DS310 HW2

Zhixuan Yang

October 15, 2019

1. Let W_t be the weight vector after t weight updates.

Let w^* be such that

$\forall X_n \in S^+, w^* \cdot X_n \geq \delta$ and $\forall X_n \in S^-, w^* \cdot X_n \leq -\delta$.

We also assume that $w^* \cdot X = 0$ passes through the origin.

$\because$ training set E is linearly separable iff

$\exists w^*$, such that $\forall X_p \in S^+, w^* \cdot X_p \geq \delta$

and $\forall X_p \in S^-, w^* \cdot X_p \leq -\delta$

$\therefore$ Whenever $w^* \in R^{n+1}$ and $\delta > 0$ exist, training set E is linearly separable (A)

Let $\forall X_{n_1} \in S^+, Z_n = X_n$, $\forall X_n \in S^-, Z_n$

$\forall X_{n_1} \in S^-, Z_n = -X_n$,

$Z = \{Z_n\}$

$(\forall X_n \in S^+, w^* \cdot X_n \geq \delta \ \& \ \forall X_n \in S^-, w^* \cdot X_n \leq -\delta)$

$\Leftrightarrow (\forall Z_n \in Z, w^* \cdot Z_n \geq \delta)$

Let $E' = \{(Z_n, 1)\}$, $W_{t+1} = W_t + \eta(d_n - y_n)Z_n$, where η is the learning rate and d is the desired output, y is the real output, learning rate $\eta > 0$, and $W_0 = [0\ 0\ \dots\ 0]^T$

since [weight update based on example $(Z_n, 1)$] $\Leftrightarrow [(d_n = 1) \wedge (y_n = -1)]$

$\therefore w^* \cdot W_{t+1} = w^* \cdot (W_t + 2\eta Z_n) = (w^* \cdot W_t) + 2\eta(w^* \cdot Z_n)$

$\because \forall Z_n \in Z, (w^* \cdot Z_k \geq \delta), w^* \cdot W_{t+1} \geq w^* \cdot W_t + 2\eta\delta$

$\therefore \forall t, w^* \cdot W_t \geq 2t\eta\delta$ (B)

$$\|W_{t+1}\|^2 = W_{t+1} \cdot W_{t+1}$$
$$= (W_t + 2\eta Z_n) \cdot (W_t + 2\eta Z_n)$$
$$= (W_t \cdot W_t) + 4\eta (W_t \cdot Z_n) + 4\eta^2 (Z_n \cdot Z_n)$$

Note that given any finite training set, the length of the training patterns is bounded. That is, $\forall Z_n, \|Z_n\| \leq L$. where $L = \max_{X_n \in S} \|X_n\|$ and $S = S^+ \cup S^-$

$$\therefore \|W_{t+1}\|^2 \leq (\|W_t^2\|) + 4\eta^2 \|Z_n\|^2 \leq \|W_t\|^2 + 4\eta^2 L^2$$

Hence $\|W_t\|^2 \leq 4t\eta^2 L^2$

$$\therefore \forall t \ \|W_t\| \leq 2\eta L \sqrt{t} \quad (C)$$

From (b) we have: $\forall t (W^* \cdot W_t) \geq 2t\eta\delta \Rightarrow \{\forall t \ 2t\eta\delta \leq W^* \cdot W_t)\} \Rightarrow$

$\{\forall t \ 2t\eta\delta \leq \|W^*\| \|W_t\| \cos\theta\} \Rightarrow \{\forall t \ 2t\eta\delta \leq \|W^*\| \|W_t\|)\}$

$\because \forall \theta \ \cos\theta \leq 1$, substituting from for an upper bound on $\|W_t\|$ from (C)

$\forall t \{2t\eta\delta \leq \|W^*\| 2\eta L\sqrt{t}\} \Rightarrow \{\forall t (\delta\sqrt{t} \leq \|W^*\| L)\} \Rightarrow t \leq \left(\frac{\|W^*\| L}{\delta}\right)^2$ (D)

From formula (D), We have prove that the bound on the number of weight updates does not depend on the learning rate, and the training set is linear separable from (A). This conversence theorm holds for any bounded learning late $\eta > 0$, and there is a condition that "$0 < A \leq \eta(t) \leq B$", thus this theorm is always suitable for this case.

Above, we have shown that learning rate does not directly influence the number of weight updates. Now, we have to prove that variant learning rates is between $0 < A \leq \eta(t) \leq B$, where A and B are fixed lower and upper bounds.

