Midterm 1 Practice Problems *(find and work with a partner)*

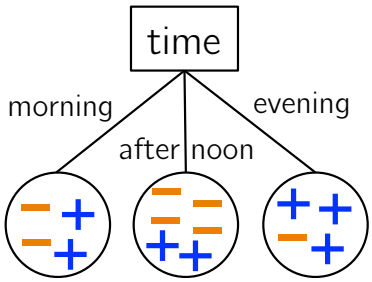
1. Trying to predict whether a student will be most interested in courses related to History, Physics, or Art is an example of what type of learning problem?

2. Trying to predict a person's height based on their genome is an example of what type of learning problem?

3. Match the decision tree component on the left with its corresponding data component on the right.

• internal nodes	class labels
• branches	feature names
• leaves	feature values

4. Say I am trying to predict if a student will like a course (`+`) or dislike it (`-`). One of the features is the time of day the course is offered. If I just choose this one feature and build a decision tree, here is how the training examples cluster at the leaves:



   (a) Using a threshold of 0.5, how would you classify a new example with value `evening` for the feature `time`?

   (b) Using a threshold of 0.5, how many *training* examples are incorrectly predicted? Divide by the total number of examples to find the error.

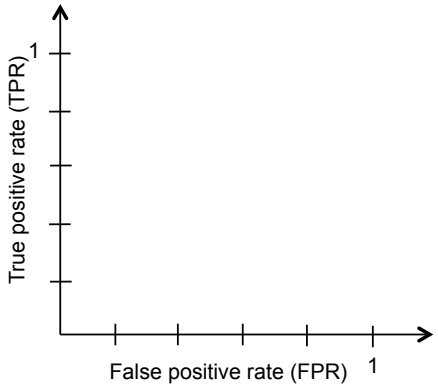
5. Student X has created a test for diabetes. The outcomes of the test are *negative* (person doesn't have diabetes) and *positive* (person does have diabetes).

   (a) They give you their model and you use it on your test data to obtain the following confusion matrix. Which cell below do you find most concerning? Briefly explain your answer.

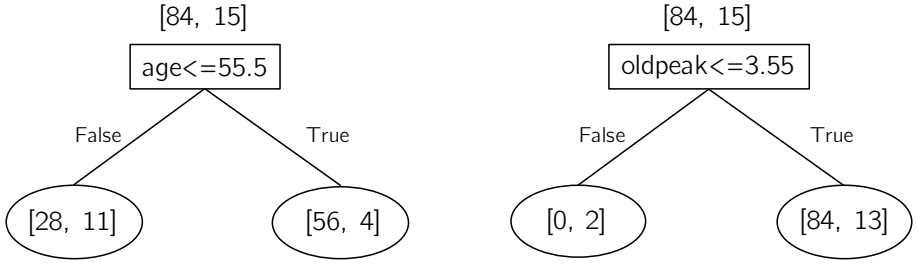
		Predicted class: Negative	Predicted class: Positive
True class	Negative	42	8
	Positive	8	12

   (b) From the confusion matrix above, calculate the FPR (False Positive Rate) and the TPR (True Positive Rate).

   (c) Use your FPR and TPR to plot a point on the ROC curve axes below (roughly). Also include the 2 other points that are always on a ROC curve. Connect these 3 points to sketch the curve.



6. Consider the two feature choices below (for a heart disease dataset), and their associated splits. Counts of label -1 vs. +1 are shown in brackets (negative first). Given a threshold of 0.5, what is the *training error* for each model? Draw out the confusion matrix for each model (for the training data).



7. *Linear Regression.* Say we have $p = 1$ and two training examples: $(x_1, y_1) = (3, 0)$ and $(x_2, y_2) = (7, 1)$, and we would like to fit a linear model to this dataset.

   (a) Draw these two examples on a coordinate system and sketch the linear function that would fit them. What are the optimal weights? ($\hat{w}_0$ and $\hat{w}_1$)

   (b) Say in our SGD method, we choose to analyze (x_2, y_2) first. Before starting SGD, we set $w_0 = 0$ and $w_1 = 0$. After analyzing (x_2, y_2), what are w_0 and w_1? Choose $\alpha = 0.1$. Plot this updated line on your graph above.

   (c) Next we consider (x_1, y_1). What are w_0 and w_1 be after this second data point? Plot this line on your graph above. At this point we have finished *one* iteration of SGD.

8. Given the movie data below ("Liked" is the response variable), create a decision tree for each feature. Which feature minimizes the training error? (using a threshold of 0.5)

Movie	Type	Length	Director	Famous actors	Liked?
m1	Comedy	Short	Adamson	No	Yes
m2	Animated	Short	Lasseter	No	No
m3	Drama	Medium	Adamson	No	Yes
m4	Animated	Long	Lasseter	Yes	No
m5	Comedy	Long	Lasseter	Yes	No
m6	Drama	Medium	Singer	Yes	Yes
m7	Animated	Short	Singer	No	Yes
m8	Comedy	Long	Adamson	Yes	Yes
m9	Drama	Medium	Lasseter	No	Yes