Example from lecture - when to play tennis?

Choosing the root of the tree

Compute the score for each of the features, choosing the majority label:

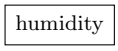
outlook		
sunny:	**yes**	**4**
	no	3
rain:	**yes**	**4**
	no	2
score:		$\frac{8}{13}$

temperature		
warm:	**yes**	**5**
	no	4
cool:	**yes**	**3**
	no	1
score:		$\frac{8}{13}$

humidity		
high:	yes	2
	no	**4**
normal:	**yes**	**6**
	no	1
score:		$\mathbf{\frac{10}{13}}$

wind		
strong:	**yes**	**3**
	no	3
weak:	**yes**	**5**
	no	2
score:		$\frac{8}{13}$

Humidity feature gives the highest score, so this will be the root of the tree.



Next we divide the data set into two based on the value of the **humidity** feature. In one set are the items with humidity **high** and in the other the items with humidity **normal**. This is the **divide** part of the divide-and-conquer strategy.

Growing the tree

Next we apply the same procedure to the both sets separately, now excluding the already used humidity feature. This is the **conquer** step of the divide-and-conquer strategy.

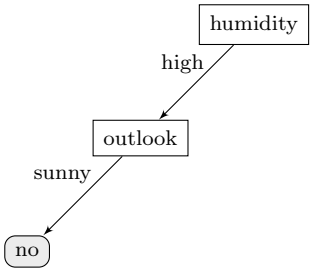
Let's first do the set with humidity **high**:

humidity = high **outlook**		
sunny:	yes	0
	no	**3**
rain:	**yes**	**2**
	no	1
score:		$\mathbf{\frac{5}{6}}$

humidity = high **temperature**		
warm:	yes	2
	no	**4**
cool:	**yes**	**0**
	no	0
score:		$\frac{4}{6}$

humidity = high **wind**		
strong:	yes	1
	no	**2**
weak:	yes	1
	no	**2**
score:		$\frac{4}{6}$

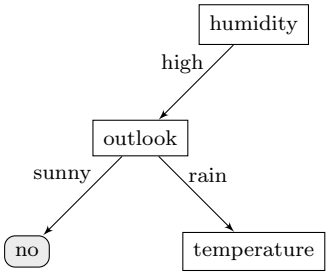
Highest score is given by the **outlook** feature, so we divide the set again, now based on the weather outlook. As the value **sunny** always leads to the answer **no** there is no need to grow this branch any further and we end it with the leaf node **no**. We proceed with growing the branch where outlook is **rain**.



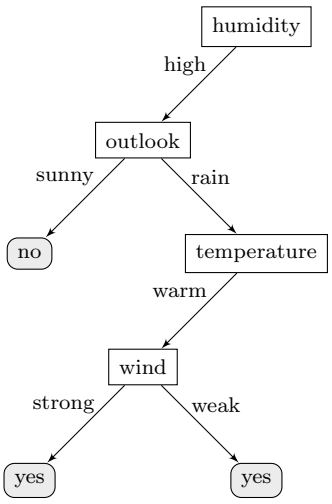
humidity = high **outlook = rain** **temperature**		
warm:	**yes**	**2**
	no	1
cool:	**yes**	**0**
	no	0
score:		$\mathbf{\frac{2}{3}}$

humidity = high **outlook = rain** **wind**		
strong:	**yes**	**1**
	no	1
weak:	**yes**	**1**
	no	0
score:		$\frac{2}{3}$

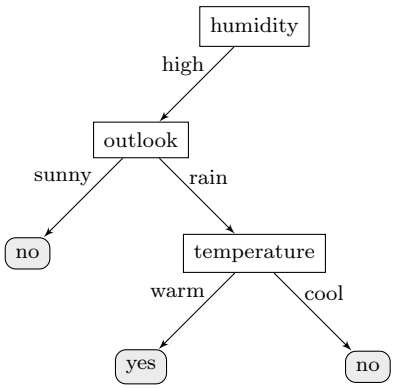
Here the scores are equal and we can break the tie arbitrarily. I'll choose the first feature **temperature** for now, although one could argue that the **wind** feature would actually be a better choice because with temperature both branches will contain ambiguity - value **warm** has mixed labels and value **cool** has no data at all. Wind feature has also mixed labels with the value **strong**, but is disambiguous with the value **weak** (although this is really a mini example and normally one data point is not enough to draw any conclusions). There are other cost or gain functions that could distinguish between the differences of these two options.



We continue with the subset where temperature is **warm**. Looking at the last feature **wind** we can see that there is noise in the data when wind is strong - there are two exactly similar data items (rain warm high strong), one having label **yes** and other **no**. Here we also break the tie by arbitrarily choosing the label **yes**. With the wind feature value **weak** the label in the leaf will be **yes**.



As the both leaves under the branch **wind** will have label **yes** then we can actually remove this branch altogether and replace it with the leaf labelled **yes**. The branch with temperature feature value **cool** has no data at all and thus we must decide randomly which label will be assigned. Let's assign now arbitrarily **no**. If we would have chosen **yes** then both branches under the feature **temperature** would have had the same value in the leaf and we could have also replaced this branch with a single leaf **yes**.



Growing the second half of the tree

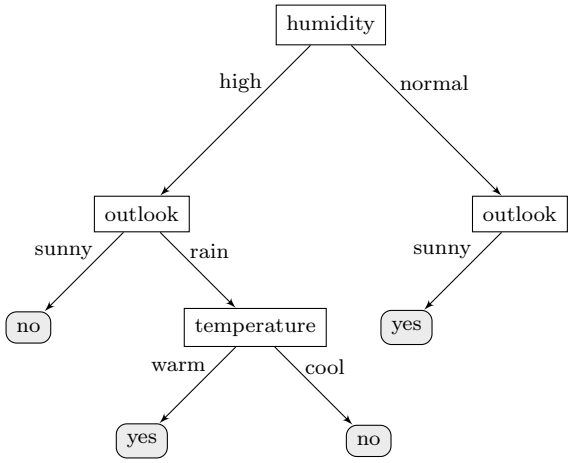
Now we are ready with the first half of the tree and must follow the same procedure with the other half - with the data items having humidity feature value **normal**.

humidity = normal outlook		
sunny:	**yes**	**4**
	no	0
rain:	**yes**	**2**
	no	1
score:		$\mathbf{\frac{6}{7}}$

humidity = normal temperature		
warm:	**yes**	**3**
	no	0
cool:	**yes**	**3**
	no	1
score:		$\frac{6}{7}$

humidity = normal wind		
strong:	**yes**	**2**
	no	1
weak:	**yes**	**4**
	no	0
score:		$\frac{6}{7}$

The scores are equal and we make an arbitrary choice choosing the feature **outlook**. With the value **sunny** the label is always **yes** and thus we end this branch with a leaf and continue with the subset of items having the feature value **rain**.



humidity = normal **outlook = rain** **temperature**		
warm:	**yes**	**1**
	no	0
cool:	**yes**	**1**
	no	1
score:		$\frac{2}{3}$

humidity = normal **outlook = rain** **wind**		
strong:	yes	0
	no	**1**
weak:	**yes**	**2**
	no	0
score:		$\frac{\mathbf{3}}{\mathbf{3}}$

The **wind** feature has accuracy 100% on this subset and thus it will be chosen and the branch can be finished with the corresponding leaves.