CIS 419 – Homework 1

Devin Stein
devstein@seas.upenn.edu

September 25, 2016

Problem 1

a) At the root node for a decision tree in this domain, what are the information gains associated with the Outlook and Humidity attributes? (Use a threshold of 75 for humidity (i.e., assume a binary split: humidity ≤ 75 / humidity > 75). Be sure to show your computations.

$H(Root) = -\frac{9}{14}\log(\frac{9}{14}) - \frac{5}{14}\log(\frac{5}{14}) = .94$
$H(Humidity) = \frac{5}{14}(-\frac{4}{5}\log(\frac{4}{5}) - \frac{1}{5}\log(\frac{1}{5})) + \frac{9}{14}(-\frac{5}{9}\log(\frac{5}{9}) - \frac{4}{9}\log(\frac{4}{9}))= .895$
$H(Outlook) = \frac{5}{14}(-\frac{2}{5}\log(\frac{2}{5}) - \frac{3}{5}\log(\frac{3}{5})) + (\frac{4}{14}(0)) + \frac{5}{14}(-\frac{3}{5}\log(\frac{3}{5}) - \frac{2}{5}\log(\frac{2}{5}))= .694$

$$InformationGain(Humidity) = .94 - .895 = .045$$

$$InformationGain(Outlook) = .94 - .694 = .247$$

b) Again at the root node, what are the gain ratios associated with the Outlook and Humidity attributes (using the same threshold as in (a))? Be sure to show your computations.

$SplitInfo(Humidity) = -\frac{9}{14}\log(\frac{9}{14}) - \frac{5}{14}\log(\frac{5}{14}) = .94$
$SplitInfo(Outlook) = -\frac{5}{14}\log(\frac{5}{14}) - \frac{4}{14}\log(\frac{4}{14}) - \frac{5}{14}\log(\frac{5}{14}) = 1.58$

$$InformationRatio(Humidity) = \frac{.045}{.94} = .048$$

$$InformationRatio(Outlook) = \frac{.247}{1.58} = .156$$

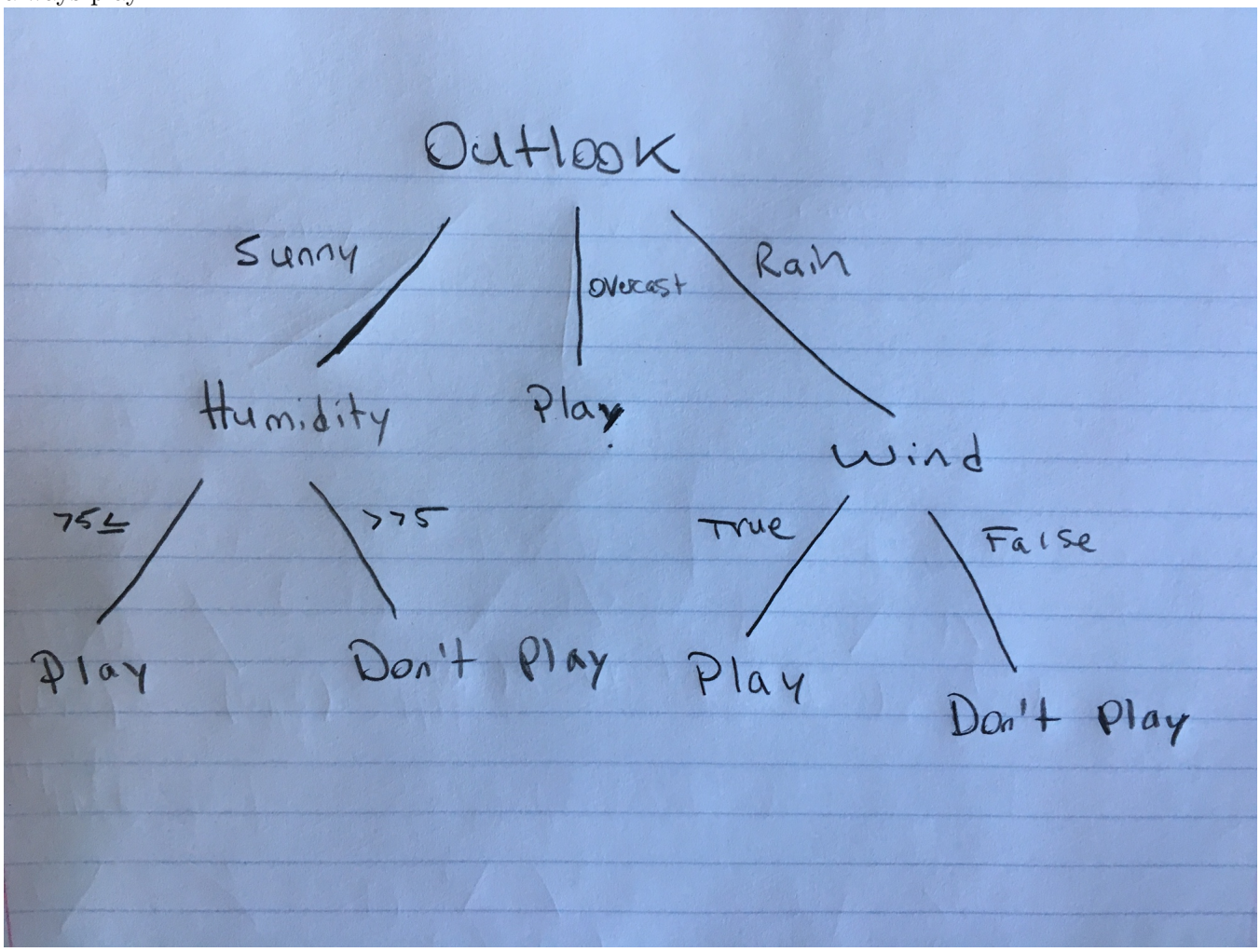
c) Draw the complete (unpruned) decision tree, showing the class predictions at the leaves.

