Yet another notes on Decision Tree

Richard Xu

August 15, 2022

1 Introduction

Despite the thousands of decision tree lecture notes, I still wanted to write one that suits my own taste. I also want to add the χ^2 test into the mix. This decision tree notes then gave me an excuse to do this because the χ^2 test was used as one of the splitting methods.

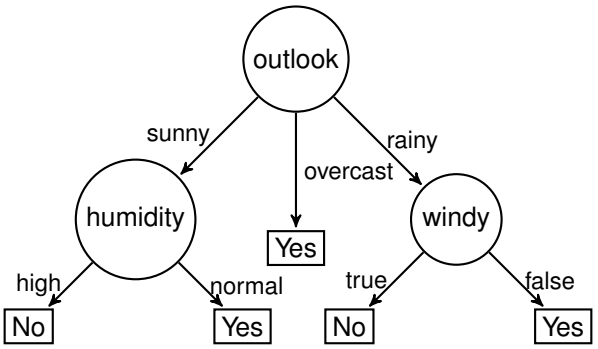
2 Decision tree

imagine you are given a training dataset, and the goal is to create classier to classify unseen data:

Table 1: database

outlook	**temperature**	**humidity**	**windy**	**play**
sunny	hot	high	FALSE	no
sunny	hot	high	TRUE	no
overcast	hot	high	FALSE	yes
rainy	mild	high	FALSE	yes
rainy	cool	normal	FALSE	yes
rainy	cool	normal	TRUE	no
overcast	cool	normal	TRUE	yes
sunny	mild	high	FALSE	no
sunny	cool	normal	FALSE	yes
rainy	mild	normal	FALSE	yes
sunny	mild	normal	TRUE	yes
overcast	mild	high	TRUE	yes
overcast	hot	normal	FALSE	yes
rainy	mild	high	TRUE	no

We are now trying to solve it using the decision tree approach. Firstly, we can build the tree in the following way:



Note that we may not need all the features involved in order to build the tree

2.1 terminologies

First, there are some terms that shouldn't be hard to remember since they're just standard tree definitions:

1. **root node**
   (a) first node of the tree
   (b) No incoming edges and zero or more outgoing edges
   (c) for example, in the example, it would be the node "outlook"
2. **internal nodes**
   (a) Has exactly one incoming edge and two or more outgoing edges
   (b) Contain attribute test
   (c) for example, it is "humidity" and "windy" nodes
3. **leaf/terminal nodes**
   (a) has exactly one incoming edge
   (b) and no outgoing edges and is assigned a class label
   (c) for example, all the "Yes" and "no" are the leaf nodes

2.2 inference of new data using the tree

Once you have established the tree, then the decision can be made by passing a new rows of data, for example:

Table 2: database

outlook	**temperature**	**humidity**	**windy**	**play**
rainy	hot	normal	FALSE	?

if we trace through the decision tree, then, the answer should be YES:

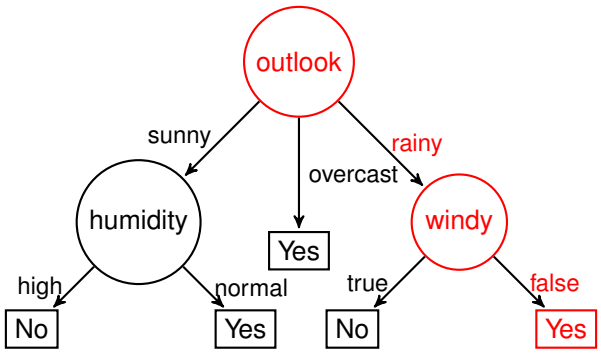


Figure 1: classification inference on new data

2.3 Effect of selecting a node/feature

Obviously, there isn't a unique way to create a tree, right? So, which tree is better for our classification purposes? For example, which node should we use as the root node?

Maybe let's look at a numerical example that can help us make the effect of node splitting more visually obvious:

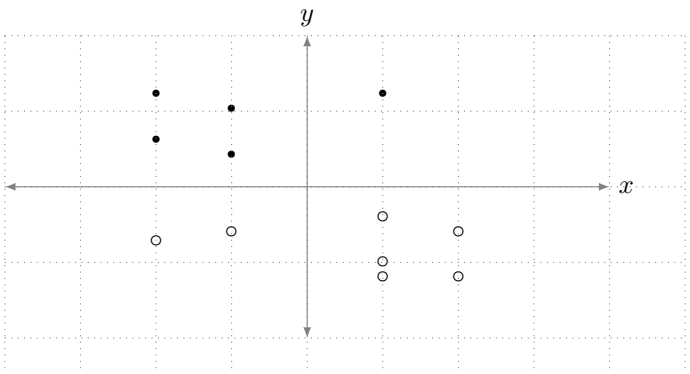


Figure 2: binary classification data-set with two features x and y

2.3.1 case one

if we select the y feature and split the y-axis into two regions: upper and lower:

