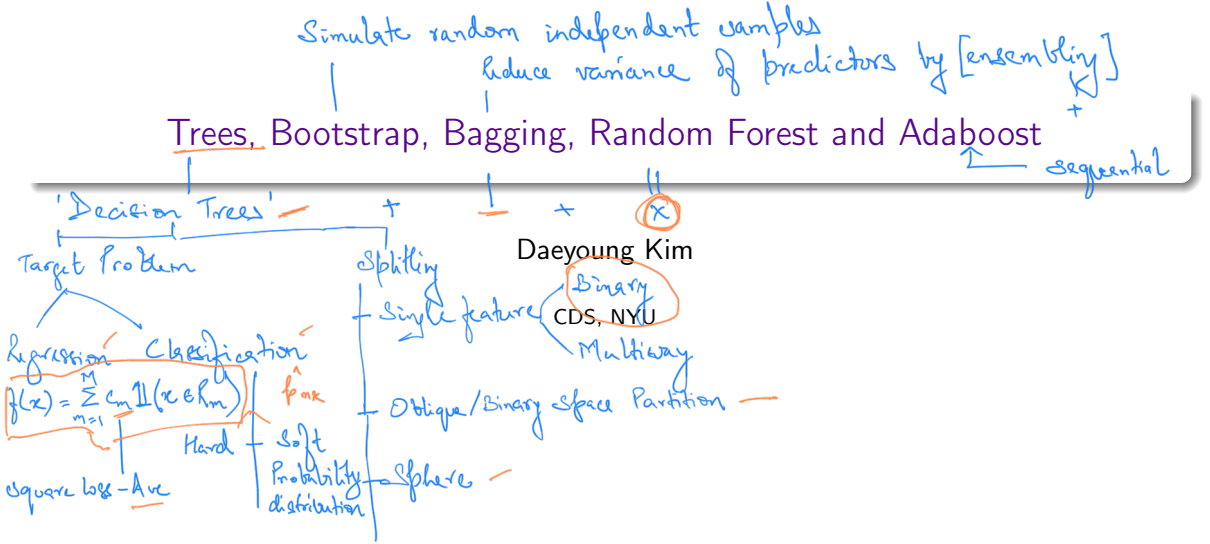

Simulate random independent samples
Reduce variance of predictors by [ensembling]
Trees, Bootstrap, Bagging, Random Forest and Adaboost
sequential
'Decision Trees'
Target Problem
Regression
Classification
Hard
Soft
Probability distribution
square loss - Ave
Splitting
Single feature
Binary
Multiway
Oblique / Binary space Partition
Sphere
Daeyoung Kim
CDS, NYU

Scratch Space

$$\mathbb{E}\left[\hat{\sigma}^2\right] = \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2\right] = \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}(x_i^2 + \bar{x}^2 - 2x_i\bar{x})\right]$$

$$= \frac{1}{N}\mathbb{E}\left[\sum_{i=1}^{N} x_i^2 + N\bar{x}^2 - 2\bar{x}\sum_{i=1}^{N} x_i\right]$$

$N\bar{x}$

$$\mathbb{E}(\bar{x}^2) = \frac{\sigma^2}{N} + \mu^2 \qquad = \frac{1}{N}\mathbb{E}\left[\sum_{i=1}^{N} x_i^2 - N\bar{x}^2\right] \leq$$

$$\boxed{\mathbb{E}(X_i^2) = \mathrm{Var}(X_i) + (\mathbb{E}(X_i))^2}$$

$\sigma^2 \quad + \mu^2$

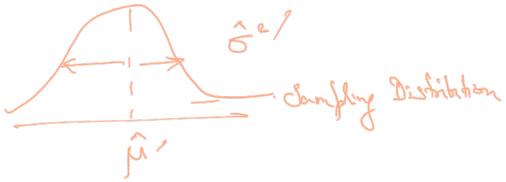
$$- N \cdot \frac{1}{N}\left[\sum_{i=1}^{N} x_i^2 + \sum_{i \neq j} x_i x_j\right]$$

$$\mathbb{E}(XY) = \mathbb{E}(X)\,\mathbb{E}(Y)$$

# Question 1: What is random?

Which of the followings are random?

- statistic
- parameter
- point estimator
- sampling distribution
- standard error
- bias and variance of a point estimator
- bootstrap sample



# [Solution] Question 1: What is random?

Which of the followings are random?

