**Know Your Data**

1. The ArnetMiner citation dataset (provided by arnetminer.org) by year 2012 can be downloaded in the attached file.
   (1) Count the number of authors, venues (conferences/journals), and publications in the datasets.
   (2) What are the min, max, Q1, Q3, and median number of publications per author? Can you plot the histogram for number of publications per author?
   (3) What are the min, max, Q1, Q3, and median number of citations per author? Can you plot the histogram for number of citations received per author?
   (4) Please plot the scatter plot between the number of publications vs. the number of citations for authors who have more than 5 publications.

**Classification for Matrix Data**

2. Decision Tree

Construct a decision tree for the following training data, where "Edible" is the class we are going to predict. Information gain is used to select the attributes. Please write down the major steps in the construction process (you need to show the information gain for each candidate attribute when a new node is created in the tree).

| Color | Size | Shape | Edible? |
|---|---|---|---|
| Yellow | Small | Round | + |
| Yellow | Small | Round | - |
| Green | Small | Irregular | + |
| Green | Large | Irregular | - |
| Yellow | Large | Round | + |
| Yellow | Small | Round | + |
| Yellow | Small | Round | + |
| Yellow | Small | Round | + |
| Green | Small | Round | - |
| Yellow | Large | Round | - |
| Yellow | Large | Round | + |
| Yellow | Large | Round | - |
| Yellow | Large | Round | - |
| Yellow | Large | Round | - |
| Yellow | Small | Irregular | + |
| Yellow | Large | Irregular | + |

3. Naïve Bayes

Consider a Naïve Bayes model for spam classification with the vocabulary V = {secret, offer, low, price, valued, customer, today, dollar, million, sports, is, for, play, healthy, pizza}, where each word in the vocabulary is considered as a feature, and their values could be either 1 or 0, denoting whether they exist in one message. We have the messages and labels in the following table:

| Messages | Class label |
|---|---|
| Million dollar offer | Spam |
| Secret offer today | Spam |
| Secret is secret | Spam |
| Low price for valued customer | non-spam |
| Play secret sports today | non-spam |
| Sports is healthy | non-spam |
| Low price pizza | non-spam |

Give the MLEs for the following parameters:$\theta_{spam} = P(C_{spam})$, $\theta_{secret|spam} = P(secret = 1|C_{spam})$, $\theta_{secret|non-spam} = P(secret = 1|C_{non-spam})$, $\theta_{sports|non-spam} = P(sports = 1|C_{non-spam})$, and $\theta_{dollar|spam} = P(dollar = 1|C_{spam})$.

4. Support Vector Machine

| # | x1 | x2 | class |
|---|---|---|---|
| 1 | 2.46 | 2.59 | 1 |
| 2 | 3.05 | 2.87 | 1 |
| 3 | 1.12 | 1.64 | 1 |
| 4 | 0.01 | 1.44 | 1 |
| 5 | 2.20 | 3.04 | 1 |
| 6 | 0.41 | 2.04 | 1 |
| 7 | 0.53 | 0.77 | 1 |
| 8 | 1.89 | 2.64 | 1 |
| 9 | -0.39 | 0.96 | 1 |
| 10 | -0.96 | 0.08 | 1 |
| 11 | 2.65 | -1.33 | -1 |
| 12 | 1.57 | -1.70 | -1 |
| 13 | 3.05 | 0.01 | -1 |
| 14 | 2.66 | -1.15 | -1 |
| 15 | 4.51 | -0.52 | -1 |
| 16 | 3.06 | -0.82 | -1 |
| 17 | 3.16 | -0.56 | -1 |
| 18 | 2.05 | -0.62 | -1 |
| 19 | 0.71 | -2.47 | -1 |
| 20 | 1.63 | -0.91 | -1 |

Given 20 data points and their class labels in the above, suppose by solving the dual form of the quadratic programming of svm, we can derive the $\alpha's$ for each data point as follows:
$\alpha_7 = 0.4952$
$\alpha_{18} = 0.0459$
$\alpha_{20} = 0.4493$
Others = 0

(1) Please point out the support vectors in the training points.
(2) Calculate the normal vector of the hyperplane: w

(3) Calculate the bias b, according to $b = \sum_{k:\alpha_k \neq 0}(y_k - w'x_k)/N_k$ , where $x_k = (x_{k1}, x_{k2})'$ indicate the support vectors and $N_k$ is the total number of support vectors.
(4) Write down the learned decision boundary function $f(x) = w'x + b$ (the hyperplane) by substituting w and b with learned values in the formula.
(5) Suppose there is a new data point $x = (-1,2)$, please use the decision boundary to predict its class label.

**Bonus Question**

5. Mutual Information and Information Gain

In information theory, mutual information between two discrete random variables is defined as:

$$I(X;Y) = \sum_x \sum_y p(x,y)\log(\frac{p(x,y)}{p(x)p(y)})$$

Which is designed for evaluating the mutual dependence of two random variables. What is the connection between mutual information and information gain we have learned in decision tree? Can you prove it? (Hint: consider Y as the class label, and X as the attribute to predict Y.)