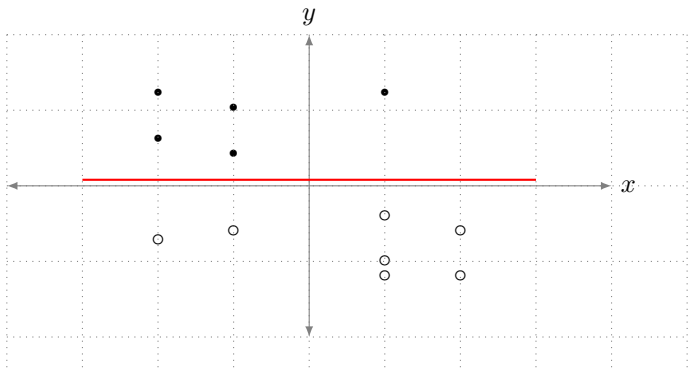


Figure 3: split at y feature

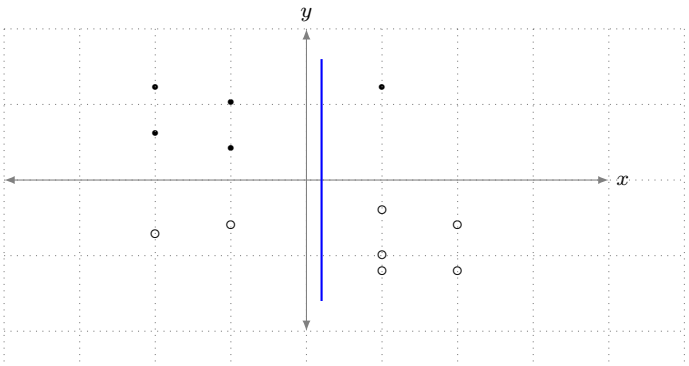
In here, it can be seen that after the split, each portion of points contains only one type of circles. In terms of probabilities, it means:

$$\begin{cases} \Pr(\text{circle} = \text{filled}) = \frac{5}{5} = 1 & \Pr(\text{circle} = \text{empty}) = \frac{0}{5} = 0 \quad \text{for upper portion} \\ \Pr(\text{circle} = \text{filled}) = \frac{0}{7} = 0 & \Pr(\text{circle} = \text{empty}) = \frac{7}{7} = 1 \quad \text{for lower portion} \end{cases} \tag{1}$$

note that in each regions, the data types are "pure".

2.3.2 case two

we can not perform the "pure" split, if we were trying to split the x-features:



therefore, after the split on x-axis, we can see:

$$\begin{cases} \Pr(\text{circle} = \text{filled}) = \frac{4}{6} & \Pr(\text{circle} = \text{empty}) = \frac{2}{6} \quad \text{for left portion} \\ \Pr(\text{circle} = \text{filled}) = \frac{1}{6} & \Pr(\text{circle} = \text{empty}) = \frac{5}{6} \quad \text{for right portion} \end{cases} \tag{2}$$

note that in each regions, the data types are "impure".

2.3.3 what is our leftover task?

Comparing the two cases, we can see that the first case is preferred because the feature split in y-axis gives us very high certainty that future data will be classified correctly than the second split, i.e., we want to have "purity" of data types after the split.

Comparing the two cases, we can see that the first case is preferred because the feature split on the y-axis gives us very high certainty that future data will be correctly classified instead of the second split on x-axis, i.e. we want choose the method of split in order to have the "purity" of data types.

So obviously we need to find an appropriate way to measure attribute selection, or the purity of the data type after the split. For example, what is a single number to represent the purity of:

$$\begin{aligned} &\Big(\Pr\big(X = \text{filled}\big) = \frac{4}{6},\ \Pr\big(X = \text{empty}\big) = \frac{2}{6}\Big) \\ \text{versus} \quad &\Big(\Pr\big(X = \text{filled}\big) = 1.0,\ \Pr\big(X = \text{empty}\big) = 0.0\Big) \end{aligned} \tag{3}$$

Today, we will discuss two types of measures:

1. Information Gain
2. Gini Index

which becomes our topic of **Attribute Selection Measures (Splitting Rules)**. Then in the next lecture, we will discuss various algorithms of how to construct the Decision Tree.

3 Information Gain

Before discussing information gain, let's take a look some of the background, in particular what is an entropy?

3.1 What is Entropy?

Remember, you've seen it when we discussed T-SNE (though I didn't ask you to remember Entropy back then). We also talked about the term Cross-Entropy in classification function last week (which we will discuss more when we talk about neural networks). However, you need to remember Entropy now!

3.1.1 Sushi Party Example

Let's take an example: imagine we will have a sushi party (just like what the department organizes every year) - you can see last year's photos (before Omicron of course!):
`https://www.math.hkbu.edu.hk/cafe/231121/photo/` (btw, see if you can spot a photo contains me!). Imagine that the students are sitting on four different tables:

1. **table 1 :** everyone says YES to sushi.
2. **table 2 :** 19 students says YES to sushi, 1 says NO to sushi.
3. **table 3 :** 10 students says YES to sushi, 10 says NO to sushi.
4. **table 4 :** 1 students says YES to sushi, 19 says NO to sushi.

So the question is, how "surprised" would you be if you picked a random person at each table and asked the student if he/she would say "yes" (or "no") to sushi?
The measure of "surprise" is also the "information".
I believe you may have the following reaction summarized in Table **??** below:

Table 3: surprised table

	love sushi (YES)	love sushi (NO)
table 1	shouldn't be any surprise!	infinitely surprised
table 2	not surprised	surprised
table 3	fairly surprised	fairly surprised
table 4	surprised	not surprised

3.2 how do we measure "surprise"?

Clearly, "surprise" is inversely proportional to probability, i.e. the more often an event occurs, the less surprised people feel about receiving the outcome of that event.

For example, if people at a table mostly say "yes" to sushi, it wouldn't be surprising if you randomly asked people at that table what they thought about sushi, and got a "yes" answer.

