

HOMEWORK PROBLEMS EXPLAINED

Shingchern D. You

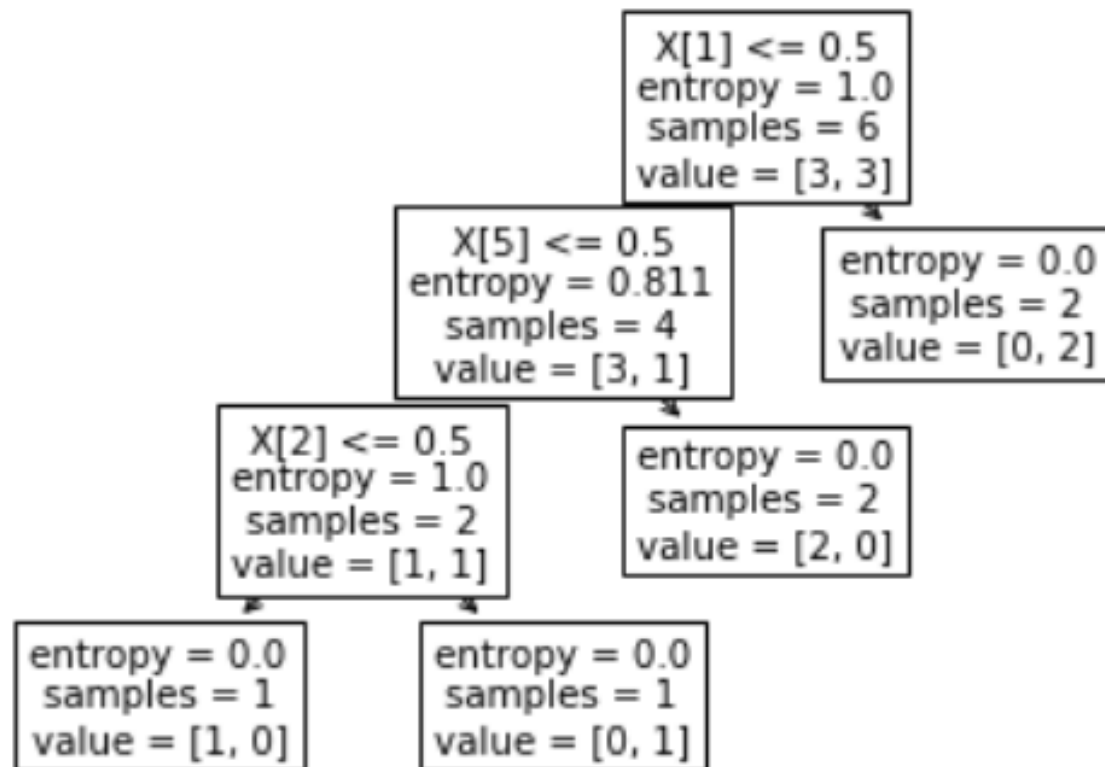
How to read decision tree

□ Here is the table

Outlook	Temperature	Humidity	Windy	Decision
Sunny	Hi	Hi	No	No play
Sunny	Hi	Hi	Yes	No play
Overcast	Hi	Lo	No	Play
Overcast	Lo	Lo	Yes	Play
Rain	Lo	Hi	No	Play
Rain	Lo	Hi	Yes	No play

