# CMSC5724: Quiz 1

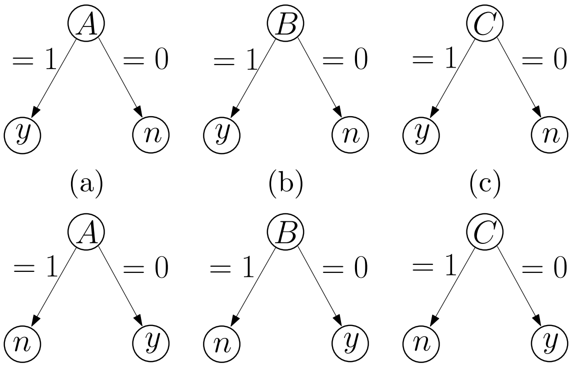
Hand-write all your solutions on paper. Take a picture of the paper **together with** your CUHK student ID card. Upload the picture to Blackboard or email it to the instructor at taoyf@cse.cuhk.edu.hk. Your must do so within 15 minutes after the quiz has started.

**Problem 1 (60%).** Consider the training data shown below. Here, $A, B$, and $C$ are attributes, and $Y$ is the class label.

| $A$ | $B$ | $C$ | $Y$ |
|---|---|---|---|
| 1 | 1 | 1 | y |
| 0 | 1 | 1 | y |
| 0 | 0 | 1 | y |
| 1 | 1 | 0 | y |
| 1 | 0 | 1 | n |
| 1 | 1 | 1 | n |
| 0 | 0 | 0 | n |
| 1 | 0 | 0 | n |

Suppose that we consider only decision trees each having 3 nodes (i.e., a root node and two leaves). Give the decision tree with the best empirical error. You need to explain your reasoning.

**Answer.** For the given input, there are only 6 possible decision trees having 3 nodes, which are:



(a) (b) (c)

(d) (e) (f)

Among them, the decision tree (b) has the lowest empirical error 1/4 and, hence, is the answer.

**Problem 2 (40%).** Use the generalization theorem (in Lecture Notes 1) to estimate the generalization error of your decision tree in Problem 1. Again, we consider only the decision trees with 3 nodes. Your estimate should be correct with probability at least 99%.

**Answer.** Les $S$ be the training set given in Problem 1 and $\mathcal{H}$ be the set of classifiers that can possibly be returned. Denote by $h$ the best decision tree we found in Problem 1. From the above solution, we know $|\mathcal{H}| = 6$ and the empirical error $err_S(h) = 1/4$.

According to the generalization theorem, with probability at least $1 - \delta$, we have

$$\begin{aligned} err_{\mathcal{D}}(h) &\le err_S(h) + \sqrt{\frac{\ln(1/\delta) + \ln|\mathcal{H}|}{2|S|}} \\ &\le 1/4 + \sqrt{\frac{\ln(1/\delta) + \ln 6}{16}}. \end{aligned}$$

By setting $\delta = 0.01$, we know with probability at least 0.99,

$$err_{\mathcal{D}}(h) \quad \leq \quad 1/4 + \sqrt{\frac{\ln(1/0.01) + \ln 6}{16}}.$$