You can also think of "surprise" as a measure of randomness, i.e. the higher the probability, the less randomness.

So let's try to define "surprise" in terms of P, which is the probability of an event.

3.2.1 First try

$$\text{surprise} = \frac{1}{P} \tag{4}$$

Then the "surprise" table should become:

Table 4: surprised table with $\frac{1}{P}$

	love sushi (YES)	love sushi (NO)
table 1	$\frac{1}{\frac{20}{20}} = 1$	$\frac{1}{\frac{0}{20}} = \infty$
table 2	$\frac{1}{\frac{19}{20}} = \frac{20}{19}$	$\frac{1}{\frac{1}{20}} = 20$
table 3	$\frac{1}{\frac{10}{20}} = 2$	$\frac{1}{\frac{10}{20}} = 2$
table 4	$\frac{1}{\frac{1}{20}} = 20$	$\frac{1}{\frac{19}{20}} = \frac{20}{19}$

While it looks okay, the numbers aren't perfect, e.g.,:

1. In table 1, since everyone said YES to sushi, it should be 0 surprise (i.e., no surprise at all) to see a randomly selected student from that table saying YES to sushi, but instead we get 1.

To circumvent this problem, we now change our measure of surprise to:

3.2.2 Second try

Let's put a $\log(\cdot)$ around it. Note that since we have two possible outcomes, we need the base 2 $\log$:

$$\begin{aligned} \text{surprise} &= \log_2\left(\frac{1}{P}\right) \\ &= -\log_2(P) \end{aligned} \tag{5}$$

and then we plot the numbers in Table **??** below:

Table 5: surprised table with $-\log_2(P)$

	love sushi (YES)	love sushi (NO)
table 1	$-\log_2\left(\frac{20}{20}\right) = 0$	$-\log_2\left(\frac{0}{20}\right) =$ undefined
table 2	$-\log_2\left(\frac{19}{20}\right) = 0.074$	$-\log_2\left(\frac{1}{20}\right) = 4.321$
table 3	$-\log_2\left(\frac{10}{20}\right) = 1$	$-\log_2\left(\frac{10}{20}\right) = 1$
table 4	$-\log_2\left(\frac{1}{20}\right) = 4.321$	$-\log_2\left(\frac{19}{20}\right) = 0.074$

It looks so much better! But remember, we have both $P_{\text{(YES)}}$ and $P_{\text{(NO)}}$, so how can we combine them? We can't just add two numbers to each table. Why? Because each of these events is associated with a different probability. So we should change the formula to:

$$\text{surprise} = -P_{\text{(YES)}} \log(P_{\text{(YES)}}) + -P_{\text{(NO)}} \log(P_{\text{(NO)}}) \tag{6}$$

This is the (weighted) average of surprises. This is entropy. We use the notation H, and many literary works actually use the notation E.

In the case P defines over multiple outcomes in an outcome space $\mathcal{S}$, then we can write Entropy H as:

$$H(P) = -\sum_{x \in \mathcal{S}} P(X = x) \log(P(X = x)) \tag{7}$$

therefore, for the sushi problem we have, the entropy of:

$$H = -P_{\text{(YES)}} \log(P_{\text{(YES)}}) + -P_{\text{(NO)}} \log(P_{\text{(NO)}}) \tag{8}$$

plotted in table below:

Table 6: Entropy for each tables

table 1	$-\log_2\left(\frac{20}{20}\right) \times \frac{20}{20} - \log_2\left(\frac{0}{20}\right) \times \frac{0}{20} = 0$
table 2	$-\log_2\left(\frac{19}{20}\right) \times \frac{19}{20} - \log_2\left(\frac{1}{20}\right) \times \frac{1}{20} = 0.074 \times \frac{19}{20} + 4.321 \times \frac{1}{20} = 0.286$
table 3	$-\log_2\left(\frac{10}{20}\right) \times \frac{10}{20} - \log_2\left(\frac{10}{20}\right) \times \frac{10}{20} = 1 \times 0.5 + 1 \times 0.5 = 1$
table 4	$-\log_2\left(\frac{1}{20}\right) \times \frac{1}{20} - \log_2\left(\frac{19}{20}\right) \times \frac{19}{20} = 4.321 \times \frac{1}{20} + 0.074 \times \frac{19}{20} = 0.286$

It is not surprising to see that the maximum entropy occurs when $P_{\text{(YES)}} = \frac{1}{2}$ and $P_{\text{(NO)}} = \frac{1}{2}$ (with entropy being 1) and least Entropy occurs when one of the probabilities is 1 (with entropy being 0). You may obtain this fact from basic calculus.

