

CS6220 Datamining Fall 2014 Homework 1, Wei Luo

Know Your Data

1. (1) The count of each item is:

Number of authors: 1484984

Number of publications: 1977248

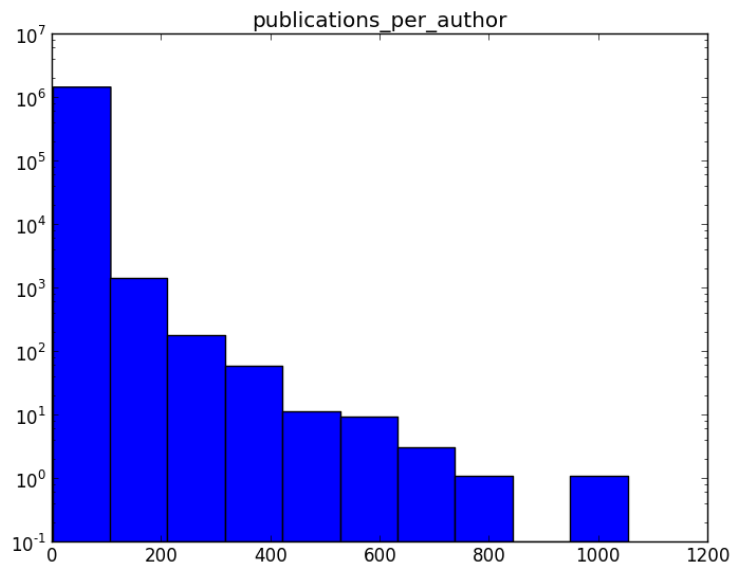
Number of venues: 255686

(2) For publications per author, the statistical values are:

min	max	q1	q3	median
-----	-----	----	----	--------

1	1054	1	2	1
---	------	---	---	---

The histogram for number of publications per author is:

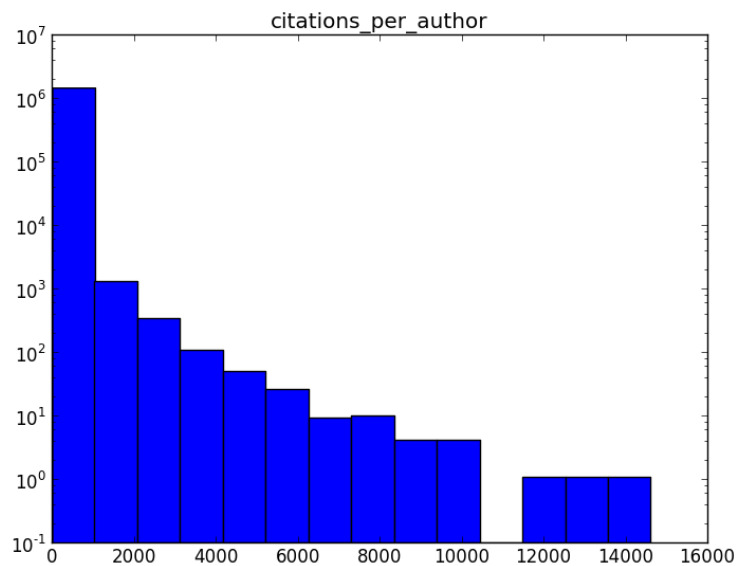


(3) For citations per author, the statistical values are:

