Homework8

Zhang Fan

December 22, 2016

1 Understanding Decision Tree

1.1

Caculate the original entropy of the information:

$$H_0 = \frac{5}{14}\log\frac{14}{5} + \frac{9}{14}\log\frac{14}{9} \approx 0.9403$$

If we branch with 'outlook',the entropy is:

$$H_o = \frac{5}{14} * (\frac{2}{5}\log\frac{5}{2} + \frac{3}{5}\log\frac{5}{3}) + \frac{5}{14} * (\frac{3}{5}\log\frac{5}{3} + \frac{2}{5}\log\frac{5}{2}) + \frac{4}{14} * (\frac{4}{4}\log\frac{4}{4}) \approx 0.6935$$

If we branch with 'temperature',the entropy is:

$$H_t = \frac{4}{14} * (\frac{2}{4}\log\frac{4}{2} + \frac{2}{4}\log\frac{4}{2}) + \frac{6}{14} * (\frac{2}{3}\log\frac{3}{2} + \frac{1}{3}\log\frac{3}{1}) + \frac{4}{14} * (\frac{3}{4}\log\frac{4}{3} + \frac{1}{4}\log\frac{4}{1}) \approx 0.9111$$

If we branch with 'humidity',the entropy is:

$$H_h = \frac{7}{14} * (\frac{3}{7}\log\frac{7}{3} + \frac{4}{7}\log\frac{7}{4}) + \frac{7}{14} * (\frac{6}{7}\log\frac{7}{6} + \frac{1}{7}\log\frac{7}{1}) \approx 0.7885$$

If we branch with 'windy',the entropy is:

$$H_w = \frac{8}{14} * (\frac{6}{8}\log\frac{8}{6} + \frac{2}{8}\log\frac{8}{2}) + \frac{6}{14} * (\frac{3}{6}\log\frac{6}{3} + \frac{3}{6}\log\frac{6}{3}) \approx 0.8922$$

So we have the 'Information Gain' of the 4 kinds of branch:

$$S_o = 0.2468$$

$$S_t = 0.0292$$

$$S_h = 0.1518$$

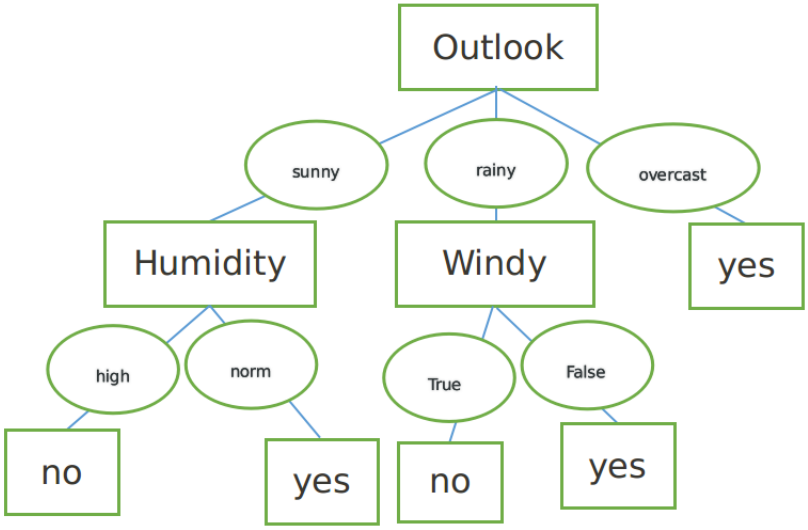
$$S_w = 0.0481$$

So,for the first step we choose 'outlook'.And getting 3 branch 'sunny','rainy' and 'overcast'. (I omit the process below) For the second step:

For the 'sunny' branch: If we branch with 'temperature',the entropy is: 0.4 If we branch with 'humidity',the entropy is: 0 If we branch with 'windy',the entropy is: 0.9510 So we choose 'humidity' for this banch.

For the 'rainy' branch: If we branch with 'temperature',the entropy is: 0.9510 If we branch with 'humidity',the entropy is: 0.9510 If we branch with 'windy',the entropy is: 0 So we choose 'windy' for this banch.

For the 'rainy' branch we can stop branching. So we stop branching here,and we have the decision tree:



1.2

From the subproblem1,we have the 'Info Gain': $S_o = 0.2468, S_t = 0.0292, S_h = 0.1518, S_w = 0.0481$.

