模式识别与统计学习 不纯度指数作业

题目：计算下图中A、B两种情况的不纯度指数，其中：

1. Entropy:
2. Misclassification rate:
3. Gini index:

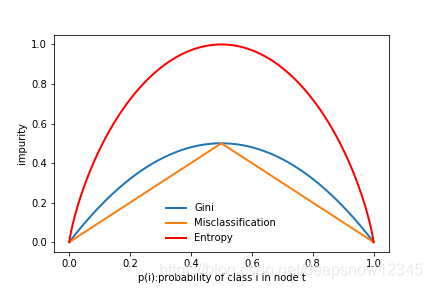
解答：

<A>

1. Entropy:
2. Misclassification rate:
3. Gini index:

<B>

1. Entropy:
2. Misclassification rate:
3. Gini index:



综合A、B计算结果和曲线图，<B>中使用Entropy和Gini Impurity计算的增益都比<A>高，所以选择<B>生成子节点。在计算过程中也发现Misclassification rate的值相同，无法判断两种分类方式的优劣，所以进行以下总结：

1. Misclassification rate是将结点内整体类别归为一类的错误比例，是单一的比例衡量； Entropy和Gini Impurity均是对结点内整体分布的衡量。
2. Misclassification rate和Gini Impurity的值域在0～0.5之间，而Entropy在0～1之间，所以在设置信息增益阈值处，Entropy可设置的更高一些。
3. Misclassification rate对于不同区分度的降低的评价是相同的，即其导数为常量。但是Entropy和Gini Impurity的导数随p(i)的增加而降低，且Entropy比Gini Impurity要更高一些。故，Entropy和Gini Impurity更公平一些。