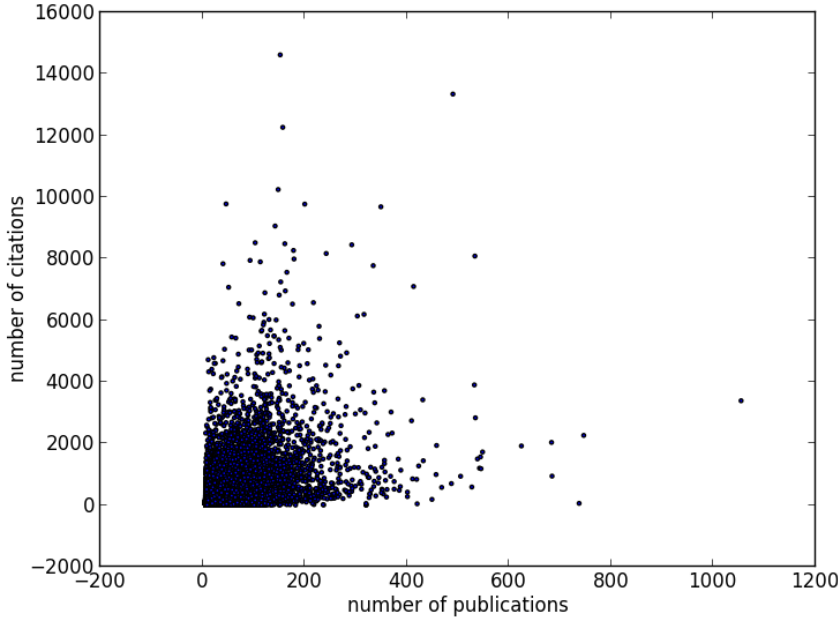
min	max	q1	q3	median
-----	-----	----	----	--------

0	14618	0	4	0
---	-------	---	---	---

The histogram for number of citations per author is:



(4) The scatter plot between the number of publications vs. the number of citations for authors who have more than 5 publications is:



Classification for Matrix Data

2. Decision Tree

From the problem we have features:

Color = {Yellow, Green}, Size = {Small, Large}, Shape = {Round, Irregular}

And label: Edible = {+, -}

Since each feature is binary, we don't have to consider it as a candidate again in the sub-trees once it is chosen for split.

And probabilities:

$$\begin{aligned} P(g) &= P(\text{Color} = \text{Green}) & P(y) &= P(\text{Color} = \text{Yellow}) \\ P(s) &= P(\text{Size} = \text{Small}) & P(l) &= P(\text{Size} = \text{Large}) \\ P(r) &= P(\text{Shape} = \text{Round}) & P(i) &= P(\text{Shape} = \text{Irregular}) \\ P(+) &= P(\text{Edible} = +) & P(-) &= P(\text{Edible} = -) \end{aligned}$$

Before split:

$$\begin{aligned} P(g) &= \frac{3}{3+13} = \frac{3}{16} & P(y) &= \frac{13}{3+13} = \frac{13}{16} & P(s) &= \frac{8}{8+8} = \frac{1}{2} & P(l) &= \frac{8}{8+8} = \frac{1}{2} \\ P(r) &= \frac{12}{12+4} = \frac{3}{4} & P(i) &= \frac{4}{12+4} = \frac{1}{4} & P(+) &= \frac{9}{9+7} = \frac{9}{16} & P(-) &= \frac{7}{9+7} = \frac{7}{16} \\ P(+|g) &= \frac{1}{1+2} = \frac{1}{3} & P(-|g) &= \frac{2}{1+2} = \frac{2}{3} & P(+|y) &= \frac{8}{8+5} = \frac{8}{13} & P(-|y) &= \frac{5}{8+5} = \frac{5}{13} \\ P(+|s) &= \frac{6}{6+2} = \frac{3}{4} & P(-|s) &= \frac{2}{6+2} = \frac{1}{4} & P(+|l) &= \frac{3}{3+5} = \frac{3}{8} & P(-|l) &= \frac{5}{3+5} = \frac{5}{8} \\ P(+|r) &= \frac{6}{6+6} = \frac{1}{2} & P(-|r) &= \frac{6}{6+6} = \frac{1}{2} & P(+|i) &= \frac{3}{3+1} = \frac{3}{4} & P(-|i) &= \frac{1}{3+1} = \frac{1}{4} \end{aligned}$$

$$H(\text{Edible}) = P(+)\log_2\left(\frac{1}{P(+)}\right) + P(-)\log_2\left(\frac{1}{P(-)}\right) = 0.9987$$

$$H(\text{Edible}|\text{Color}) = \sum_i P(\text{Color}_i) \sum_j P(\text{Edible}_j|\text{Color}_i) \log_2 \frac{1}{P(\text{Edible}_j|\text{Color}_i)} = 0.9532$$