- statistic is any function of the data. It is random since we're considering the data is random. arbitrary
- parameter is any function of the distribution. It is not random.
- point estimator is some statistic which is to serve as a "best guess" of an unknown parameter. It is random.
- sampling distribution is not random
- standard error is a parameter of the sampling distribution. It is not random.
- bias and variance of a point estimator are still parameters of the sampling distribution of the point estimator. It is not random.
- bootstrap sample are random samples from the data. It is random.

# Question 2: Bias and Variance

$\frac{d\ell}{d\hat{\mu}} = +\frac{1}{\hat{\sigma}^2}\sum_{i=1}^{N}(x_i-\hat{\mu}) = 0$ $\quad \frac{d\ell}{d\hat{\sigma}^2} = -\frac{N}{2\hat{\sigma}^2} + \frac{1}{2(\hat{\sigma}^2)^2}\sum_{i=1}^{N}(x_i-\hat{\mu})^2 = 0$

Suppose we have samples $X_1, \ldots, X_N$ from Normal$(\mu, \sigma^2)$. (independent) pdf $\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

- What is the maximum likelihood estimator $\hat{\mu}, \hat{\sigma}^2$ of the parameter $\mu$, $\sigma^2$?
- Are the maximum likelihood estimator unbiased? If not, how do we fix it?
- What is the sampling distribution of $\hat{\mu}$? What is the variance of the estimator $\hat{\mu}$?

MLE $\longrightarrow$ $\hat{\theta} = \arg\max_{\hat{\theta}} L(\hat{\theta}|D) = \arg\max_{\hat{\theta}} \sum_{i=1}^{N} \log p(D|\hat{\theta})$

$\left\{\frac{1}{(2\pi\hat{\sigma}^2)^{N/2}} \exp\left(-\frac{\sum_{i=1}^{N}(x_i-\hat{\mu})^2}{2\hat{\sigma}^2}\right)\right\} \quad -\frac{N}{2}\log 2\pi - \frac{N}{2}\log\hat{\sigma}^2 - \frac{1}{2\hat{\sigma}^2}\sum_{i=1}^{N}(x_i-\hat{\mu})^2$

# [Solution] Question 2: Bias and Variance

Suppose we have samples $X_1, \ldots, X_N$ from Normal$(\mu, \sigma^2)$.

- What is the maximum likelihood estimator $\hat{\mu}, \hat{\sigma}^2$ of the parameter $\mu$, $\sigma^2$?

$$\hat{\mu} = \bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i; \qquad \hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})^2$$

- Are the maximum likelihood estimator unbiased? If not, how do we fix it?

$$\mathbb{E}\hat{\mu} = \mu; \qquad \mathbb{E}\hat{\sigma}^2 = \frac{N-1}{N}\sigma^2$$

So $\hat{\mu}$ is unbiased but $\hat{\sigma}^2$ is not unbiased. $\frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})^2$ is an unbiased estimator of $\sigma^2$.

- What is the sampling distribution of $\hat{\mu}$? What is the variance of the estimator $\hat{\mu}$?
Sampling distribution is Normal $(\mu, \sigma^2/N)$. Variance of the estimator is $\sigma^2/N$.

# Question 3: Bias and Variance 2

Let $X_1, \ldots, X_n$ be an i.i.d. sample from a distribution with mean $\mu$ and variance $\sigma^2$. How large must $n$ be so that the sample mean has standard error smaller than .01?

$$\hat{\sigma} = \sqrt{\mathrm{Var}(\hat{\theta})} = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$$

# [Solution] Question 3: Bias and Variance 2

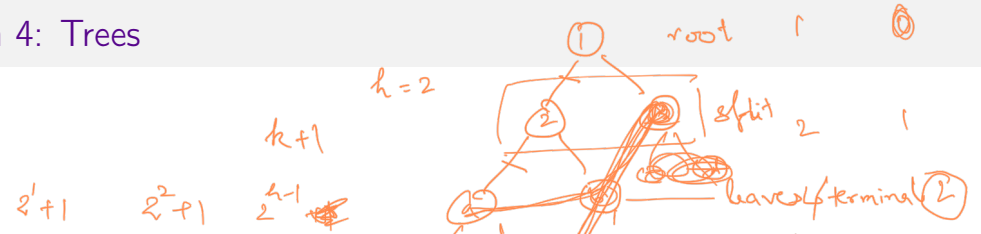
Recall that the sample mean has variance