The Perceptron algorithm changes We as follows:

$W_{t+1} \leftarrow W_t \pm \eta X_n$

suppose we choose η such that at each weight update,

$W_{t+1} X_n = (W_t + \eta_t X_n) \cdot X_n$

$W_{t+1} \cdot X_n > 0$ if $W_t X_n < 0$ and $C_n = 1$,

$\therefore \eta(t) = \left[\frac{|W_t \cdot X_n|}{X_n \cdot X_n} \right]$ (E)

since the training set is finite, therefore, this equation proves that $0 < A \leq \eta(t) \leq B$, where A and B are fixed lower and upper bounds respectively. All in All, We have Prove that the training data are linearly separable from (A), and the number of weight updates does not depend on learning rate from (D), and $0 < A \leq \eta(t) \leq B$ from (E), Thus, the resulting variant of the Perceptron learning algorithm is guaranteed to terminate with a weight vector.

Additionally, the Perceptron algorithm is quite robust as a learning model. It can be shown that the algorithm converges even when η is allowed to fluctuate arbitrarily over time as long as $0 < \eta_t \leq B$ where B is the upper bound on the learning rate, since We have prove that $0 < A \leq \eta(t) \leq B$, therefore, this proves that η is allowed to fluctuate arbitrarily.

2. Ordered 4-tuples of attributes values corresponding to

Outlook (Sunny, Overcast, Rain)
Temperature (Hot, Mild, Cool)
Humidity (High, Normal)
Wind (Strong, Weak)

and we have a class label PlayTennis, and it corresponds to

PlayTennis (Yes (+), No (-))

and we have 14 instances from Training Data

Instance	class label	
I_1 (S, H, I, W)	—	$\hat{H}(X) = -\frac{5}{14}\log_2\frac{5}{14} - \frac{9}{14}\log_2\frac{9}{14} \approx 0.940$
I_2 (S, H, I, S)	—	$\hat{H}(X \mid \text{Outlook}=S) = -\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5} = 0.971$
I_3 (O, H, I, W)	+	$\hat{H}(X \mid \text{Outlook}=O) = -\frac{0}{4}\log_2\frac{0}{4} - \frac{4}{4}\log_2\frac{4}{4} = 0$
I_4 (R, M, I, W)	+	$\hat{H}(X \mid \text{Outlook}=R) = -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5} = 0.971$
I_5 (R, C, N, W)	+	$\hat{H}(X \mid \text{Outlook}) = \frac{5}{14}(0.971) + \frac{4}{14}(0) + \frac{5}{14}(0.971) = 0.694$
I_6 (R, C, N, S)	—	$I(\text{Outlook}) = 0.940 - 0.694 = 0.246$
I_7 (O, C, N, S)	+	$\hat{H}(X \mid \text{Temperature}=H) = -\frac{2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4} = 1$
I_8 (S, M, I, W)	—	$\hat{H}(X \mid \text{Temperature}=M) = -\frac{2}{6}\log_2\frac{2}{6} - \frac{4}{6}\log_2\frac{4}{6} = 0.918$
I_9 (S, C, N, W)	+	$\hat{H}(X \mid \text{Temperature}=C) = -\frac{1}{4}\log_2\frac{1}{4} - \frac{3}{4}\log_2\frac{3}{4} = 0.811$
I_{10} (R, M, N, W)	+	$\hat{H}(X \mid \text{Temperature}) = \frac{4}{14}(1) + \frac{6}{14}(0.918) + \frac{4}{14}(0.81) = 0.911$
I_{11} (S, M, N, S)	+	$I(\text{temperature}) = 0.940 - 0.911 = 0.029$
I_{12} (O, M, I, S)	+	
I_{13} (O, H, N, W)	+	$\hat{H}(X \mid \text{Humidity}=I) = -\frac{4}{7}\log_2\frac{4}{7} - \frac{3}{7}\log_2\frac{3}{7} = 0.985$
I_{14} (R, M, I, S)	—	$\hat{H}(X \mid \text{Humidity}=N) = -\frac{1}{7}\log_2\frac{1}{7} - \frac{6}{7}\log_2\frac{6}{7} = 0.592$

$\hat{H}(\text{Humidity}) = \frac{7}{14}*(0.985) + \frac{7}{14}(0.592) = 0.789$

$I(\text{Humidity}) = 0.940 - 0.789 = 0.151$

$\hat{H}(X|Wind=t) = -\frac{3}{6}\log_2\frac{3}{6} - \frac{3}{6}\log_2\frac{3}{6} = 1$

