Assignment 2 **Spring 2023**
CIS530 Advanced Data Mining

Due: 02/26/2022 23:59PM Submit to myCourses

Note:

1. *Please provide necessary derivations and steps. Solutions with results only (e.g., a number) will get 0 pts!* ***Type the solutions for all questions in Microsoft word****. DO NOT write in a piece of paper.*
2. *Please provide independent source files named by the question #.*
3. *You are allowed to use MATLAB or Python for experiments.*
4. *Only one student needs to submit solution to myCourses on behalf of the group.*
5. *Please list group ID and member names on the cover page.*

Q1. (6 pts) What are typical reasons for overfitting? Please list at least three reasons.

Q2. (10 pts) In Slides-5, page 9, there is a question "In fruit example, if each box contained the same fraction of apples and oranges then $p(F|B) = p(F)$" Can you prove that?

Q3. (14 pts) True or False

a. From a Bayesian perspective, the probability is a quantification of uncertainty
b. Uniform distribution will provide the smallest entropy for discrete variables
c. With more nodes included in the decision tree, the testing error will become smaller and smaller
d. If we toss a coin three times and always get "head", then from Bayesian perspective, we believe $\Pr(x \text{ is head}) = 1$
e. There might be more than one decision trees that fit the same data.
f. In decision tree, higher impurity means lower Gini values.

Q4. (15 pts) Consider the below joint probability distribution between x and y.

		y = 0	y = 1
x	0	1/3	1/3
	1	0	1/3

a. Please compute the marginal probabilities $p(x)$ and $p(y)$ for each value x and y take on.
b. Are x and y independent? Prove your result.
c. Compute conditional probabilities $p(x|y)$ and $p(y|x)$ for each value x and y take on.

Q5. (15 pts) Recall the fruit problem in Slides-5 page 8, can you find the probability of $p(B = b|F = a)$? Please also indicate the prior, likelihood, and posterior terms in your solutions.

Q6. (20 pts) Given N independent and identically distributed (i.i.d.) observations $(x_1, x_2, \dots x_N) \in \mathbb{R}^{1\times N}$ that follow the Gaussian distribution $N(\mu, \sigma^2)$. Please answer the following questions:

a. Formulation of joint probability Pr $(x_1, x_2, \dots x_N)$
b. Log-likelihood function of the above joint probability
c. Maximum likelihood solution to μ and σ^2
d. Are the solutions to question (c) biased estimations or not, and why?
e. Consider the curve fitting problem $y = \mathbf{w}x$. Following Slides-5, we assume the observations of y_i are means for Gaussians $N(t_i|y_i = \mathbf{w}x_i, \sigma^2)$ that will be able to model the probability for different target values t_i, and different Gaussians share $\sigma^2 = 1/\beta$. Please provide the maximum likelihood estimation for **w (**derivations and details of each step are required).

Q7. (20 pts) First, please compute the: **(1) entropy; (2) Gini; (3) classification error** of cases (a) and (b). Assume they are data at a node of decision tree.

a. C0: 8 and C1: 12
b. C0: 6 and C1: 14

Second, assume we plan to split the data for case (a) and (b) in the follow ways. Could you calculate the **information gain** for each? Do you suggest a split based on these, and why?

a. C0:1 and C1:9 + C0: 7 and C1:3
b. C0:3 and C1:3 + C0: 3 and C1:11