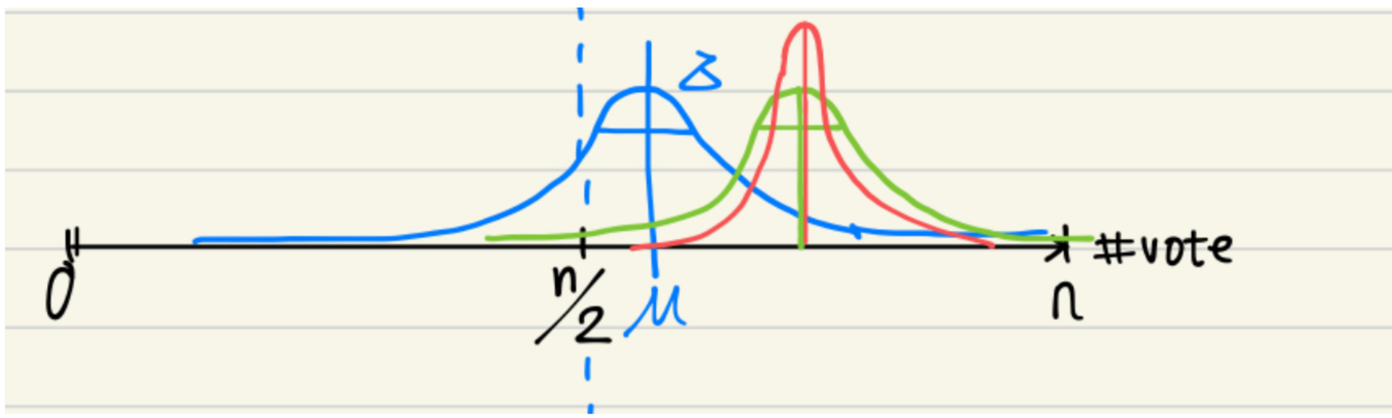
Random Forsts

- Why do random forests and bagging improve the stability of predictors.
- Fix an input feature vector x.
- Consider repeatedly running a random forest with n trees. Let #vote denote the number of votes for class 1. Clearly #vote is between 0 and n. Each time you run the random forest algorithm you get a different value for #vote for the label of the example x. Suppose you run the random forest algorithm many times, this would give you a distribution over #votes, the curves below represent the histogram of #vote. for different examples x_1 (green and red lines) and x_2 (blue line)
- Denote the fraction of votes that are 1: $q =$ #vote$/n$ (recall from your statistics class that #vote and q are random variables, while $p(x)$ is a constant
-
- 

- Suppose the $p(x)$ is the probability that a randomly generated tree predicts 1 on the input x
- The mean of q is $\mu = p(x)$, the standard-deviation of q is $\sigma = \sqrt{\frac{p(x)*(1-p(x))}{n}}$
- Suppose $p(x_1)=3/4$
    - The green line corresponds to $n = 100$ for which $\mu = 3/4$ and $\sigma = \sqrt{\frac{3/16}{100}} \approx 0.04$
    - The red line corresponds to $n = 400$ for which $\mu = 3/4$ and $\sigma = \sqrt{\frac{3/16}{400}} \approx 0.02$
    - In other words, the red curve is two times narrower than the green curve. (the figure is for qualitative, not quantitative, illustration)
- The final label, for both bagging and random forests, is the majority vote. The label for x_1 is very likely to be 1. The probability that it is zero is equal to the area under the green or red curves that is below 0. As the std corresponding to the red line is smaller, the probability of a 1 prediction decreases.
- The blue line corresponds to x_2, for which $p(x_2) = 0.6$ and $n = 10$, in this case the mean and std of q are $\mu = 0.6$ and $\sigma = \sqrt{\frac{24/100}{10}} \approx 0.15$
- The gap between the mean and the threshold is $0.6 - 0.5 = 0.1$, and the standard deviation is 0.15. This means that the probability of a 0 prediction is pretty high.

- Combining these, we find out that there are two factors that control the probability of making the a-typical prediction:
  - the distance between the true probability $p(x)$ and 0.5. THe larger this distance - the lower the probability of making the a-typical prediction.
  - the number of trees n - The higher this number the lower the probability of making a-typical predictions.
- **Excercise:** You are given $p(x)$ and n, write a formula for the probability that the output of a random tree is 1. Assume that n is large enough that the central limit theorem gives a good approximation and use the notation for the CDF of the normal distribution: $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2}\, dt$
  - $Pr(\text{label} = 1) \approx$