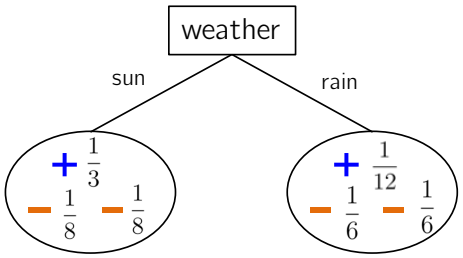
Midterm 2 Practice Problems *(find and work with a partner)*

1. *AdaBoost with Decision Stumps.* Say I am at iteration t of AdaBoost with $n = 6$. I train a classifier with the current weights (shown for each example below) and this is the resulting decision tree:



   (a) If I use a threshold of 0.5, what are the labels of each leaf? In other words, if a new example had *weather="sun"*, would I classify it as $+1$ or -1? And if the example had *weather="rain"*?

   (b) What is the weighted training error ϵ_t for this classifier?

   (c) We compute the score of this model as $\alpha_t = \frac{1}{2} \ln\left(\frac{1-\epsilon_t}{\epsilon_t}\right)$. What is the intuition behind using this scoring function?

2. *Ensembles reducing error.* Say I have $T = 5$ classifiers and each one has an $r = \frac{1}{3}$ chance of being incorrect. If I run all of them on a new example, the number of votes for the wrong class could be $R = 0, 1, 2, 3, 4$, or 5. Of these options, which would result in an incorrect classification overall? Compute $P(R = k)$ for each of these options, then use this result to compute the overall probability of an incorrect classification.

3. With $n = 2$ training examples, how many unique datasets can I generate with bagging? What about $n = 3$?

4. Say you are given the three confusion matrices below, which correspond to three different thresholds (i.e. if the threshold is t, I declare the label positive if $p(y|x) \geq t$).

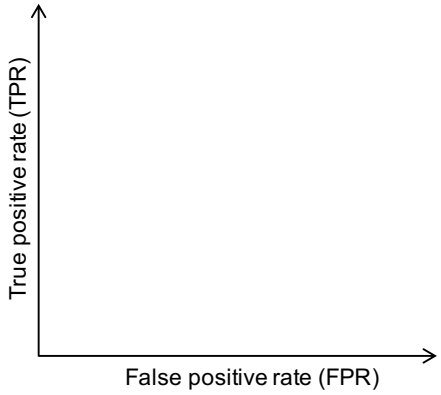
Predicted class

True class	N	P
N	3	7
P	2	18

	N	P
N	7	3
P	5	15

	N	P
N	9	1
P	10	10

(a) For each confusion matrix, add the corresponding point on the axes below to create a ROC curve (also include the two points that are always on a ROC curve).



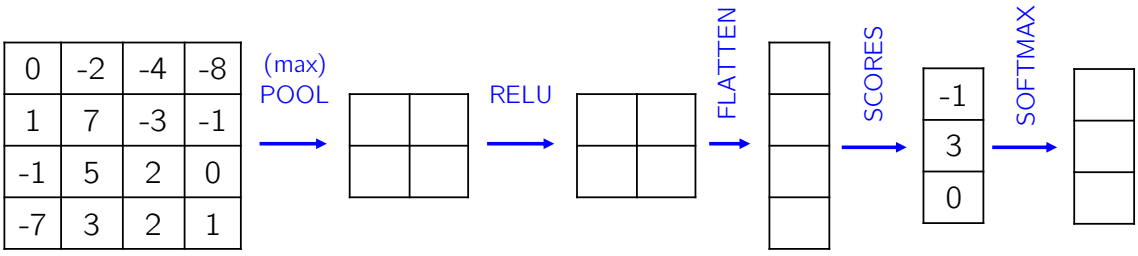
(b) Which confusion matrix corresponds to a low threshold t? Which one corresponds to a high threshold?

(c) For this example, what confusion matrix would correspond to the "best" point on an ideal ROC curve?

(d) Why do we apply the threshold during testing but not training?

(e) Think of a scenario (that we have not discussed in class/lab) where a low threshold would be desirable. Where a high threshold would be desirable?

5. Say I have the following output of a CONV layer on the left. Assume no bias terms throughout.



(a) If my original input was also 4×4 and I used one convolutional filter with size 3×3 (no bias), how much zero padding would I need? How many parameters would I need to learn just for this CONV layer?

(b) Fill in the steps POOL (2×2 with stride 2), RELU, FLATTEN.

(c) Say the scores are as given above for three potential labels $\hat{y} \in \{1, 2, 3\}$. Compute the SOFTMAX function to obtain a probability distribution over these three classes. What would you choose for the predicted label $\hat{y}$?

(d) If the true class was in fact $y = 2$, what is the cross-entropy loss?

(e) In the input had been a matrix of zeros, what would the scores be? What would the probability distribution (output of SOFTMAX) be?