$$H(\text{Edible}|\text{Size}) = \sum_i P(\text{Size}_i) \sum_j P(\text{Edible}_j|\text{Size}_i) \log_2 \frac{1}{P(\text{Edible}_j|\text{Size}_i)} = 0.8829$$

$$H(\text{Edible}|\text{Shape}) = \sum_i P(\text{Shape}_i) \sum_j P(\text{Edible}_j|\text{Shape}_i) \log_2 \frac{1}{P(\text{Edible}_j|\text{Shape}_i)} = 0.9528$$

$$IG(\text{Color}) = H(\text{Edible}) - H(\text{Edible}|\text{Color}) = 0.0455$$

$$IG(\text{Size}) = H(\text{Edible}) - H(\text{Edible}|\text{Size}) = 0.1158$$

$$IG(\text{Shape}) = H(\text{Edible}) - H(\text{Edible}|\text{Shape}) = 0.0459$$

We choose feature Size for split at *root* node since it has the most information gain.

Now for the left sub-tree, there are 8 data points, given that their Size is Small. (We are at *node*₁, the left child of *root* node)

$$\begin{aligned} P(+) &= \frac{3}{4} & P(-) &= \frac{1}{4} \\ P(g) &= \frac{1}{4} & P(y) &= \frac{3}{4} & P(r) &= \frac{3}{4} & P(i) &= \frac{1}{4} \\ P(+|g) &= \frac{1}{2} & P(-|g) &= \frac{1}{2} & P(+|y) &= \frac{5}{6} & P(-|y) &= \frac{1}{6} \\ P(+|r) &= \frac{2}{3} & P(-|r) &= \frac{1}{3} & P(+|i) &= 1 & P(-|i) &= 0 \end{aligned}$$

$$H(\text{Edible}) = P(+)\log_2\left(\frac{1}{P(+)}\right) + P(-)\log_2\left(\frac{1}{P(-)}\right) = 0.8113$$

$$H(\text{Edible}|\text{Color}) = \sum_i P(\text{Color}_i) \sum_j P(\text{Edible}_j|\text{Color}_i) \log_2 \frac{1}{P(\text{Edible}_j|\text{Color}_i)} = 0.7375$$

$$H(\text{Edible}|\text{Shape}) = \sum_i P(\text{Shape}_i) \sum_j P(\text{Edible}_j|\text{Shape}_i) \log_2 \frac{1}{P(\text{Edible}_j|\text{Shape}_i)} = 0.6887$$

$$IG(\text{Color}) = H(\text{Edible}) - H(\text{Edible}|\text{Color}) = 0.0738$$

$$IG(\text{Shape}) = H(\text{Edible}) - H(\text{Edible}|\text{Shape}) = 0.1226$$

We choose feature Shape for split at *node*₁ since it has more information gain.

Now we move one step further to the left sub-tree, there are 6 data points, given that their Size is Small and Shape is Round. (We are at *node*₂, the left child of *node*₁)

Since there is only one feature Color left as candidate, we use it for split. Since $P(-|g) = 1$ and $P(+|y) = \frac{3}{4}$, we make a leaf node of $-$ at the Green branch, a leaf node of $+$ at the Yellow branch.

We go back to the right child of *node*₁, there are 2 data points, given that their Size is Small and Shape is Irregular. Since all labels are $+$, we make a leaf node of $+$ here.

Now we go back to the right child node of *root* node. There are 8 data points, given that their Size is Large. (We are at *node*₃, the right child of *root* node)

$$\begin{aligned} P(+) &= \frac{3}{8} & P(-) &= \frac{5}{8} \\ P(g) &= \frac{1}{8} & P(y) &= \frac{7}{8} & P(r) &= \frac{3}{4} & P(i) &= \frac{1}{4} \\ P(+|g) &= 0 & P(-|g) &= 1 & P(+|y) &= \frac{3}{7} & P(-|y) &= \frac{4}{7} \\ P(+|r) &= \frac{1}{3} & P(-|r) &= \frac{2}{3} & P(+|i) &= \frac{1}{2} & P(-|i) &= \frac{1}{2} \end{aligned}$$