Then we compute the 'Split Info'(with mark 'V':

$$V_o = \frac{5}{14}\log\frac{14}{5} + \frac{5}{14}\log\frac{14}{5} + \frac{4}{14}\log\frac{14}{4} \approx 1.5774$$

$$V_t = \frac{6}{14}\log\frac{14}{6} + \frac{4}{14}\log\frac{14}{4} + \frac{4}{14}\log\frac{14}{4} \approx 1.5567$$

$$V_h = \frac{7}{14}\log\frac{14}{7} + \frac{7}{14}\log\frac{14}{7} = 1$$

$$V_w = \frac{8}{14}\log\frac{14}{8} + \frac{6}{14}\log\frac{14}{6} \approx 0.9852$$

Caculate the 'Gain Ratio'(with mark 'G'):
$G_o = 0.1565, G_t = 0.0188, G_h = 0.1518, S_w = 0.0488$.
We also choose 'Outlook' for the first step.

1.3

Caculate the original entropy of the information:

$$H_0 = \frac{5}{14} * \frac{9}{14} + \frac{9}{14} * \frac{5}{14} \approx 0.4592$$

If we branch with 'outlook',the entropy is:

$$H_o = \frac{5}{14} * (\frac{2}{5} * \frac{3}{5} + \frac{3}{5} * \frac{2}{5}) + \frac{5}{14} * (\frac{3}{5} * \frac{2}{5} + \frac{2}{5} * \frac{3}{5}) + \frac{4}{14} * (\frac{4}{4} * (1 - \frac{4}{4})) \approx 0.1714$$

If we branch with 'temperature',the entropy is:

$$H_t = \frac{4}{14} * (2 * \frac{2}{4} * \frac{2}{4}) + \frac{6}{14} * (\frac{2}{3} * \frac{1}{3} + \frac{1}{3} * \frac{2}{3}) + \frac{4}{14} * (2 * \frac{3}{4} * \frac{1}{4}) \approx 0.4405$$

If we branch with 'humidity',the entropy is:

$$H_h = \frac{7}{14} * (2 * \frac{3}{7} * \frac{4}{7}) + \frac{7}{14} * (2 * \frac{6}{7} * \frac{1}{7}) \approx 0.3673$$

If we branch with 'windy',the entropy is:

$$H_w = \frac{8}{14} * (2 * \frac{6}{8} * \frac{2}{8}) + \frac{6}{14} * (2 * \frac{3}{6} \log \frac{3}{6}) \approx 0.4286$$

So we have the 'Information Gain' of the 4 kinds of branch:

$$S_o = 0.2778$$

$$S_t = 0.0187$$

$$S_h = 0.0919$$

$$S_w = 0.0306$$

So,for the first step we choose 'outlook'.And getting 3 branch 'sunny','rainy' and 'overcast'. (I omit the process below) For the second step:

For the 'sunny' branch: If we branch with 'temperature',the entropy is: 0.2 If we branch with 'humidity',the entropy is: 0 If we branch with 'windy',the entropy is: 0.4667 So we choose 'humidity' for this banch.

For the 'rainy' branch: If we branch with 'temperature',the entropy is: 0.4667 If we branch with 'humidity',the entropy is: 0.4667 If we branch with 'windy',the entropy is: 0 So we choose 'windy' for this banch.

For the 'rainy' branch we can stop branching. So we stop branching here,and we have the decision tree the same with the ID3.

2 Learning Ensemble Methods

2.1

1. Initialize $f_0(x) = 0$. 2. For $m = 1$ to M: we would like to reduce l over the data set through manipulating f . (1) First we compute the gradient of the loss function with respect to f

$$(g_m)_i = \frac{\partial}{\partial \hat{y_i}} (\frac{1}{2}(y_i - \hat{y_i})^2) = -(y_i - \hat{y_i})$$

and get g_m .