3.2.3 For Reference: $\log$ base conversion formula

$\log_a x$ converts to base b by:

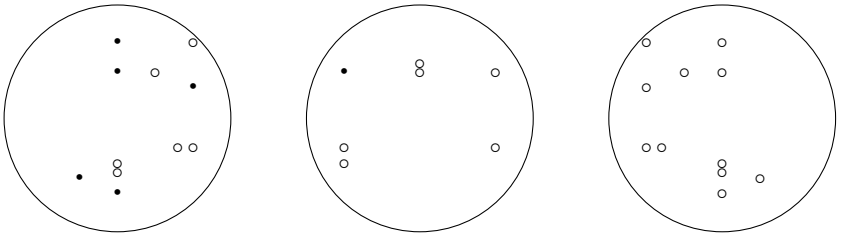
$$\frac{\log_b x}{\log_b a} \tag{9}$$

for example:

$$\log_3 7 = \frac{\log_{10} 7}{\log_{10} 3} = 1.771 \tag{10}$$

3.3 Apply Entropy to Decision tree

Firstly, we have more definitions for each of the cases:



Impure Slightly impure Pure

So obviously after you split a node, you want it from impure to as big as possible. But how much of this "change" is measurable? I'll show you the formula for the 2-branch case (left and right). Of course, it's easy to generalize this to the multi-branch case.

$$\text{GAIN} = H_{\text{parents}} - \left(\Pr_{\text{left}} H_{\text{left}} + \Pr_{\text{right}} H_{\text{right}}\right) \tag{11}$$

Instead of dig into each of the symbols, it is best to be illustrated through a small dataset:

Table 7: small dataset to illustrate Information Gain

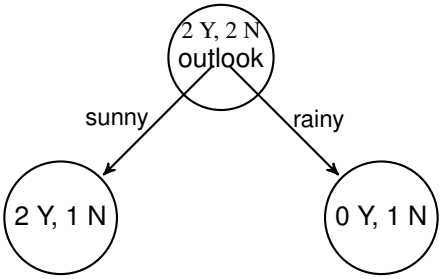
outlook	**temperature**	**humidity**	**play**
sunny	cool	high	yes
sunny	cool	low	yes
rainy	hot	high	no
sunny	hot	low	no

3.3.1 why Information Gain instead of just Entropy

Note that all the examples below will have the same entropy for the parents since they (each of the attributes) are using entire data of the same table. However, when considering nodes of different branches, they may not have the same entropy.

It is also time to discuss the advantage of using information GAIN instead of just looking at the Entropy alone. We can consider diminishing returns here. For example, when a node already has very low entropy (from which we can already make decisions), it seems unnecessary to split it further. So this makes GAIN a very suitable metric for deciding whether a node should be split.

3.3.2 split on outlook



1. before split

$$\begin{aligned} H_{\text{parents}} &= -P_{(\text{Y})}\log_2(P_{(\text{Y})}) + -P_{(\text{N})}\log_2(P_{(\text{N})}) \\ &= -\frac{1}{2}\log_2\left(\frac{1}{2}\right) + -\frac{1}{2}\log_2\left(\frac{1}{2}\right) \\ &= 1 \end{aligned} \tag{12}$$

2. after split:
   left child: $2Y, 1N$:

$$\begin{aligned} H_{\text{left}} &= -P_{(\text{Y})}\log_2(P_{(\text{Y})}) + -P_{(\text{N})}\log_2(P_{(\text{N})}) \\ &= -\frac{2}{3}\log_2\left(\frac{2}{3}\right) + -\frac{1}{3}\log_2\left(\frac{1}{3}\right) \\ &= 0.918 \end{aligned} \tag{13}$$

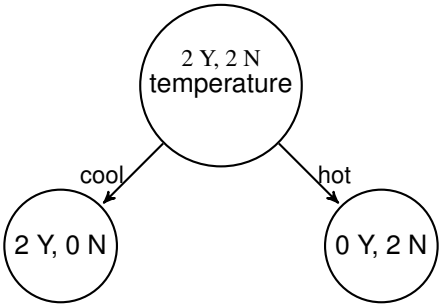
   right child: $1N$:

$$H_{\text{right}} = 0 \tag{14}$$

3. combine them together:

$$\begin{aligned} \text{GAIN} &= 1 - \left(\text{Pr}_{\text{left}} H_{\text{left}} + \text{Pr}_{\text{right}} H_{\text{right}}\right) \\ &= 1 - \left(\frac{3}{4}H_{\text{left}} + \frac{1}{4}H_{\text{right}}\right) \\ &= 1 - \left(\frac{3}{4} \times 0.918 + \frac{1}{4} \times 0\right) \\ &= 0.311 \end{aligned} \tag{15}$$

3.3.3 split on temperature



1. before split

$$\begin{aligned} H_{\text{parents}} &= -P_{(\text{Y})} \log_2(P_{(\text{Y})}) + -P_{(\text{N})} \log_2(P_{(\text{N})}) \\ &= -\frac{1}{2} \log_2 \left(\frac{1}{2}\right) + -\frac{1}{2} \log_2 \left(\frac{1}{2}\right) \\ &= 1 \end{aligned} \tag{16}$$

2. after split:
   left child: $2Y, 0N$:

$$H_{\text{left}} = 0 \tag{17}$$

   right child: $0Y, 2N$:

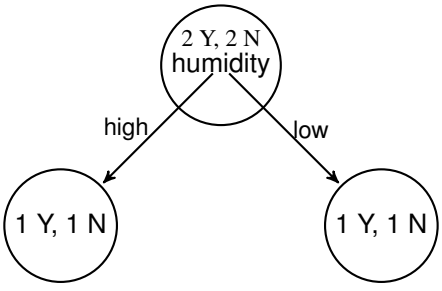
$$H_{\text{right}} = 0 \tag{18}$$

3. combine them together:

$$\begin{aligned} \text{GAIN} &= 1 - \left(\text{Pr}_{\text{left}} H_{\text{left}} + \text{Pr}_{\text{right}} H_{\text{right}}\right) \\ &= 1 - \left(\frac{1}{2} H_{\text{left}} + \frac{1}{2} H_{\text{right}}\right) \\ &= 1 - \left(\frac{1}{2} \times 0 + \frac{1}{2} \times 0\right) \\ &= 1 \end{aligned} \tag{19}$$

   which is the most possible GAIN

3.3.4 split on humidity



1. before split

$$\begin{aligned} H_{\text{parents}} &= -P_{(\text{Y})}\log_2(P_{(\text{Y})}) + -P_{(\text{N})}\log_2(P_{(\text{N})}) \\ &= -\frac{1}{2}\log_2\left(\frac{1}{2}\right) + -\frac{1}{2}\log_2\left(\frac{1}{2}\right) \\ &= 1 \end{aligned} \tag{20}$$