$$\text{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{\text{Var}(X_1)}{n} = \frac{\sigma^2}{n},$$

with standard error $\sigma/\sqrt{n}$. Thus we have

$$\sigma/\sqrt{n} < .01 \iff n > 10000\sigma^2.$$

# Question 4: Trees




1. How many regions (leaves) will a tree with k node splits have?
2. What is the maximum number of regions a tree of height k can have? Recall that the height of a tree is the number of edges in the longest path from the root to any leaf.
3. Give an upper bound on the depth needed to exactly classify $n$ distinct points in $\mathbb{R}^d$. [Hint: In the worst case each leaf will have a single training point.]

# [Solution] Question 4: Trees

1. Given a fixed tree, if we split a leaf node we add a single leaf to the tree. Thus $k$ splits corresponds to $k+1$ leaves.
2. A tree of height $k$ can have at most $2^k$ regions (leaves).
3. A tree of height $\lceil \log_2(n) \rceil$ is sufficient to distinguish all possible values for the first feature. At each leaf we can then put another tree of this height that distinguishes the second feature, and so forth. These give an upper bound of $d\lceil \log_2(n) \rceil$.

# Question 4: Trees 2

Suppose we are looking at a fixed node of a classification tree, and the class labels are, sorted by the first feature values,

$$4, 1, 0, 0, 1, 0, 2, 3, 3.$$

We are currently testing splitting the node into a left node containing $4, 1, 0, 0, 1, 0$ and a right node containing $2, 3, 3$. For each of the following impurity measures, give the value for the left and right parts, along with the total score for the split.

1. Gini index.
2. Entropy.

# [Solution] Question 4: Trees 2

1. Gini:
   - Left: $3/6(3/6)+2/6(4/6)+1/6(5/6)=22/36$
   - Right: $1/3(2/3)+2/3(1/3)=4/9$
   - Total: $6(22/36)+3(4/9)=30/6=5$
2. Entropy:
   - Left: $-3/6\log(3/6)-2/6\log(2/6)-1/6\log(1/6)$
   - Right: $-1/3\log(1/3)-2/3\log(2/3)$
   - Total: $6[-3/6\log(3/6)-2/6\log(2/6)-1/6\log(1/6)]+3[-1/3\log(1/3)-2/3\log(2/3)]$.

# Question 5: Bootstrap

Let $X_1, \ldots, X_{2n+1}$ be an i.i.d. sample from a distribution. To estimate the median of the distribution, you can compute the sample median of the data.

1. How do we compute an estimate of the variance of the sample median?
2. How do we compute an estimate of a 95% confidence interval for the median.

# [Solution] Question 5: Bootstrap

Let $X_1, \ldots, X_{2n+1}$ be an i.i.d. sample from a distribution. To estimate the median of the distribution, you can compute the sample median of the data.

1. How do we compute an estimate of the variance of the sample median?
   - Draw $B$ bootstrap samples $D^1, \ldots, D^B$ each of size $2n+1$. The samples are formed by drawing uniformly with replacement from the original data set $X_1, \ldots, X_{2n+1}$. We will make a total of $B(2n+1)$ draws.
   - For each $D^i$ compute the corresponding median $\hat{m}_i$.
   - Compute the sample variance of the $B$ medians $m_1, \ldots, m_B$.
2. How do we compute an estimate of a 95% confidence interval for the median.
   - Draw $B$ bootstrap samples $D^1, \ldots, D^B$ each of size $2n+1$. The samples are formed by drawing uniformly with replacement from the original data set $X_1, \ldots, X_{2n+1}$. We will make a total of $B(2n+1)$ draws.
   - For each $D^i$ compute the corresponding median $\hat{m}_i$.
   - Compute the 2.5% and 97.5% sample quantiles of the list $\hat{m}_1, \ldots, \hat{m}_B$. Use these as the estimates of the left and right endpoints of the confidence interval, respectively.

# Question 6: Bagging and Random Forest

- A slide titled "Averaging Independent Prediction Functions" (Lec. 10b p.6/20) states that $\mathrm{Var}(\hat{f}_{avg}(x)) = \frac{1}{B^2}\mathrm{Var}\left(\sum_{b=1}^{B}\hat{f}_b(x)\right) = \frac{1}{B}\mathrm{Var}\left(\hat{f}_1(x)\right)$. Justify each of the two equality signs.
- The above equality gives some intuition as to why bagging might reduce variance. But really, but situation is more complicated: the bootstrap samples used for bagging are not independent. Why not?
- If the variance of the individual predictors that we are bagging is $\sigma^2$, and the correlation between them is $\rho^2$, what is the variance of the bagged predictor?
- Bagging decision trees leads us to the highly popular random forests. However, to make bagging for decision trees work well, we need one more key idea. What is it?