(2) Fit regression model to negative gradient $-g_m$:

$$p_m = arg \min_{p \in mathscrF} \sum_{i=1}^{N} [(y_i - \hat{y_i}) - p(x_i)]^2$$

(3) Choose fixed stepsize $v_m = v \in 0, 1]$, or take a line search strategy:

$$v_m = arg\min_{v>0} \sum_{i=1}^{N} \frac{1}{2}[(y_i - f_{m-1}(x_i) + vp_m(x_i)]^2$$

(4) Take this step:

$$f_m(x) = f_{m-1}(x) + v_m p_m(x).$$

3. Return f_M which is our final hypothesis function.
The distance equals to:

$$\|\bar{x} - x^*\|_2 = \left\| \frac{w^T \bar{x} + b}{w} \right\|_2$$

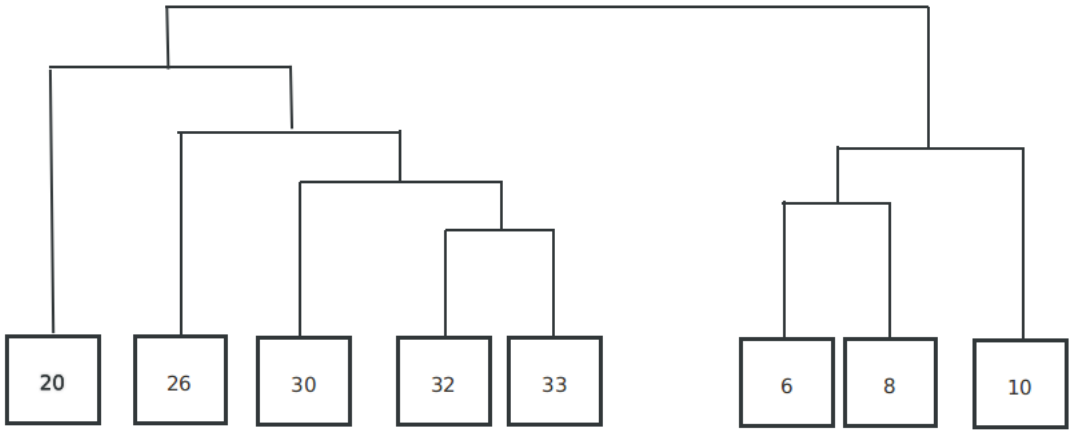
2.2

$\frac{2}{\|w\|}$

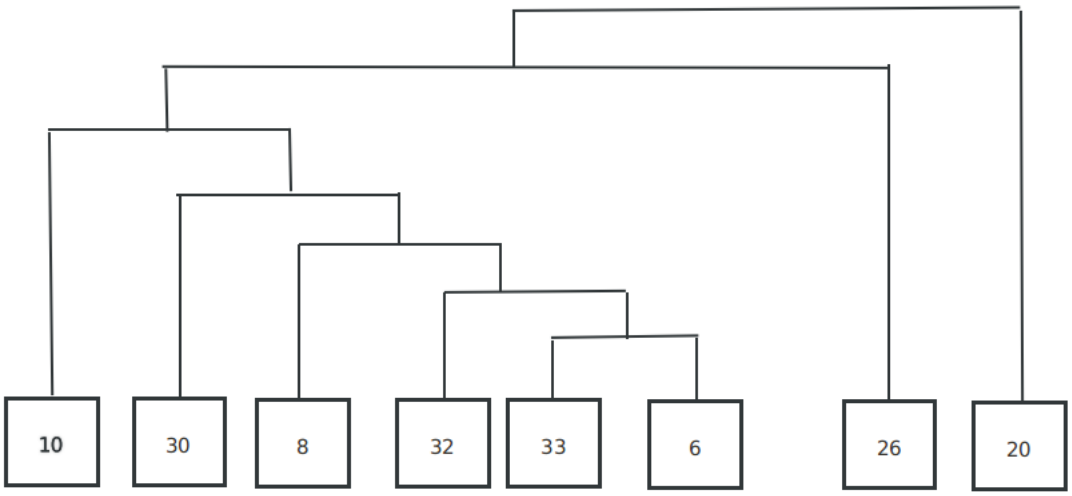
3 Clustering

3.1 Hierarchical clustering

(a) single linkage



(b) complete linkage



3.2

The first centroid is $\{6\}$,Then the second centroid is $\{33\}$, calculate the other points' distance to the centroid are:

$$\{2, 4, 13, 7, 3, 1\}$$

So we choose $\{20\}$ to be the third centroid.Then we will get three clusters:$\{6, 8, 10\}, \{20, 26\}, \{30, 32, 33\}$.

3.3

```
>> load fisheriris;
>> eva = evalclusters(meas,'kmeans','CalinskiHarabasz','KList',[1:6])

eva =

  CalinskiHarabaszEvaluation with properties:

    NumObservations: 150
        InspectedK: [1 2 3 4 5 6]
    CriterionValues: [1x6 double]
          OptimalK: 3
```

So $K = 3$.