=1]\log\frac{1}{\mathbb{P}_S[Y=1]} + \mathbb{P}_S[Y=0]\log\frac{1}{\mathbb{P}_S[Y=0]} = \log(2)$

$$H[Y|X_i] = \mathbb{P}_S[X_i=1]H[Y|X_i=1] + \mathbb{P}_S[X_i=0]H[Y|X_i=0] =$$
$$\mathbb{P}_S[X_i=1]\left(\mathbb{P}_S[Y=1|X_i=1]\log\frac{1}{\mathbb{P}_S[Y=1|X_i=1]} + \mathbb{P}_S[Y=0|X_i=1]\log\frac{1}{\mathbb{P}_S[Y=0|X_i=1]}\right)$$
$$+\mathbb{P}_S[X_i=0]\left(\mathbb{P}_S[Y=1|X_i=0]\log\frac{1}{\mathbb{P}_S[Y=1|X_i=0]} + \mathbb{P}_S[Y=0|X_i=0]\log\frac{1}{\mathbb{P}_S[Y=0|X_i=0]}\right)$$

- $H[Y|X_1] = 0.75\left(\frac{2}{3}\log\left(\frac{3}{2}\right) + \frac{1}{3}\log(3)\right) + 0.25\cdot 0$
- $H[Y|X_2] = H[Y|X_3] = 0.5(0.5\log(2) + 0.5\log(2)) + 0.5(0.5\log(2) + 0.5\log(2)) = \log(2)$

So -> $\arg\max_{i\in\{1,2,3\}} G(S,i) = 1$,and X_1 is the root.

Now we divide the data to sets:

$S_0 = \{points\ with\ X_1 = 0\} = \{((0,0,1),0)\}$

$S_1 = \{points\ with\ X_1 = 1\} = \{((1,1,1),1),((1,0,0),1),((1,1,0),0)\}$

S_0 has a unique label so it is a leaf labeled 0.

We do the same calculation as before on S_1 on the set $\{X_2, X_3\}$ and we chose the next split node:

	X_2	X_3	Y
0	1/3	2/3	1/3
1	2/3	1/3	2/3

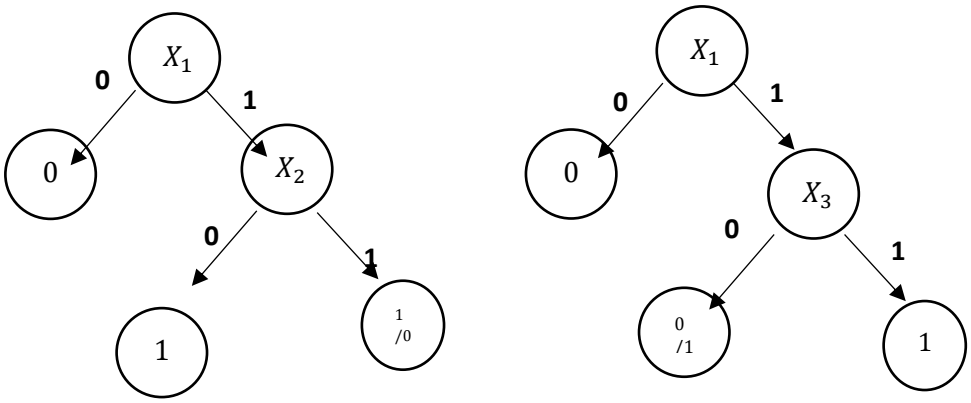
Conditional probabilities: $\mathbb{P}_{S_1}[Y=1|X_i=j]$

	$X_2=1$	$X_2=0$	$X_3=1$	$X_3=0$
$Y=1$	0.5	1	1	0.5

So,

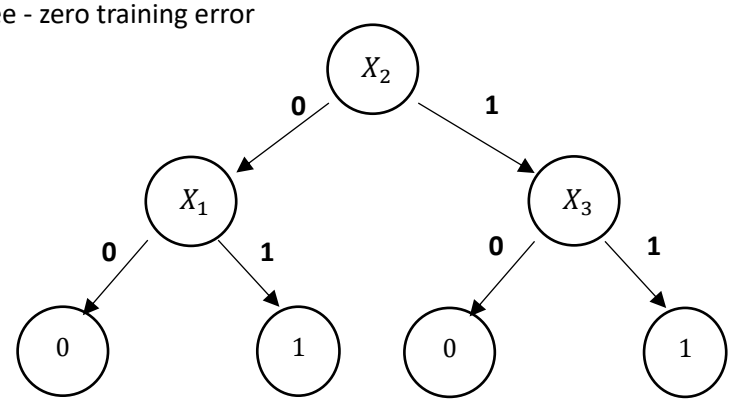
- $H[Y|X_2] = 2/3(\log 2) + 1/3(0) = 2/3\cdot\log(2)$
- $H[Y|X_3] = 1/3(0) + 2/3(\log 2) = 2/3\cdot\log(2)$