$$H(\text{Edible}) = P(+)\log_2\left(\frac{1}{P(+)}\right) + P(-)\log_2\left(\frac{1}{P(-)}\right) = 0.9544$$

$$H(\text{Edible}|\text{Color}) = \sum_i P(\text{Color}_i) \sum_j P(\text{Edible}_j|\text{Color}_i) \log_2 \frac{1}{P(\text{Edible}_j|\text{Color}_i)} = 0.7099$$

$$H(\text{Edible}|\text{Shape}) = \sum_i P(\text{Shape}_i) \sum_j P(\text{Edible}_j|\text{Shape}_i) \log_2 \frac{1}{P(\text{Edible}_j|\text{Shape}_i)} = 0.9387$$

$$IG(\text{Color}) = H(\text{Edible}) - H(\text{Edible}|\text{Color}) = 0.0157$$

$$IG(\text{Shape}) = H(\text{Edible}) - H(\text{Edible}|\text{Shape}) = 0.2445$$

We choose feature Shape for split at *node*₃ since it has more information gain.

Now we move one step further to the left sub-tree, there are 6 data points, given that their Size is Large and Shape is Round. All the Color for this level is Yellow, so we stop splitting. And since $P(-) = \frac{3}{4}$, we make a leaf node of $-$ here.

Now we go back and see the right child of *node*₃, there are 2 data points, given that their Size is Large and Shape is Irregular. (We are at *node*₄, the right child of *node*₃)

Since there is only one feature Color left as candidate, we use it for split. Since $P(-|g) = 1$ and $P(+|y) = 1$, we make a leaf node of $-$ at the Green branch, a leaf node of $+$ at the Yellow branch.

Then we finished building the decision tree.

3. Naïve Bayes

By looking in to three features {secret, sports, dollar}, the message table can be abstracted as:

secret	sports	dollar	label
0	0	1	spam
1	0	0	spam
1	0	0	spam
0	0	0	non-spam
1	1	0	non-spam
0	1	0	non-spam
0	0	0	non-spam

Then by counting, we can calculate MLEs as:

$$\theta_{spam} = P(C_{spam}) = 3/7$$

$$\theta_{secret|spam} = P(secret = 1|C_{spam}) = 2/3$$

$$\theta_{secret|non-spam} = P(secret = 1|C_{non-spam}) = 1/4$$

$$\theta_{sports|non-spam} = P(sports = 1|C_{non-spam}) = 2/4$$

$$\theta_{dollar|spam} = P(dollar = 1|C_{spam}) = 1/3$$

4. Support Vector Machine

(1) The support vectors are data points {7,18,20}:

$$(0.53, 0.77), (2.05, -0.62), (1.63, -0.91)$$

(2) $\mathbf{w} = [w_0, w_1, w_2]$

$$w_0 + 0.53w_1 + 0.77w_2 = 1$$

$$w_0 + 2.05w_1 - 0.62w_2 = -1$$

$$w_0 + 1.63w_1 - 0.91w_2 = -1$$

Solve the equations, we get $w_0 = 0.6687, w_1 = -0.5661, w_2 = 0.8198$

So $\mathbf{w} = [0.6687, -0.5661, 0.8198]$

(3) For calculating bias b:

$$x[\alpha_k \neq 0] = [(0.53, 0.77), (2.05, -0.62), (1.63, -0.91)]$$

$$w' = [-0.5661, 0.8198], y = [1, -1, -1], N_k = 3$$

$$b = \sum_{k:\alpha_k \neq 0} (y_k - w'x_k) / N_k = 0.6687$$

(4) The learned decision boundary function is:

$$f(x) = f((x_1, x_2)) = 0.6687 - 0.5661x_1 + 0.8198x_2$$

(5) For data point $x = (-1, 2)$:

$$f(x) = 0.6687 - 0.5661x_1 + 0.8198x_2 = 2.8744 > 1$$

So the predicted label for x is 1.