# [Solution] Question 6: Bagging and Random Forest

- A slide titled "Averaging Independent Prediction Functions" states that $\mathrm{Var}(\hat{f}_{avg}(x)) = \frac{1}{B^2}\mathrm{Var}\left(\sum_{b=1}^{B}\hat{f}_b(x)\right) = \frac{1}{B}\mathrm{Var}\left(\hat{f}_1(x)\right)$. Justify each of the two equality signs.
  $\mathrm{Var}\left(\frac{1}{B}\sum_{b=1}^{B}\hat{f}_b(x)\right)$. $\hat{f}_1(x), \hat{f}_2(x), \ldots, \hat{f}_B(x) = \frac{1}{B^2}\mathrm{Var}\left(\sum_{b=1}^{B}\hat{f}_b(x)\right)$ are independent prediction functions. So

$$\frac{1}{B^2}\mathrm{Var}\left(\sum_{b=1}^{B}\hat{f}_b(x)\right) = \frac{1}{B^2}B * \mathrm{Var}\left(\hat{f}_1(x)\right) = \frac{1}{B}\mathrm{Var}\left(\hat{f}_1(x)\right)$$

- The above equality gives some intuition as to why bagging might reduce variance. But really, but situation is more complicated: the bootstrap samples used for bagging are not independent. Why not?
  A bootstrap sample from $\mathcal{D}_n = (x_1, \ldots, x_n)$ is a sample of size $n$ drawn with replacement from $\mathcal{D}_n$. So there can be overlaps between bootstrap samples.

# [Solution] Question 6: Bagging and Random Forest

- If the variance of the individual predictors that we are bagging is $\sigma^2$, and the correlation between them is $\rho^2$, what is the variance of the bagged predictor?

$$\text{Var}\left(\frac{1}{B}\sum_{b=1}^{B}\hat{f}_b\right) = \frac{1}{B^2} * \left(B * \text{Var}(\hat{f}_1(x)) + B(B-1) * \text{Cov}(\hat{f}_1(x), \hat{f}_2(x))\right) = \frac{1}{B}\sigma^2 + \frac{B-1}{B}\rho^2$$

- Bagging decision trees leads us to the highly popular random forests. However, to make bagging for decision trees work well, we need one more key idea. What is it?
  We randomly sample features when building trees to reduce the covariance between trees.

# Question 7: Adaboost-Concept Check[1]

Decide whether each of the statements below is true or false.

F
- If a weak classifier has a weighted error rate $\epsilon \leqslant 1/3$, it can only misclassify up to $1/3$ of the training points.

F
- The error rate of the ensemble classifier never increases from one round to the next.

F
- Adaboost accounts for outliers by lowering the weights of training points that are repeatedly misclassified.

F
- When you update weights, the training point with the smallest weight in the previous round will always increase in weight.

[1] From MIT exams

# Adaboost Algorithm

Given training set $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$.

1. Initialize observation weights $w_i = 1,\ i = 1, 2, \ldots, n$.
2. For $m = 1$ to $M$:
   1. Base learner fits weighted training data and returns $G_m(x)$
   2. Compute **weighted empirical 0-1 risk**:

$$\text{err}_m = \frac{1}{W} \sum_{i=1}^{n} w_i 1(y_i \neq G_m(x_i)) \quad \text{where } W = \sum_{i=1}^{n} w_i.$$

   3. Compute $\alpha_m = \ln\left(\frac{1-\text{err}_m}{\text{err}_m}\right)$ **[classifier weight]**
   4. Set $w_i \leftarrow w_i \cdot \exp\left[\alpha_m 1(y_i \neq G_m(x_i))\right], \quad i = 1, 2, \ldots, n$ **[example weight adjustment]**
3. Ouptut $G(x) = \text{sign}\left[\sum_{m=1}^{M} \alpha_m G_m(x)\right]$.

# [Solution] Question 7: Adaboost-Concept

Decide whether each of the statements below is true or false.

