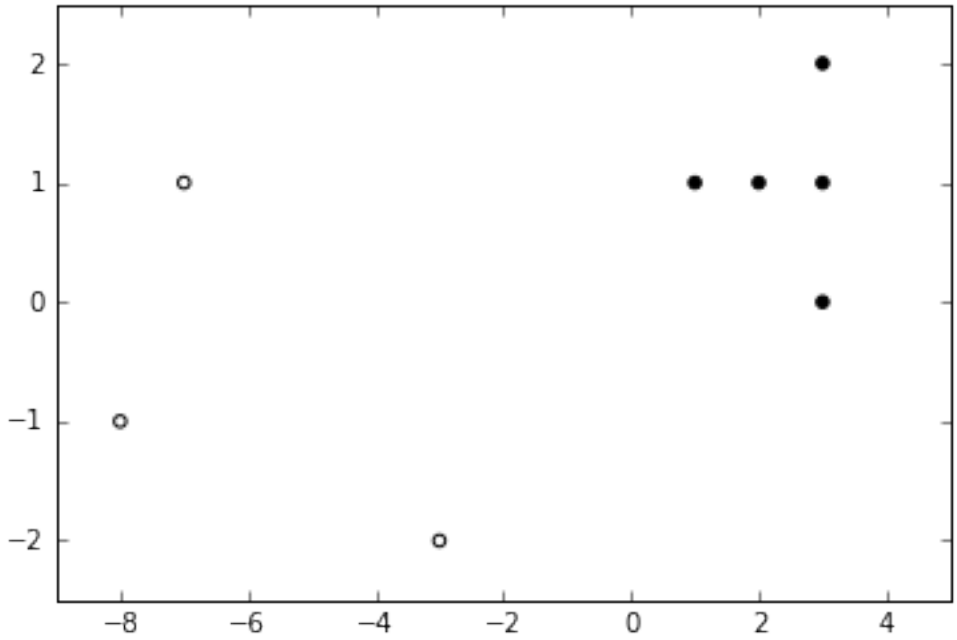
Classification using Decision Trees

Decision trees are another way of classifying! The cool thing about decision trees is that you can use it for many types of problems, including ones with continuous variables (like our points (x, y), or height, or time, or price) and problems with categorical variables (categories like yes/no, buy/hold/sell, fitted/loose, etc). Decision trees usually involved drawing out a tree, and at each branch you make a decision. When you play the game "Twenty Questions" you are using a decision tree.

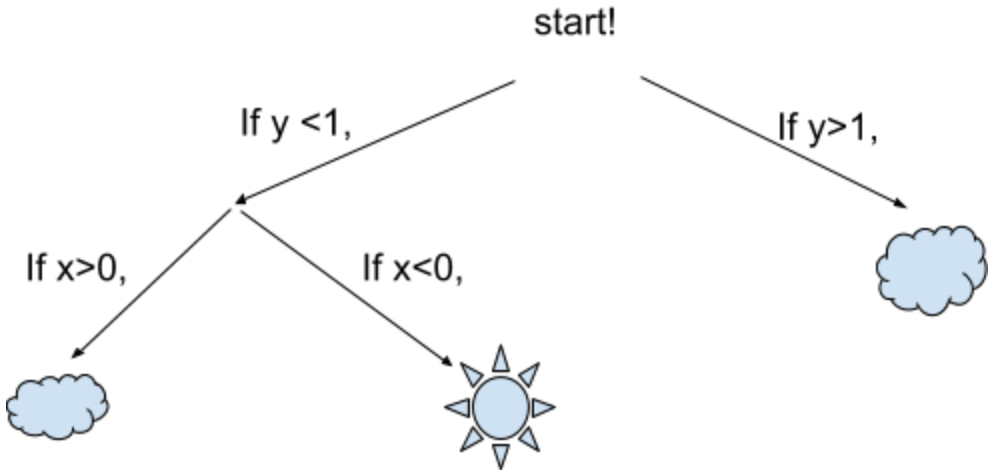
Big idea: Pick the right questions to separate the groups

Fake data first: Here is some fake data, designed to be nice. Can you ask a question about one variable (just the x coordinate, or just the y coordinate) that would allow you to decide if you should have a black dot or a white dot?



Talk with your neighbor – do you agree?

Go backward this time: I'll give you a decision tree, and you draw a set of points that agrees with the decision tree rules.



Draw the set of points (x, y) that should be sunny and the set of points that should be cloudy according to this decision tree:

The sun and cloud symbols above are fun but also point to a set of examples people have used in textbooks about machine learning and data mining (this also means learning from data). Here's a version of this classic example:

> You are trying to figure out if your little sister's soccer game will happen today. Last year, you know the following happened:

Temperature	Rain	Humidity	Play
90	no	high	no
73	no	low	yes
81	yes	high	no
67	no	high	yes
72	yes	high	yes
77	no	low	yes
96	no	low	no
81	yes	high	no
58	yes	high	yes
72	no	medium	yes

1. On Saturday, the forecast is for 90 degrees, high humidity, and no rain. Using your intuition and the data above, do you think the game will happen?

2. On Sunday, the forecast is for 70 degrees, low humidity, and rain. Using your intuition and the data above, do you think the game will happen?

3. Based on the data above, make a decision tree that tells you if a game will happen, based on the weather.

> Compare your decision tree with your neighbor. How can you tell if one decision tree is better or worse than another?

4. Looking at the soccer game data, what do you think is the most important factor in determining whether or not a game occurs?

5. How does your answer to 4 affect how you would construct your decision tree?

6. Can you think of a way to use the data to show mathematically that your answer to 4 is correct?