QUIZ3

1. **Decision Tree**
   Impurity functions play an important role in decision tree branching. For binary classification problems, let μ_+ be the fraction of positive examples in a data subset, and $\mu_- = 1 - \mu_+$ be the fraction of negative examples in the data subset. The Gini index is $1 - \mu_+^2 - \mu_-^2$. What is the maximum value of the Gini index among all $\mu_+ \in [0, 1]$?

   A. 0.5

   B. 0.75

   C. 0.25

   D. 0

   E. 1

2. Following Question 1, there are four possible impurity functions below. We can normalize each impurity function by dividing it with its maximum value among all $\mu_+ \in [0, 1]$ For instance, the classification error is simply $\min(\mu_+, \mu_-)$ and its maximum value is 0.5. So the normalized classification error is $2\min(\mu_+, \mu_-)$. After normalization, which of the following impurity function is equivalent to the normalized Gini index?

   A. the squared regression error (used for branching in classification data sets), which is by definition $\mu_+(1 - (\mu_+ - \mu_-))^2 + \mu_-(-1 - (\mu_+ - \mu_-))^2$.

   B. the entropy, which is $-\mu_+ \ln \mu_+ - \mu_- \ln \mu_-$, with $0 \log 0 \equiv 0$.

   C. the closeness, which is $1 - |\mu_+ - \mu_-|$.

   D. the classification error $\min(\mu_+, \mu_-)$.

   E. none of the other choices

3. **Random Forest**
   If bootstrapping is used to sample $N' = pN$ examples out of N examples and N is very large. Approximately how many of the N examples will not be sampled at all?

   A. $(1 - e^{-1/p}) \cdot N$

   B. $(1 - e^{-p}) \cdot N$

   C. $e^{-1} \cdot N$

   D. $e^{-1/p} \cdot N$

   E. $e^{-p} \cdot N$