$\hat{H}(X|Wind=W) = -\frac{6}{8}\log_2\frac{6}{8} - \frac{2}{8}\log_2\frac{2}{8} = 0.811$

$\hat{H}(X|Wind) = \frac{6}{14}*(1) + \frac{8}{14}(0.811) = 0.892$

$I(Wind) = 0.940 - 0.892 = 0.048$

Based on the results, Outlook has the largest information gain, therefore Outlook is the root of this tree.
since overcast of outlook is Pure, which means it has 4 for Yes, and 0 for No, and then we don't need to think about this.
Then, we have find the child under sunny,

$\hat{H}(temperature=h|S) = -\frac{2}{2}\log_2\frac{2}{2} - \frac{0}{2}\log_2\frac{0}{2} = 0.$

$\hat{H}(temperature=m|S) = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2} = 1$

$\hat{H}(temperature=c|S) = -\frac{0}{1}\log_2\frac{0}{1} - \frac{1}{1}\log_2\frac{0}{1} = 0$

$I(temperature|S) = 0.971 - \frac{2}{5}*0 - \frac{2}{5}*1 - \frac{1}{5}*0 = 0.571$

$\hat{H}(Humidity=h|S) = -\frac{3}{3}\log_2\frac{3}{3} - \frac{0}{3}\log_2\frac{0}{3} = 0.$

$\hat{H}(Humidity=n|S) = -\frac{0}{2}\log_2\frac{0}{2} - \frac{2}{2}\log_2\frac{2}{2} = 0$

$I(Humidity|S) = 0.971 - 0 - 0 = 0.971$

$\hat{H}(Wind=t|S) = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2} = 1$

$\hat{H}(Wind=W|S) = -\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3} = 0.918$

$I(Wind|S) = 0.971 - (\frac{2}{5})*1 - (\frac{3}{5})(0.918) = 0.0202$

Therefore, Humindity is the child of sunny because it has largest information gain.
Next, we have to find child of rain.

$\hat{H}(\text{temperature}=h|r) = 0$

$\hat{H}(\text{temperature}=m|r) = -\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3} = 0.918$

$\hat{H}(\text{temperature}=c|r) = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2} = 1$

$I(\text{temperature}|r) = 0.971 - \frac{3}{5}(0.918) - (\frac{2}{5})(1) = 0.0202$

$\hat{H}(\text{Humidity}=i|r) = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2} = 1$

$\hat{H}(\text{Humidity}=n|r) = -\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3} = 0.918$

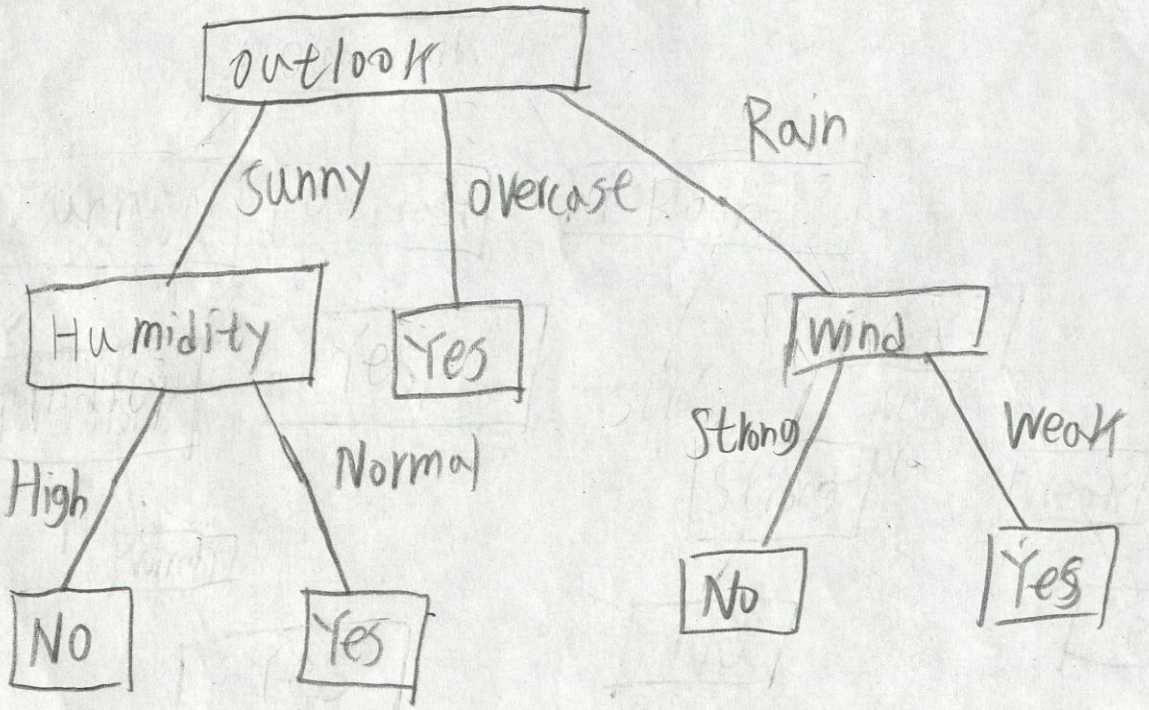
$I(\text{Humidity}|r) = 0.971 - \frac{3}{5}(0.918) - \frac{2}{5}(1) = 0.0202$