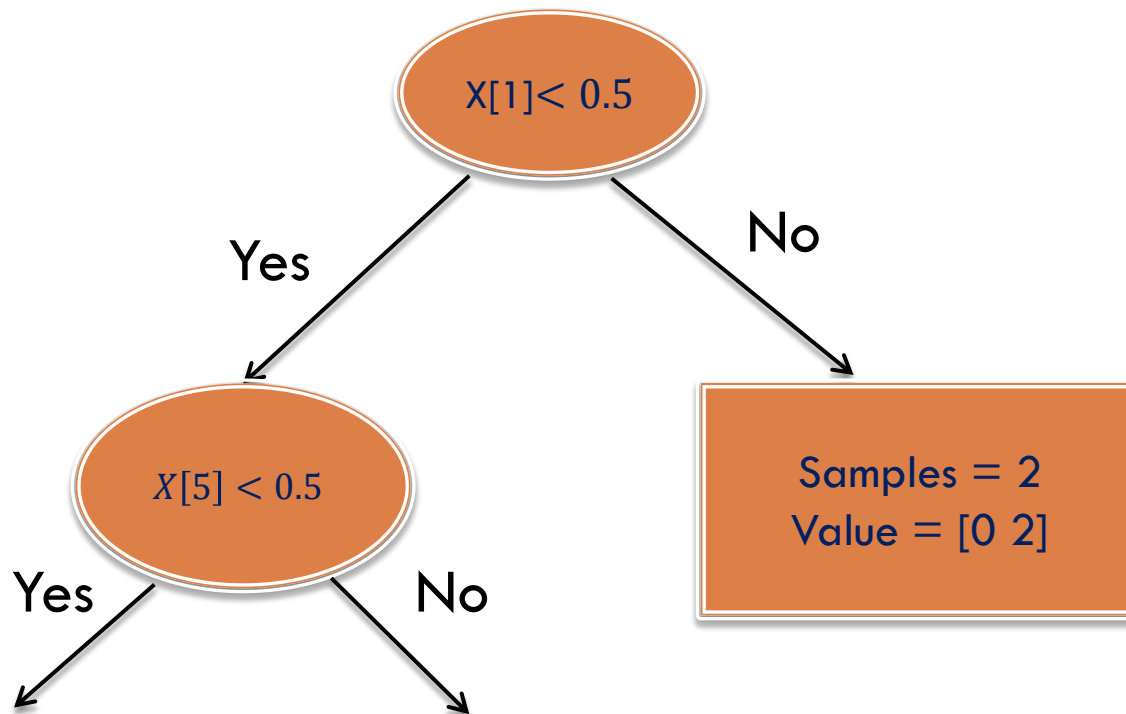
How to read decision tree

□ The generated tree



General format

- Left edge is **yes**, right edge is **no**



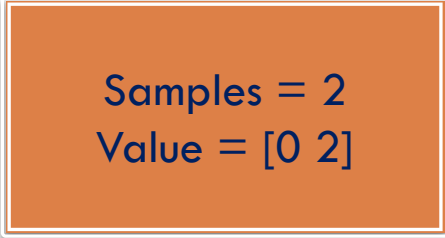
Remember one-hot encoding

- $X[1] = 1$ means overcast
- Right edge of $(X[1] < 0.5)$ means overcast

Sunny	Overcast	Rainy	Temp	Humi	Windy
X[0]	X[1]	X[2]	X[3]	X[4]	X[5]
1	0	0	1	1	0
			More		

Leaf node

- How about leaf node



Samples = 2
Value = [0 2]

- Two samples in this leaf
- Number of samples in 1st category (no) is 0
- Number of samples in 2nd category (yes) is 2
- Final answer is **Yes** for this leaf
 - ▣ Recall we encode answer of **no** as 0, and **yes** as 1

Is this answer correct

- Check out the original table (it is correct)

Outlook	Temperature	Humidity	Windy	Decision
Sunny	Hi	Hi	No	No play
Sunny	Hi	Hi	Yes	No play
Overcast	Hi	Lo	No	Play
Overcast	Lo	Lo	Yes	Play
Rain	Lo	Hi	No	Play
Rain	Lo	Hi	Yes	No play