- If a weak classifier has a weighted error rate $\epsilon \leqslant 1/3$, it can only misclassify up to $1/3$ of the training points.
  **False**. We use weighted error rate.
- The error rate of the ensemble classifier never increases from one round to the next.
  **False**. There is no guarantee for error rate not to increase from one round to the next.
- Adaboost accounts for outliers by lowering the weights of training points that are repeatedly misclassified.
  **False**. Adaboost will increase the weights of training points that are repeatedly misclassified.
- When you update weights, the training point with the smallest weight in the previous round will always increase in weight.
  **False**. It will increase weight only if it is misclassified in the current round.

# Question 8: Adaboost-Algorithm[2]

Consider building an ensemble of decision stumps Gm with the AdaBoost algorithm,

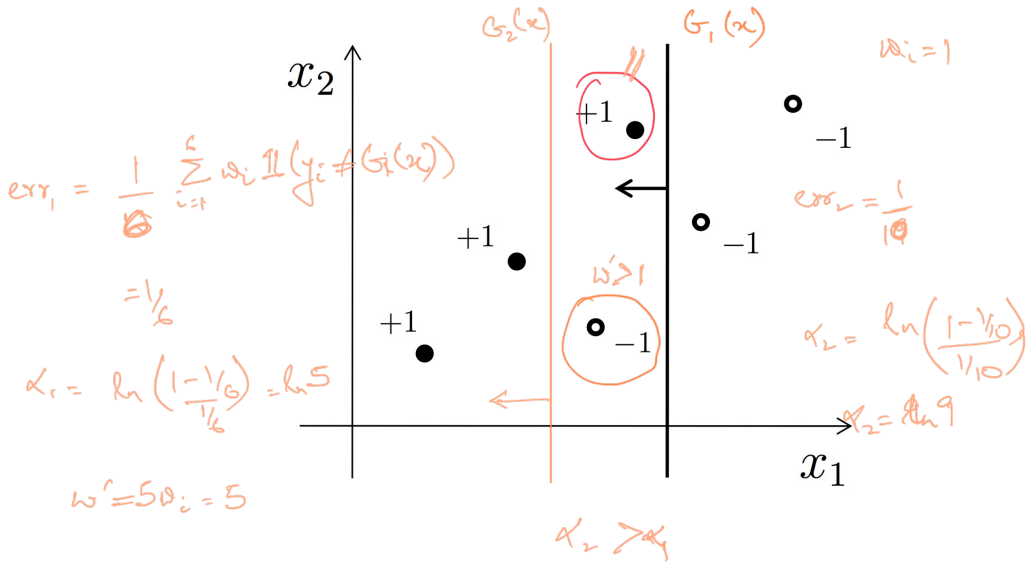
$$f(x) = \text{sign}\left(\sum_{m=1}^{M} \alpha_m G_m(x)\right)$$

displays a few labeled point in two dimensions as well as the first stump we have chosen. A stump predicts binary $\pm 1$ values, and depends only on one coordinate value (the split point). The little arrow in the figure is the normal to the stump decision boundary indicating the positive side where the stump predicts $+1$. All the points start with uniform weights.

- Circle all the point(s) in Figure 1 whose weight will increase as a result of incorporating the first stump (the weight update due to the first stump).
- Draw in the same figure a possible stump that we could select at the next boosting iteration. You need to draw both the decision boundary and its positive orientation.
- Will the second stump receive higher coefficient in the ensemble than the first? In other words, will $\alpha_2 > \alpha_1$? Briefly explain your answer. (no calculation should be necessary).

[2] From CMU exams

# Question 8:Adaboost-Algorithm



# [Solution] Question 8:Adaboost-Algorithm

- Circle all the point(s) in Figure 1 whose weight will increase as a result of incorporating the first stump (the weight update due to the first stump).
  (sol.) The only misclassified negative sample.
- Draw in the same figure a possible stump that we could select at the next boosting iteration. You need to draw both the decision boundary and its positive orientation.
  (sol.) The second stump will also be a vertical split between the second positive sample (from left to right) and the misclassified negative sample, as drawn in the figure.
- Will the second stump receive higher coefficient in the ensemble than the first? In other words, will $\alpha_2 > \alpha_1$? Briefly explain your answer. (no calculation should be necessary).
  (sol.) $\alpha_2 > \alpha_1$ because the point that the second stump misclassifies will have a smaller relative weight since it is classified correctly by the first stump.