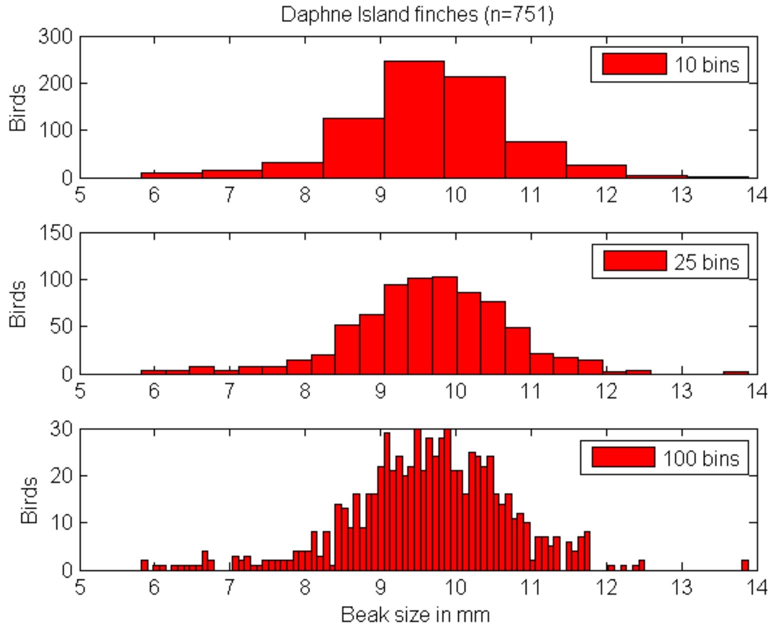
NTUST, CSIE
Machine Learning (CS5087701), Fall 2018

Homework 1 (16pts)

Due date: Nov. 6 (Q1.1-3) & Nov. 13 (the rest)

Question 1.1. [2pts] Given the following three histograms for data visualization, answer a few questions related to some basic concepts of machine learning.

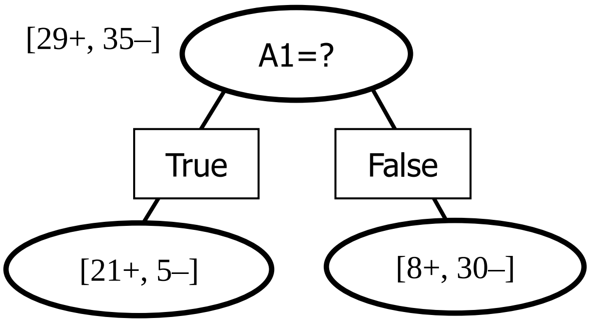


(a) If the middle histogram is the true distribution, using the top or the bottom histogram as the model may lead to what consequence?

(b) Argue why the middle histogram could be the best choice for visualization. In fact, the middle one is also the one most likely to be the best model among three, why? Why the bottom one is unlikely to be the true model? hint: There is no single answer for this. You may use (but not limited to) the concepts of *underfitting*, *overfitting*, *Occam's razor*, *VC dimension* to explain your idea!

Question 1.2. [2pts] Explain the pros and cons of using *k*-nearest neighbors as a machine learning model.

Question 1.3. [2pts] Consider the following tree splitting and two questions.



(a) A statement says: "To decide the best attribute in each splitting of decision induction, instead of computing *information gain*, you just need to compute the expected (average) entropy in the lower level". Do you think it is correct or not?

(b) The similar question is asked again, but the criterion "information gain" is substituted by "gain ratio". What is your answer then?

Question 1.4. [10pts] Analyze the following two datasets:

the *Bank Marketing* dataset
(http://archive.ics.uci.edu/ml/datasets/Bank+Marketing) from
UCI (http://archive.ics.uci.edu/ml/), and
the Spooky Author Identification from Kaggle
(https://www.kaggle.com/c/spooky-author-identification).

You are recommended to use *C*4.5 (or *C*5.0 that you can use in our lab), the decision tree from scikit-learn of Python or from Weka. After your analysis, you should write a mini report of around one or two pages with a discussion section. In your report, you should include the following items:

(a) List all the parameters for the models that you used.

(b) The prediction accuracy with cross-validation and possible different data partitions.

(c) Explain the result you obtain, e.g., why you have a particular attribute as the root of the tree, the tree size etc.

(d) Which dataset is the one we obtain a better result from decision trees? Why?

(e) Give the reasons why the result is good (or bad) for different experimental settings (pre-pruning, post-pruning strategies, etc.).

(f) (Bonus) Can you suggest any approach for re-building the tree or revising the tree so that the prediction result is better? (hint: manually selecting some particular attributes, transforming the attributes from categorical ones to numerical ones or the other way around.)