Homework 3

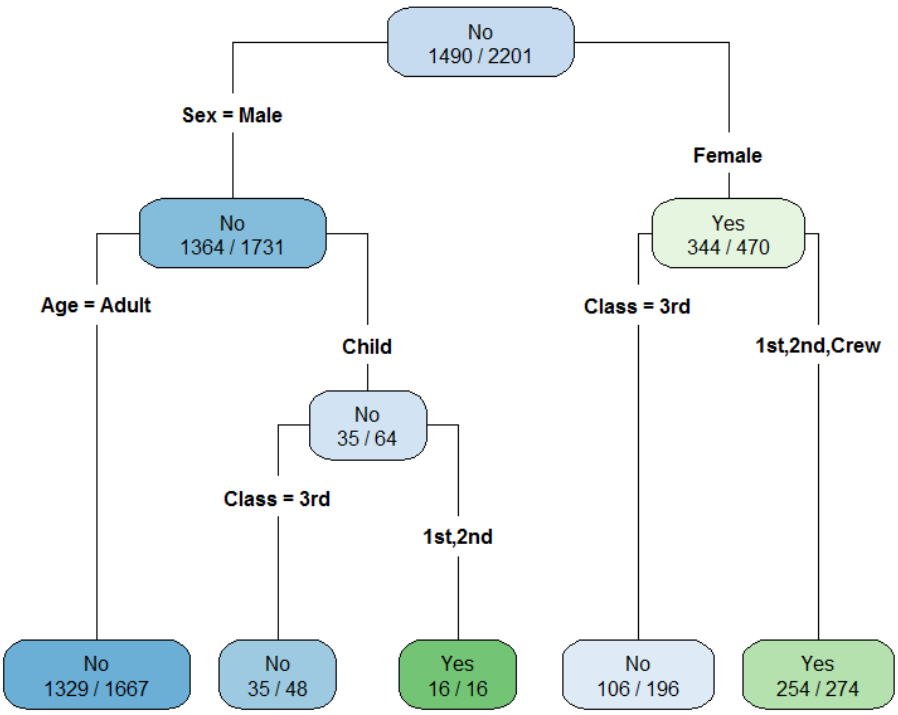
DSA1101
Introduction to Data Science

October 21, 2018

Name:

Matriculation card number:

Problem 1 (15 points). *Decision Trees*
Consider the following decision tree for predicting survival on *Titanic*. In each node, the number on the right refers to the total number of data points falling into that node, while the number on the left refers to the number of data points with outcome labels that are the same as the predicted label for that node.



(a) (5 points) Calculate the base entropy D_{survived} for predicting survival at the root node.

$$\begin{aligned} D_{\text{survived}} &= -\left\{ \frac{1490}{2201} \log_2 \left(\frac{1490}{2201} \right) + \frac{711}{2201} \log_2 \left(\frac{711}{2201} \right) \right\} \\ &\approx 0.908 \end{aligned}$$

(b) (8 points) Calculate the conditional entropy $D_{\text{survived|sex}}$ for predicting survival using the feature variable *Sex* as decision variable with the split at *Male* versus *Female*.

$P(Sex = Male) = \frac{1731}{2201}$, $P(Sex = Female) = \frac{470}{2201}$

$$\begin{aligned} D_{\text{survived|sex}} = -\Bigg\{ & \frac{1731}{2201} \left[\frac{1364}{1731} \log_2 \left(\frac{1364}{1731} \right) + \frac{367}{1731} \log_2 \left(\frac{367}{1731} \right) \right] \\ & + \frac{470}{2201} \left[\frac{126}{470} \log_2 \left(\frac{126}{470} \right) + \frac{344}{470} \log_2 \left(\frac{344}{470} \right) \right] \Bigg\} \\ = & \approx 0.766 \end{aligned}$$

(c) (2 points) Hence calculate the entropy reduction (or equivalently information gain) associated with using *Sex* to predict survival on *Titanic*.
$\approx 0.908 - 0.766 = 0.142$

Problem 2 (10 points). *Naïve Bayes Classifier*
Consider the following dataset on whether to play golf, given factors such as temperature, humidity, and wind.

Play	Temperature	Humidity	Wind
yes	cool	normal	FALSE
no	cool	normal	TRUE
yes	hot	high	FALSE
no	mild	high	FALSE
yes	cool	normal	FALSE
yes	cool	normal	FALSE
yes	cool	normal	FALSE
yes	hot	normal	FALSE
yes	mild	high	TRUE
no	mild	high	TRUE

(a) (10 points) Predict whether a person will play golf on a day with wind, mild temperature and high humidity using a naïve Bayes classifier. Show your working for computing the probability scores for making the prediction.

$P(Play = yes) = 0.7,\ P(Play = no) = 0.3$

$$P(Play = yes|Wind = T, Humidity = high, Temp = mild) \propto$$
$$0.7 \times \frac{1}{7} \times \frac{2}{7} \times \frac{1}{7}$$
$$\approx 0.0041$$

$$P(Play = no|Wind = T, Humidity = high, Temp = mild) \propto$$
$$0.3 \times \frac{2}{3} \times \frac{2}{3} \times \frac{2}{3}$$
$$\approx 0.089$$

The preson is predicted not to play golf since $0.089 > 0.0041$