2. after split:
   left child: $2Y, 1N$:

$$\begin{aligned} H_{\text{left}} &= -P_{(\text{Y})}\log_2(P_{(\text{Y})}) + -P_{(\text{N})}\log_2(P_{(\text{N})}) \\ &= -\frac{1}{2}\log_2\left(\frac{1}{2}\right) + -\frac{1}{2}\log_2\left(\frac{1}{2}\right) \\ &= 1 \end{aligned} \tag{21}$$

   right child: $1N$:

$$\begin{aligned} H_{\text{left}} &= -P_{(\text{Y})}\log_2(P_{(\text{Y})}) + -P_{(\text{N})}\log_2(P_{(\text{N})}) \\ &= -\frac{1}{2}\log_2\left(\frac{1}{2}\right) + -\frac{1}{2}\log_2\left(\frac{1}{2}\right) \\ &= 1 \end{aligned} \tag{22}$$

3. combine them together:

$$\begin{aligned} \text{GAIN} &= 1 - \Big(\Pr_{\text{left}} H_{\text{left}} + \Pr_{\text{right}} H_{\text{right}}\Big) \\ &= 1 - \Big(\frac{2}{4} H_{\text{left}} + \frac{2}{4} H_{\text{right}}\Big) \\ &= 1 - \Big(\frac{1}{2} \times 1 + \frac{1}{2} \times 1\Big) \\ &= 0 \end{aligned} \tag{23}$$

   which is the least possible GAIN

3.4 more generalized definition of information gain

after you look at the two branch case of:

$$\text{GAIN} = H_{\text{parents}} - \Big(\Pr_{\text{left}} H_{\text{left}} + \Pr_{\text{right}} H_{\text{right}}\Big) \tag{24}$$

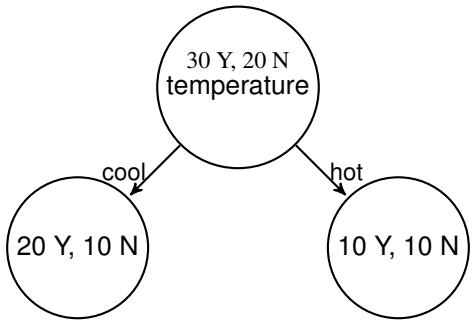
It is not difficult to formally express information gain: let $\mathcal{S}$ be all the data and $\Pr(\cdot)$ be the distribution of the data. Let $\Pi = \{\Pi_1, \dots \Pi_k\}$ be a partition of $\mathcal{S}$, where $a \in \{1, \dots, k\}$. $\pi_a \equiv \frac{|\Pi_a|}{|\mathcal{S}|}$ be the proportion of data in a^{th} partition.

$$\begin{aligned} \text{GAIN}(\mathcal{S}, \Pi) &= H(\Pr(\mathcal{S})) - \sum_a \pi_a H\big(\Pr(\mathcal{S}_a)\big) \\ &= H\big(\Pr(\mathcal{S})\big) - \mathbb{E}_{a\sim\pi}\big[H\big(\Pr(\mathcal{S}_a)\big)\big] \\ &= H\big(\Pr(\mathcal{S})\big) - H\big(\Pr(\mathcal{S}|\Pi)\big) \end{aligned} \tag{25}$$

3.5 Information Gain Ratio

One disadvantage of information gain is that the information gain will favor those nodes with a large number of branches. Consider an extreme case where there are $100 datapoints, and it has 100$ branches, and each node contains one data point (so it must have entropy 0). So the sum of entropies is 0, and it will definitely be selected! It creates an over-fitting problem.

Therefore we need to "normalize" each of the nodes somehow. What are we going to normalize it with? The answer once again lies within the Entropy. However, the entropy is no longer computed by the number of "YES" and "NO". It is infact computed by the number of data that goes into each of the branches. For example:



Because of the 50 total data is split into 30 and 20 (regardless of "YES" or "NO"), then we just use this total number to compute this so-called Intrinsic Information, or H_I:

$$\begin{aligned} H_I &= -\Big(\frac{30}{50}\log\Big(\frac{30}{50}\Big) + \frac{20}{50}\log\Big(\frac{20}{50}\Big)\Big) \\ &\approx 0.673 \end{aligned} \tag{26}$$

after that, we will need to compute the Infomration Gain Ratio GR as:

$$\text{GR} = \frac{\text{GAIN}}{H_I} \tag{27}$$

4 GINI index

Just like information gain, another metric is called the GINI index. GINI index can be used to search for splits that might lead to a "purer" distribution. The GINI index is defined as (again, we use two branch cases):

$$\text{GINI} = 1 - \Big(P^2_{\text{(YES)}} + P^2_{\text{(NO)}}\Big) \tag{28}$$

As in the case of information gain, we also combine the left and right branches of the GINI index, using:

$$\text{GINI} = \text{Pr}_{\text{left}}\text{GINI}_{\text{left}} + \text{Pr}_{\text{right}}\text{GINI}_{\text{right}} \tag{29}$$

The attribute providing smallest GINI is chosen to split the node.