Both categories nonzero

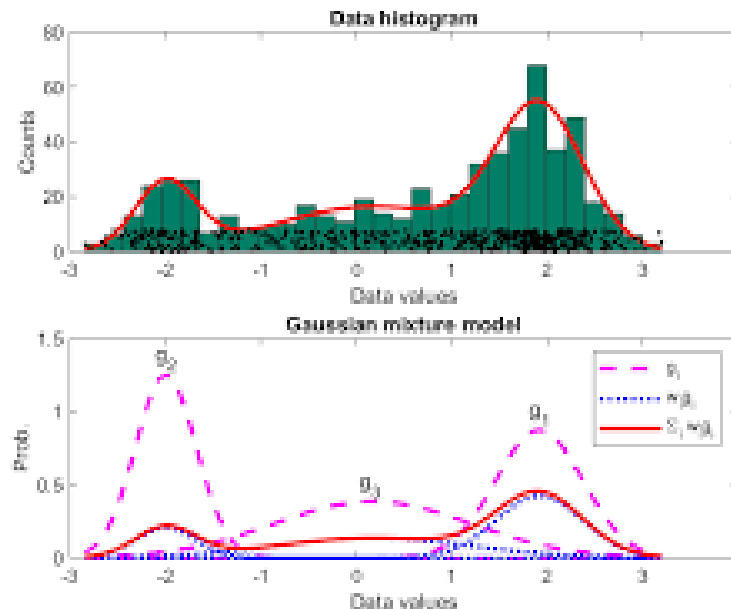
- If we have the following leaf

Samples = 5
Value = [2 3]

- Which category to use
 - ▣ 2nd category is chosen (why?)
 - ▣ Majority voting
- This situation happens if we limit the growth of the tree (to avoid overfitting)

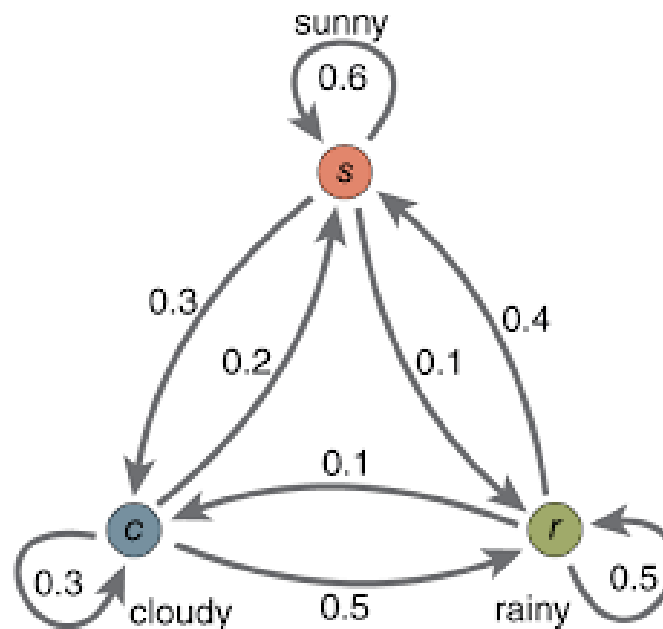
GMM

- `gmm0 = GaussianMixture(n_components=3)`
- `n_components` means number of Gaussian mixtures



HMM

- `testHMM1 =`
`hmm.MultinomialHMM(n_components=3,n_iter=100)`
- `n_components` means the number of hidden states



HMM



- `n_iter` means maximum number of iterations to run the EM algorithm
 - ▣ Only a few iterations is needed for small problems

Feature selection

- We run the feature selection program and found 3 features can reach accuracy of 95%
 - ▣ Similar performance as using 9 (all features)
- If the chosen features are [0,1,5], what does it mean

Feature selection

- From original document
 - 1. Sample code number: id number
 - 2. **Clump Thickness: 1 - 10**
 - 3. **Uniformity of Cell Size: 1 - 10**
 - 4. Uniformity of Cell Shape: 1 - 10
 - 5. Marginal Adhesion: 1 - 10
 - 6. Single Epithelial Cell Size: 1 - 10
 - 7. **Bare Nuclei: 1 - 10**
 - 8. Bland Chromatin: 1 - 10
 - 9. Normal Nucleoli: 1 - 10
 - 10. Mitoses: 1 - 10

Feature selection

- Later on, if a sample is needed we just need to examine only three parameters
 - ▣ **Clump Thickness: 1 - 10**
 - ▣ **Uniformity of Cell Size: 1 – 10**
 - ▣ **Bare Nuclei: 1 - 10**
- Save time & energy
- Although I have no knowledge in breast cancer, I know the above parameters are more important than others