We get $I(Y;X_2) = I(Y;X_3)$ and the split can be done on both of them. Then we need to stop building the tree and we determine the value of leaves by the majority of labeling. In any case, the train error is at least $1/4$. Let's see our 2 optional trees:



Where we write 0/1 since the majority of both are equal. In any case, it can be shown that we have only 1 error for the data set for this leaf, which gives us error of at least $1/4$.

a. Decision tree - zero training error



(2)

a.

$$D_{KL}(p,q) = \sum_{x,y\in X} p(x,y)\cdot\log\left(\frac{p(x,y)}{q(x,y)}\right) = \sum_{x,y\in X} p_1(x)p_2(y)\cdot\log\left(\frac{p_1(x)p_2(y)}{q_1(x)q_2(y)}\right)$$
$$= \sum_{x,y\in X} p_1(x)p_2(y)\cdot\left(\log\left(\frac{p_1(x)}{q_1(x)}\right)+\log\left(\frac{p_2(y)}{q_2(y)}\right)\right)$$
$$= \sum_{x,y\in X} p_1(x)p_2(y)\log\left(\frac{p_1(x)}{q_1(x)}\right)+\sum_{x,y\in X} p_1(x)p_2(y)\log\left(\frac{p_2(y)}{q_2(y)}\right)$$
$$= \sum_{y\in X} p_2(y)\sum_{x\in X} p_1(x)\log\left(\frac{p_2(x)}{q_2(x)}\right)+\sum_{x\in X} p_1(x)\sum_{y\in X} p_2(y)\log\left(\frac{p_2(y)}{q_2(y)}\right)$$
$$= D_{KL}(p_1,q_1)+D_{KL}(p_2,q_2)$$

b.

$$I(Y;X) = H[Y]-H[Y|X]$$
$$= -\sum_{y\in\mathcal{X}}\mathbb{P}_Y[y]\log\mathbb{P}_Y[y]+\sum_{x\in\mathcal{X}}\mathbb{P}_X[x]\sum_{y\in\mathcal{X}}\mathbb{P}_{X,Y}[y|x]\log\mathbb{P}_{X,Y}[y|x]$$
$$= -\sum_{y\in\mathcal{X}}\mathbb{P}_Y[y]\log\mathbb{P}_Y[y]+\sum_{x,y\in\mathcal{X}}\mathbb{P}_X[x]\mathbb{P}_{X,Y}[y|x]\log\mathbb{P}_{X,Y}[y|x]$$
$$= \sum_{x,y\in X}\mathbb{P}_X[x]\mathbb{P}_{X,Y}[y|x]\log\big(\mathbb{P}_{X,Y}[y|x]\big)-\sum_{y\in Y}\log(\mathbb{P}_Y[y])\sum_{x\in X}\mathbb{P}_X[x]\mathbb{P}_{X,Y}[y|x]$$
$$= \sum_{x,y\in X}\mathbb{P}_X[x]\mathbb{P}_{X,Y}[y|x]\log\big(\mathbb{P}_{X,Y}[y|x]\big)-\mathbb{P}_X[x]\mathbb{P}_{X,Y}[y|x]\log(\mathbb{P}_Y[y])$$
$$= \sum_{x,y\in X}\mathbb{P}_X[x]\mathbb{P}_{X,Y}[y|x]\log(\mathbb{P}_X[x])+\mathbb{P}_X[x]\mathbb{P}_{X,Y}[y|x]\log\big(\mathbb{P}_{X,Y}[y|x]\big)$$
$$-\mathbb{P}_X[x]\mathbb{P}_{X,Y}[y|x]\log(\mathbb{P}_X[x])-\mathbb{P}_X[x]\mathbb{P}_{X,Y}[y|x]\log(\mathbb{P}_Y[y])$$
$$= \sum_{x,y\in X}\mathbb{P}_X[x]\mathbb{P}_{X,Y}[y|x]\log\big(\mathbb{P}_X[x]\mathbb{P}_{X,Y}[y|x]\big)-\mathbb{P}_X[x]\mathbb{P}_{X,Y}[y|x]\log(\mathbb{P}_X[x]\mathbb{P}_Y[y])$$
$$= \sum_{x,y\in X}\mathbb{P}_{X\times Y}[(x,y)]\log\mathbb{P}_{X\times Y}[(x,y)]-\mathbb{P}_{X\times Y}[(x,y)]\log\mathbb{P}_{X\otimes Y}[(x,y)]$$
$$= \sum_{x,y\in X}\mathbb{P}_{X\times Y}[(x,y)]\log\frac{\mathbb{P}_{X\times Y}[(x,y)]}{\mathbb{P}_{X\otimes Y}[(x,y)]} = D_{KL}\big(P_{X\times Y},P_{X\otimes Y}\big)$$