4.1 Some examples of GINI index

It's also useful to look at GINI index calculation with Entropy together:

4.1.1 example one

Table 8: example of GINI and Entropy

class	count
YES	0
NO	6

1. GINI

$$\begin{aligned} \text{GINI} &= 1 - \left(P_{(\text{YES})}^2 + P_{(\text{NO})}^2\right) \\ &= 1 - \left[\left(\frac{0}{6}\right)^2 + \left(\frac{6}{6}\right)^2\right] \\ &= 0 \end{aligned} \tag{30}$$

2. Entropy

$$\begin{aligned} H &= -P_{(\text{YES})} \log_2(P_{(\text{YES})}) + -P_{(\text{NO})} \log_2(P_{(\text{NO})}) \\ &= -\left(\frac{0}{6}\right) \log_2\left(\frac{0}{6}\right) + -\left(\frac{6}{6}\right) \log_2\left(\frac{6}{6}\right) \\ &= 0 \end{aligned} \tag{31}$$

they both equal zero!

4.1.2 example two

Table 9: example of GINI and Entropy

class	count
YES	1
NO	5

1. GINI

$$\begin{aligned} \text{GINI} &= 1 - \left(P_{(\text{YES})}^2 + P_{(\text{NO})}^2\right) \\ &= 1 - \left[\left(\frac{1}{6}\right)^2 + \left(\frac{5}{6}\right)^2\right] \\ &= 0.278 \end{aligned} \tag{32}$$

2. Entropy

$$\begin{aligned} H &= -P_{(\text{YES})} \log_2(P_{(\text{YES})}) + -P_{(\text{NO})} \log_2(P_{(\text{NO})}) \\ &= -\left(\frac{1}{6}\right) \log_2\left(\frac{1}{6}\right) + -\left(\frac{5}{6}\right) \log_2\left(\frac{5}{6}\right) \\ &= 0.650 \end{aligned} \tag{33}$$

4.1.3 example three

Table 10: example of GINI and Entropy

class	count
YES	3
NO	3

1. GINI

$$\begin{aligned} \text{GINI} &= 1 - \left(P_{(\text{YES})}^2 + P_{(\text{NO})}^2\right) \\ &= 1 - \left[\left(\frac{3}{6}\right)^2 + \left(\frac{3}{6}\right)^2\right] \\ &= 0.5 \end{aligned} \tag{34}$$

2. Entropy

$$\begin{aligned} H &= -P_{(\text{YES})}\log_2(P_{(\text{YES})}) + -P_{(\text{NO})}\log_2(P_{(\text{NO})}) \\ &= -\left(\frac{3}{6}\right)\log_2\left(\frac{3}{6}\right) + -\left(\frac{3}{6}\right)\log_2\left(\frac{3}{6}\right) \\ &= 1 \end{aligned} \tag{35}$$

5 χ^2 testing

As I told you, most of the topics in this course are related. Here is one more proof: we can also use the χ^2 test to determine the criteria for splitting. Remember, the χ^2 test uses the X^2 statistic to determine the difference between n_i and E_i:

$$X^2 = \sum_{i=1}^{k} \frac{(n_i - E_i)^2}{E_i} \tag{36}$$

and also remember that, for some underlying assumption of p_i, we have:

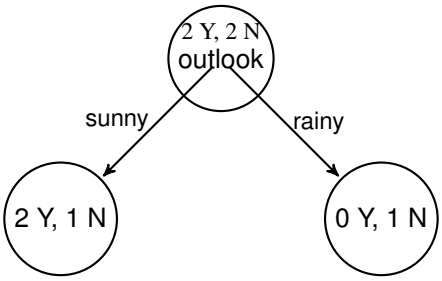
$$E_i = Np_i \tag{37}$$

So in here, let's use p_i to be the uniform distribtuion, i.e., if you have just "YES" or "NO", then it would be $\left(p_1 = \frac{1}{2}, p_2 = \frac{1}{2}\right)$. Therefore it makes E_i becomes the average of $\{n_i\}$. In the "YES" and "NO" case:

$$\left(E_{\text{YES}}, E_{\text{NO}}\right) = \frac{n_{\text{YES}} + n_{\text{NO}}}{2} \tag{38}$$

Just as before, the higher value of X^2, the larger the difference between $\left(E_{\text{YES}}, E_{\text{NO}}\right)$ and $\left(n_{\text{YES}}, n_{\text{NO}}\right)$, the higher the X^2 statistics. So natually, we would want to pick the split branch that gives the highest sum of X^2

5.1 example split on outlook



5.1.1 "sunny" branch

$$\begin{aligned} E_{\text{YES}} &= \frac{2+1}{2} = 1.5 \\ E_{\text{NO}} &= \frac{2+1}{2} = 1.5 \end{aligned} \tag{39}$$

$$\begin{aligned} \implies X^2_{\text{sunny}} &= \frac{(2-1.5)^2}{1.5} + \frac{(1-1.5)^2}{1.5} \\ &\approx 0.33 \end{aligned} \tag{40}$$

5.1.2 "rainy" branch

$$\begin{aligned} E_{\text{YES}} &= \frac{0+1}{2} = 0.5 \\ E_{\text{NO}} &= \frac{0+1}{2} = 0.5 \end{aligned} \tag{41}$$

$$\begin{aligned} \implies X^2_{\text{rainy}} &= \frac{(0-0.5)^2}{0.5} + \frac{(1-0.5)^2}{0.5} \\ &= 1.0 \end{aligned} \tag{42}$$

5.1.3 X^2 statistics for outlook

$$\begin{aligned} X^2_{\text{outlook}} &= X^2_{\text{sunny}} + X^2_{\text{rainy}} \\ &\approx 1.33 \end{aligned} \tag{43}$$

Then you repeat the process for the remaining nodes, and to pick the one has the highest X^2-statistics value.