$\hat{H}(\text{wind}=t|r) = -\frac{2}{2}\log_2\frac{2}{2} - \frac{0}{2}\log_2\frac{0}{2} = 0$

$\hat{H}(\text{wind}=w|r) = -\frac{0}{3}\log_2\frac{0}{3} - \frac{3}{3}\log_2\frac{3}{3} = 0$

$I(\text{wind}|r) = 0.971 - 0 - 0 = 0.971$

Based on the results, wind is the child of rain because it has biggest information gain. Finally, our decision tree should be the following



3. To construct a decision tree to minimize the cost of correct classification of data samples, my assumption is that we use marginal information gain instead of information gain. The function of marginal information gain is $\frac{I(F)^2}{C(F)}$, where C is the cost of measuring feature F, and I is the information gain of feature F. The run-time cost of a feature at a position is the sum of cost of seeing the feature and the cost of moving to its position. And below is the original information gain of each feature we got from Problem 2.

I(outlook) = 0.246 I(temperature) = 0.029 I(Humidity) = 0.15)

I(wind) = 0.048

And cost of each feature is given.

Cost(outlook) = 1 cost(temperature) = 2 Cost(Humidity) = 2.

cost(wind) = 3.

The marginal information gains of each feature are below

$MI(outlook) = \frac{(0.246)^2}{1} = 0.0605$ $MI(temperature) = \frac{0.029^2}{2} = 0.00042$

$MI(Humidity) = \frac{(0.151)^2}{2} = 0.0114$ $MI(wind) = \frac{0.048^2}{3} = 0.00077$

Based on the above results, outlook is the root of cost-effective decision tree, because it has biggest marginal information gain.

since overcast of outlook has 4 Yes and 0 No, so, the "Yes" would be the branch of the tree and stop here.

Let's find the first child of Sunny. and We have already got the information gain under "sunny, and they are,

I(temperature/s) = 0.571 I(Humidity/s) = 0.971, I(wind/s) = 0.0202.

And Because of the formula of marginal information gain $\frac{I(S,A)^2}{C(A)}$, then the marginal information gain of each attribute under sunny is.

$MI(temperature/s) = \frac{(0.571)^2}{1} = 0.326$ $MI(Humidity/s) = \frac{(0.971)^2}{1} = 0.943$

$MI(Wind|S) = \frac{(0.0202)^2}{1} = 0.00041$

Based on the above results, Humidity should be the child of Sunny.

Let's calculate the child of rain. the original information gains of each features under rain from problem 2 are.