(3)

(a) Take expectation of both sides of (1) wrt D:

$$\gamma \le \sum_{i=1}^{n} D(i)y_i\sum_{j=1}^{k} a_j h_j(x_i) = \sum_{j=1}^{k} a_j\sum_{i=1}^{n} D(i)y_i h_j(x_i)$$

Since $a_j \ge 0, \sum_{j=1}^{k} a_j = 1$, there exists j such that $\sum_{i=1}^{n} D(i)y_i h_j(x_i) \ge \gamma$. If $y_i = h_j(x_i)$ then $y_i h_j(x_i) = 1$ and otherwise $y_i h_j(x_i) = -1$. Therefore,

$$\sum_{i=1}^{n} D(i)y_i h_j(x_i) = \sum_{\substack{1\le i\le n\\ y_i=h_j(x_i)}} D(i) - \sum_{\substack{1\le i\le n\\ y_i\ne h_j(x_i)}} D(i) = 1-2\sum_{\substack{1\le i\le n\\ y_i\ne h_j(x_i)}} D(i) \ge \gamma$$

Hence, $\mathbb{P}_{i\sim D}\big[y_i \ne h_j(x_i)\big] = \sum_{\substack{1\le i\le n\\ y_i\ne h_j(x_i)}} D(i) \le \frac{1}{2}-\frac{\gamma}{2}$

(b) Set $k = 4d - 1, a_i = \frac{1}{4d-1}\ \forall i$, see the following hypotheses from $\mathcal{H}$:

$$\forall j = 1,2,\dots,d:\ h_{2j-1}(x) = \begin{cases} 1 & x \geq b_j \\ -1 & x < b_j \end{cases}$$

$$\forall j = 1,2,\dots,d:\ h_{2j}(x) = \begin{cases} 1 & x \leq b_j \\ -1 & x > b_j \end{cases}$$

$$\forall j = 2d+1,\dots,4d-1:\ h_j(x) = -1$$

Therefore, for $\gamma = \frac{1}{4d-1} > 0$ it holds:

- If $y_i = +1$ then:

$$y_i \sum_{j=1}^{k} a_j h_j(x_i) = y_i \sum_{j=1}^{2d} \frac{1}{4d-1} h_j(x_i) + y_i \sum_{j=2d+1}^{4d-1} \frac{1}{4d-1} h_j(x_i)$$
$$= \frac{2d}{4d-1} - \frac{2d-1}{4d-1} = \frac{1}{4d-1} = \gamma$$

- If $y_i = -1$ then:

$$y_i \sum_{j=1}^{k} a_j h_j(x_i) = y_i \sum_{j=1}^{2d} \frac{1}{4d-1} h_j(x_i) + y_i \sum_{j=2d+1}^{4d-1} \frac{1}{4d-1} h_j(x_i) = 0 + \frac{2d-1}{4d-1}$$
$$\geq \frac{1}{4d-1} = \gamma$$

(4)

the objective $f(a) = \frac{1}{2}\|y - Xa\|_2^2 + \lambda\|a\|_1$. So, we can write f as:

$$f(a) = \frac{1}{2}\|y - Xa\|_2^2 + \lambda\|a\|_1 \Rightarrow f(a) = \sum_{j=1}^{d} -z_j a_j + \frac{1}{2}\sigma_j a_j^2 + \lambda|a_j|$$

Where $z_j = \sum_{i=1}^{n} y_i X_{ij}$.

- $\|y - Xa\|_2^2 = \langle y - Xa, y - Xa\rangle = \langle y, y\rangle - 2\langle y, Xa\rangle + \langle Xa, Xa\rangle = n - 2\sum_{j=1}^{d}\sum_{i=1}^{n} y_i X_{ij}\, a_j + a^T X^T X a = n + 2\sum_{j=1}^{d}\left(-\sum_{i=1}^{n} y_i X_{ij}\right)a_j + \sum_{j=1}^{d}\sigma_j a_j^2$
- $\lambda\|a\|_1 = \sum_{j=1}^{d}\lambda|a_j|$
- We can remove n from the objective function since the minimum does not depend on additive constant.