6 Algorithm to build the decision tree

in the next lecture, we will discuss algorithms to construct the decision tree

1. ID3 (Iterative Dichotomiser), C4.5, C5.0 (Ross Quinlan 1986,1993)
2. CART (Classification and Regression Trees) (Leo Briemen et al 1984)
3. CHAID (J. A. Hartigan, 1975)

6.1 Decision tree algroithm in a glance

1. Choose **best** attribute to split the remaining cases and make that attribute a decision node.
2. Repeat this process recursively for each child.
3. Stop when the stopping criteria meets. for example, All (or most) the cases have the same target attribute value.

At each node, available attributes are evaluated on the basis of separating the classes of the training examples. An impurity measure is used. The typical impurity measures are the ones you have already studied so far:

1. Information gain (ID3/C4.5)
2. Information gain ratio (C4.5)
3. Gini Index (CART)
4. χ^2 test (CHAID)

7 χ^2 testing explanation

7.1 Contingency Table

7.1.1 joint $\rightarrow$ marginal

imagine we have a contingency table with observation counts a, b, c, d and we also know: $a+b+c+d = N$

Table 11: contingency table

	$X=0$	$X=1$	**sum**
$Y=0$	a	b	$a+b$
$Y=1$	c	d	$c+d$
sum	$a+c$	$b+d$	N

we may consider the above as "un-normalized" joint probability density (but X and Y are **not** independent). After we normalize them, we obtained:

Table 12: joint density

	$X=0$	$X=1$	**sum**
$Y=0$	$\frac{a}{N}$	$\frac{b}{N}$	$\frac{a+b}{N}$
$Y=1$	$\frac{c}{N}$	$\frac{d}{N}$	$\frac{c+d}{N}$
sum	$\frac{a+c}{N}$	$\frac{b+d}{N}$	1

Note that column **sum** and row **sum** contain marginal probabilities. Joint $\rightarrow$ Marginal density has many-to-one relationship. The reverse Marginal $\rightarrow$ Joint has one-to-many relationship!

7.1.2 marginal $\rightarrow$ joint

It is easy to check, by keeping the same marginal density from Table **??**, if we assume $\Pr(X)$ and $\Pr(Y)$ to be **independent**, then we have derived the joint densities at Table **??**:

Table 13: independent variables

	$X=0$	$X=1$	**sum**
$Y=0$	$\frac{(a+b)(a+c)}{N^2}$	$\frac{(a+b)(b+d)}{N^2}$	$\frac{a+b}{N}$
$Y=1$	$\frac{(c+d)(a+c)}{N^2}$	$\frac{(c+d)(b+d)}{N^2}$	$\frac{c+d}{N}$
sum	$\frac{a+c}{N}$	$\frac{b+d}{N}$	1

where column sum represents $\Pr(Y)$ and row sum represents $\Pr(X)$ which is the same as Table **??**:

$$\begin{cases} P(X=0) & = \frac{a+b}{N} \\ P(X=1) & = \frac{c+d}{N} \end{cases} \tag{44}$$

and

$$\begin{cases} P(Y=0) & = \frac{a+c}{N} \\ P(Y=1) & = \frac{b+d}{N} \end{cases} \tag{45}$$

7.1.3 Putting the two together

Putting the **independent** and **dependent** tabular entries together, where they share the same marginal distribution, but differ in joint densities. Therefore, our task is to check to see if there are "significant" difference between the two:

	$X=0$	$X=1$	**sum**
$Y=0$	$\frac{a}{N}$, $\frac{(a+b)(a+c)}{N^2}$	$\frac{b}{N}$, $\frac{(a+b)(b+d)}{N^2}$	$\frac{a+b}{N}$
$Y=1$	$\frac{c}{N}$, $\frac{(c+d)(a+c)}{N^2}$	$\frac{d}{N}$, $\frac{(c+d)(b+d)}{N^2}$	$\frac{c+d}{N}$
sum	$\frac{a+c}{N}$	$\frac{b+d}{N}$	1

7.2 what is χ^2 testing?

Forget about checking for independence of discrete random variables for now and let's look at vanilla χ^2 testing.

