HW #9

1. a. Linear classifiers are not as expressive as decision trees because they can only model linear decision boundaries, ~~w~~ unless you use transformations like basis expansion.

   b. SVMs are more expressive than linear classifiers, but not as expressive as decision trees. They can create quadratic / nonlinear boundaries, but are still limited to quadratic functions.

   c. NN classifiers are similarly expressive because w/ a large training set and small neighborhood sizes, they can also represent any decision boundary. Like decision trees, however, they may not generalize well.

   d. They are not as expressive because they assume each class follows a Gaussian distribution, so this assumption may be incorrect, and they wouldn't be able to create decision boundaries as expressive.

2. $d(n-1)$ possibilities

3. $G = 2(0.2)(0.8) = \boxed{0.32}$

4. a. Given $W = \sum_i \lambda_i$, $W_j = \sum_{i \mid y^{(i)} = j} \lambda_i$, using Gini index and $p_j = \frac{W_j}{W}$;

$$G = 1 - \sum_{j=1}^{k} p_j^2 = 1 - \sum_{j=1}^{k} \left(\frac{W_j}{W}\right)^2$$

$$= \boxed{1 - \frac{\sum_{j=1}^{k} W_j^2}{W^2}}$$

$\rightarrow$ p_j now considers weights for each pt, sum of pts × weights of one class over pts × weights of every pt

b. $\pi_j = \frac{\sum_{i \mid y^{(i)} = j} \lambda_i}{\sum_i \lambda_i}$ $\rightarrow$ points w/ higher weights now contribute more

$$\mu_j = \frac{\sum_{i \mid y^{(i)} = j} \lambda_i x^{(i)}}{\sum_{i \mid y^{(i)} = j} \lambda_i}$$ $\rightarrow$ weighted average

$$\Sigma_j = \frac{\sum_{i \mid y^{(i)} = j} \lambda_i (x^{(i)} - \mu_j)(x^{(i)} - \mu_j)^T}{\sum_{i \mid y^{(i)} = j} \lambda_i}$$ $\rightarrow$ data pts w/ higher weights influence cov matrix more

c. $\min_{w,b} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \xi_i$ $\rightarrow$ normal optimization

$$\min_{w,b} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \lambda_i \xi_i$$

modified function to include weights λ_i, so now misclassification of higher weighted pts cause larger penalty, which affects the decision boundary.

5. a. False, boosting reduces training error, but does not necessarily cause zero test error, which depends on how well the final classifier generalizes to new data.

b. True, since each weak classifier has error rate of $\frac{1}{2} - \varepsilon$, it will keep reducing training error till it equals zero. The algorithm weighs misclassified points more heavily, so it focuses on them until all points are correctly classified.

c. True, the final classifier is a linear combination of weak classifiers from class H, so it will also belong to class H.

6. a. ~~True,~~ This ~~benef~~ is a benefit of random forests, where each decision tree is trained independently and thus can be trained in parallel.

b. Boosted decision trees, which trains trees sequentially, improving ~~perf~~ performance between trees and focusing on misclassified points from ~~of~~ previous trees, so they are more highly optimized.

c. Boosted decision trees, because each tree improves accuracy as it reduces error from previous trees.