The derivative w.r.t a: $\nabla_j f(a) = -z_j + \sigma_j a_j + \lambda\delta_j$ where $\delta_j = \begin{cases} 1 & a_j > 0 \\ 0 & a_j = 0 \\ -1 & a_j < 0 \end{cases}$.

Since f is convex function in a then there is minimum where $\nabla_j f(a) = 0$. We shell show that $\hat{a}_j^{lasso} = \frac{sign(z_j)}{\sigma_j}\max(0, |z_j| - \lambda)$ gives $\nabla_j f(\hat{a}^{lasso}) = 0$.

Consider the first case where $sign(z_j) = +1$. So, $\hat{a}_j^{lasso} \geq 0$.

- If $\hat{a}_j^{lasso} = 0$ then $\delta_j = 0,\ z_j - \lambda \leq 0 \Longrightarrow \nabla_j f(\hat{a}_{lasso}) = -z_j + \sigma_j \frac{1}{\sigma_j}\max(0, z_j - \lambda) = 0$
- If $\hat{a}_j^{lasso} > 0$ then $\delta_j = 1,\ z_j - \lambda > 0 \Longrightarrow \nabla_j f(\hat{a}_{lasso}) = -z_j + \sigma_j \frac{1}{\sigma_j}\max(0, z_j - \lambda) + \lambda = \max(-(z_j - \lambda), 0) = 0$

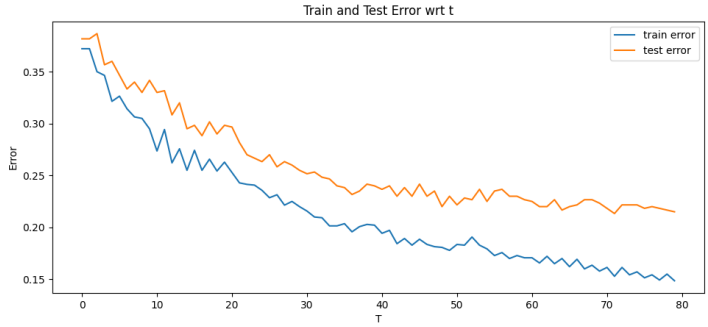
In the same manner, if $sign(z_j) = -1$, Then:

$\nabla_j f(\hat{a}_{lasso}) = -z_j + \sigma_j \frac{sign(z_j)}{\sigma_j}\max(0, |z_j| - \lambda) + \lambda\delta_j = -\max(0, -z_j - \lambda) - (z_j + \lambda) = -\max(-(z_j + \lambda), 0) = 0$

Second part:

(1)

(a) AdaBoost and plotting train and test error over iterations:



(b) Weak classifiers during the 10 iterations AdaBoost:

Weak classifier t=0: h_pred=1.0	h_index=26.0[bad]	theta=0.0
Weak classifier t=1: h_pred=-1.0	h_index=31.0[many]	theta=0.0
Weak classifier t=2: h_pred=-1.0	h_index=22.0[life]	theta=0.0
Weak classifier t=3: h_pred=1.0	h_index=311.0[worst]	theta=0.0
Weak classifier t=4: h_pred=-1.0	h_index=37.0[great]	theta=1.0
Weak classifier t=5: h_pred=1.0	h_index=372.0[boring]	theta=0.0
Weak classifier t=6: h_pred=-1.0	h_index=282.0[perfect]	theta=0.0
Weak classifier t=7: h_pred=1.0	h_index=292.0[supposed]	theta=0.0
Weak classifier t=8: h_pred=-1.0	h_index=196.0[performances]	theta=0.0
Weak classifier t=9: h_pred=1.0	h_index=88.0[script]	theta=0.0

Weak classifiers - expect to help to classify reviews:

All the blue weak classifiers are good ones because each of them represents the review as a good or bad review based on specific words. For positive words like "great" or "perfect," the algorithm chooses a classifier that return a value of -1 if these words appear once or not at all. On the other hand, for negative words like "bad," "worst," or "boring," the classifier return a value of 1 only if each of these words appears once in the review.

Weak classifiers - not expect to help to classify reviews:

The orange weak classifiers are surprising because each of them utilizes generic words like "many," "supposed," or "script" that don't have a strong inherent meaning for determining good or bad reviews. the algorithm associates the words "supposed" and "script" with negative reviews, while considering "many" as a word associated with positive reviews.

**Section C results didn't came up so well, so I decided not to include them.