If we have N random observations, that are classified into k classes. Each with the number of (random) observations fall into class i to be: n_i, where $i = 1, 2, \ldots, k$

null hypothesis assumes that true probability that an observation falls into i^{th} class is p_i, and of course:

$$\sum_{i=1}^{k} p_i = 1 \tag{46}$$

the expected numbers (computed without seen any observation n_i are):

$$E_i = Np_i \qquad \forall i \quad \text{(expectation of multinomial distribution)} \tag{47}$$

$$\begin{aligned} \sum_{i=1}^{k} E_i &= N \sum_{i=1}^{k} p_i \quad \text{theoretical} \\ &= \sum_{i=1}^{k} n_i \quad \text{observed} \\ &= N \end{aligned} \tag{48}$$

Under circumstance of **null hypothesis** being correct, as $N \to \infty$, limiting distribution of X^2 is χ^2 distribution, where X^2 is defined as follows:

$$X^2 = \sum_{i=1}^{k} \frac{(n_i - E_i)^2}{E_i} \tag{49}$$

X^2 follows χ^2 distribution with $k-1$ degrees of freedom. $k-1$ DOF because $\{p_i\}$ has $k-1$ DOF

7.3 Apply to two-way contingency table

Apply vanilla χ^2 testing into our problem of checking for independence of random discrete variable:

7.3.1 what is the degree of freedom

since we have:

$$\begin{cases} \Pr(X) & \equiv \{\Pr(X=1), \ldots, \Pr(X=k_X)\} \\ \Pr(Y) & \equiv \{\Pr(Y=1), \ldots, \Pr(Y=k_Y)\} \end{cases} \tag{50}$$

therefore the degree of freedom must be:

$$(k_X - 1) \times (k_Y - 1) \tag{51}$$

which is the (number of columns -1) $\times$ (number of rows -1)

7.3.2 How does it relate to vanilla χ^2 testing

instead of checking the difference between $\{p_1, \ldots, p_k\}$ with $\{\frac{n_1}{N}, \ldots, \frac{n_k}{N}\}$ as in the case of vanilla χ^2 testing, it is tailored to this setting by showing the correspondence:

$$\begin{aligned}
\{n_i\} \longrightarrow \quad & N \times \Big\{ \frac{a}{N}, \frac{b}{N}, \frac{c}{N}, \frac{d}{N} \Big\} \\
& = \{a, b, c, d\} \\
\{p_i\} \longrightarrow \quad & \Big\{ \frac{(a+b)(a+c)}{N^2}, \frac{(a+b)(b+d)}{N^2}, \frac{(c+d)(a+c)}{N^2}, \frac{(c+d)(b+d)}{N^2} \Big\} \\
\{E_i\} \longrightarrow \quad & N \times \Big\{ \frac{(a+b)(a+c)}{N^2}, \frac{(a+b)(b+d)}{N^2}, \frac{(c+d)(a+c)}{N^2}, \frac{(c+d)(b+d)}{N^2} \Big\} \\
& = \Big\{ \frac{(a+b)(a+c)}{N}, \frac{(a+b)(b+d)}{N}, \frac{(c+d)(a+c)}{N}, \frac{(c+d)(b+d)}{N} \Big\}
\end{aligned} \tag{52}$$

and we can compute X^2 in the exact same fashion:

$$X^2 = \sum_{i=1}^{k} \frac{(n_i - E_i)^2}{E_i}$$

where

$$\begin{aligned}
& n_i \in \{a, b, c, d\} \\
& E_i \in \Big\{ \frac{(a+b)(a+c)}{N}, \frac{(a+b)(b+d)}{N}, \frac{(c+d)(a+c)}{N}, \frac{(c+d)(b+d)}{N} \Big\}
\end{aligned} \tag{53}$$

of course, it can trivially extended to situations of having more than 2 rows and 2 columns.

7.3.3 p-value

p-value for testing shows us:

$$\Pr(\chi^2 \geq X^2) \tag{54}$$

Obviously, if p-value is small, say, less than 5%, then we can shall reject null-hypothesis, i.e., X and Y are not independent.

7.4 mutual information

Think about using KL to achieve the same purpose.

$$
\begin{aligned}
\mathbf{I}(X;Y) &= \mathrm{KL}\big(\Pr(x,y) \| \Pr(x)\Pr(y)\big) \\
&= \sum_{x\in\mathcal{X},y\in\mathcal{Y}} \Pr(x,y)\log\frac{\Pr(x,y)}{\Pr(x)\Pr(y)} \\
&= \sum_{x\in\mathcal{X},y\in\mathcal{Y}} \Pr(x,y)\log\frac{\Pr(x,y)}{\Pr(x)} - \sum_{x\in\mathcal{X},y\in\mathcal{Y}} \Pr(x,y)\log\Pr(y) \\
&= \sum_{x\in\mathcal{X},y\in\mathcal{Y}} \Pr(x)\Pr(y|x)\log\Pr(y|x) - \sum_{x\in\mathcal{X},y\in\mathcal{Y}} \Pr(x,y)\log\Pr(y) \\
&= \sum_{x\in\mathcal{X}} \Pr(x)\Big(\sum_{y\in\mathcal{Y}} \Pr(y|x)\log\Pr(y|x)\Big) - \sum_{y\in\mathcal{Y}}\Big(\sum_{x\in\mathcal{X}}\Pr(x,y)\Big)\log\Pr(y) \\
&= -\sum_{x\in\mathcal{X}} \Pr(x)\mathrm{H}(Y \mid X=x) - \sum_{y\in\mathcal{Y}}\Pr(y)\log\Pr(y) \\
&= -\mathrm{H}(Y \mid X) + \mathrm{H}(Y) \\
&= \mathrm{H}(Y) - \mathrm{H}(Y \mid X)
\end{aligned}
\tag{55}
$$

so the result shows that mutual information is the difference of two entropy, where $\mathrm{H}(Y \mid X)$ is Entropy using conditional distribution $\Pr(Y \mid X)$