$I(temperature|r) = 0.0202$, $I(Humidity|r) = 0.0202$, $I(wind|r) = 0.971$

And, the marginal information gains are

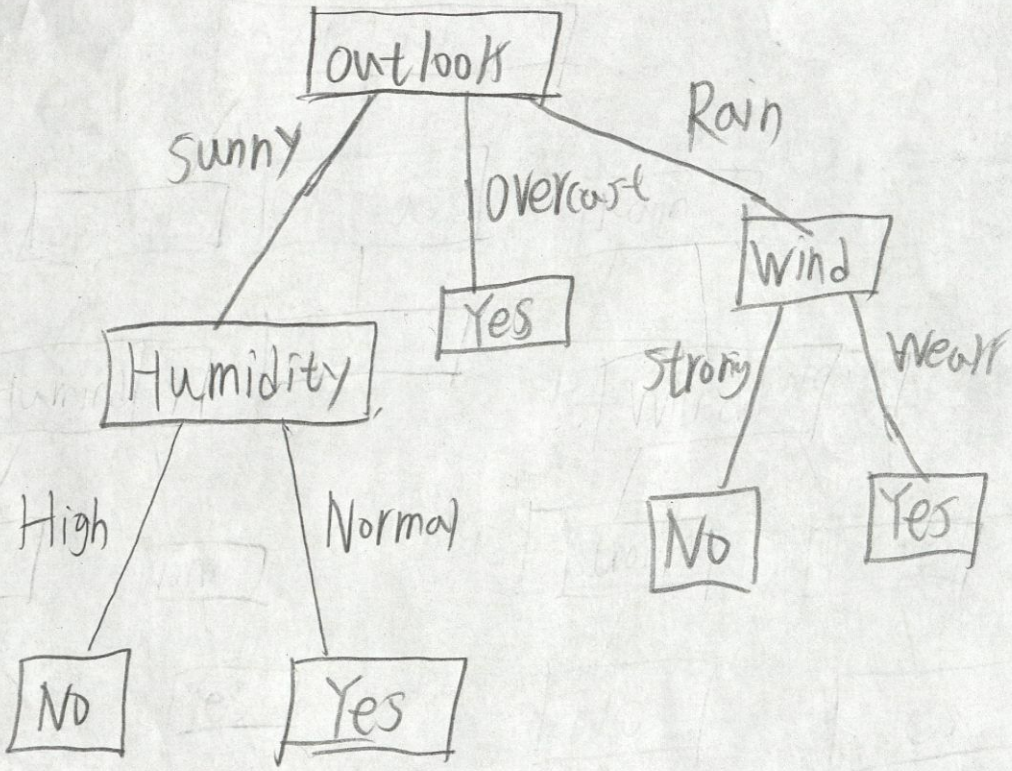
$MI(temperature|r) = \frac{(0.0202)^2}{1} = 0.0004$

$MI(Humidity|r) = (0.0202)^2 = 0.00041$

$MI(wind|r) = \frac{(0.971)^2}{1} \frac{1}{} = 0.943.$

Therefore, Wind should be the child of rain because it has the largest marginal information gain.

Therefore, the cost-effective decision tree should look like the following



4. (a) I think Random Forest is the best choice for this Problem because of following reasons. (1) the data is mixed, which means this dataset has either numerical variable or categorical, and random forest algorithm is good at deal with this situation. (2) the task is to performing accurate diagnosis of patients, and random forest is suiting for this kind of classification. (3) Additional task is that "to obtain insight regarding the relationships between features for different diseases.", and random forest learning algorithm has an advantage that it provides the feature importance after training, therefore, we can obtain the insight from feature importance using random forest. Above three reasons are why I prefer random forests algorithm for this problem.

Random forests.

(b) I think the random forests algorithm is also applicable for this Problem because of following reasons. (1) Since the dataset is very large the random forest algorithm would divide the very large dataset into several subsets, therefore, we don't have to worry about the size problem. (2) our task is Predicting users' interest. Thus, we can predict the user's interet based on the decision trees created by random forest. (3) Variables could be mixed, and Random forests algorithm is suitable for Processing mixed data. overall, the choice of algorithm should depending on the real situation. But, ideally, I think random forests algorithm is the best choice for this problem

Random forests.

(c) I think Decision Tree is the best choice for this case. First, the problem says "they can automate the decision to approve or deny a loan based on a "simple rule,", and the key word is simple. Among the four algorithms, perceptron algorithm and naive bayes algorithm require mathematical inference, so they do not meet the "simple rule" requirment. Thus, we have to compare decision tree and random forest. Another requirment of this problem is "the decision-making process be transparent,", and one of disadvantage of random forest is that this algorithm is very difficult to interpret the resulting model because of its randomness. For instance, when a client asks us why we deny his request, we can show the model of decision tree algorithm to him, and tell him which part caused his request to be rejected. If we are using the same problem using random forest, we wouldn't be able to quickly find out which part was causing this result because random forest is a collection of many decision trees, and then we don't have a easy way to figure out which tree made this prediction. All in all, the "simple rule" requirment and the "decison-making process be transparent" requirment lead to decision tree algorithm is the best choice for this problem.

Decision Tree