$InformationRatio(Sunny, Humidity) > InformationRatio(Sunny, Wind) > InformationRatio(Sunny, Te$

$InformationRatio(Rain, Wind) > InformationRatio(Rain, Humidity) > InformationRatio(Rain, Temper$

Splitting on (Sunny, Humidity) gives the most information on a perfect Play/Don't Play split. Similarly, Splitting (Rain, Wind) gives is perfect. Lastly, only looking at (Overcast) tells you to

always play.


Outlook
Sunny
Overcast
Rain
Humidity
Play
Wind
75≤
>75
True
False
Play
Don't Play
Play
Don't Play

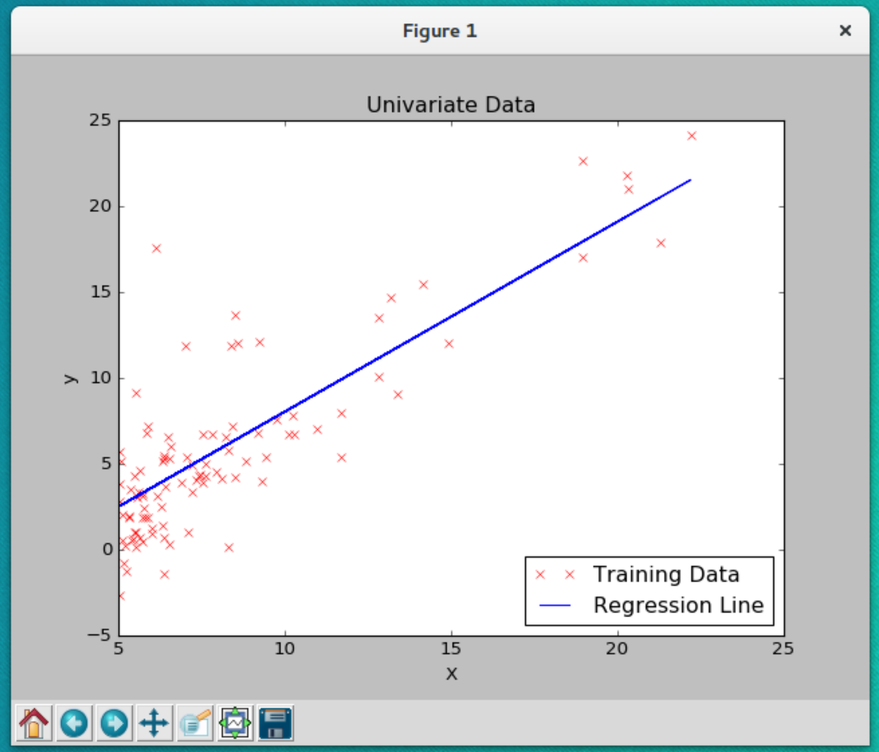
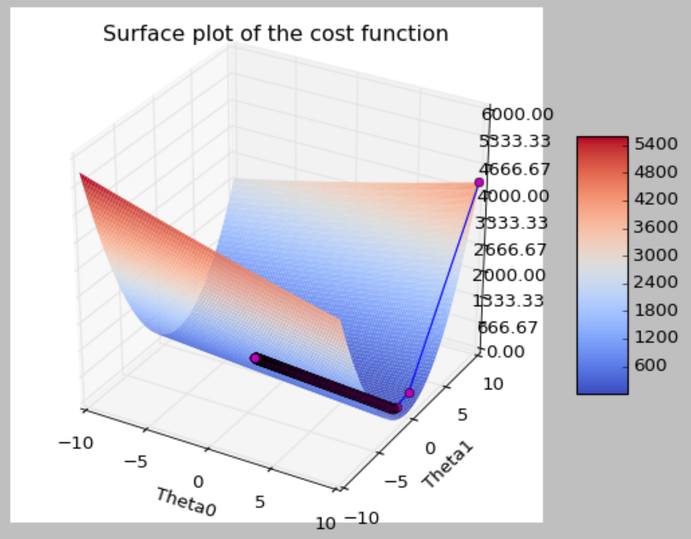


Figure 1
Univariate Data
25
20
15
10
5
0
−5
y
5
10
15
20
25
X
Training Data
Regression Line

Figure 1



Home Pictures

Recent
Home
Documents
Downloads
Music
Pictures
Videos
Trash
Network
Blade
Computer

Screenshot from 2016-09-26 16-48-58.png

"Screenshot from 2016-09-26 16-48-58.png" selected (2.8 MB)

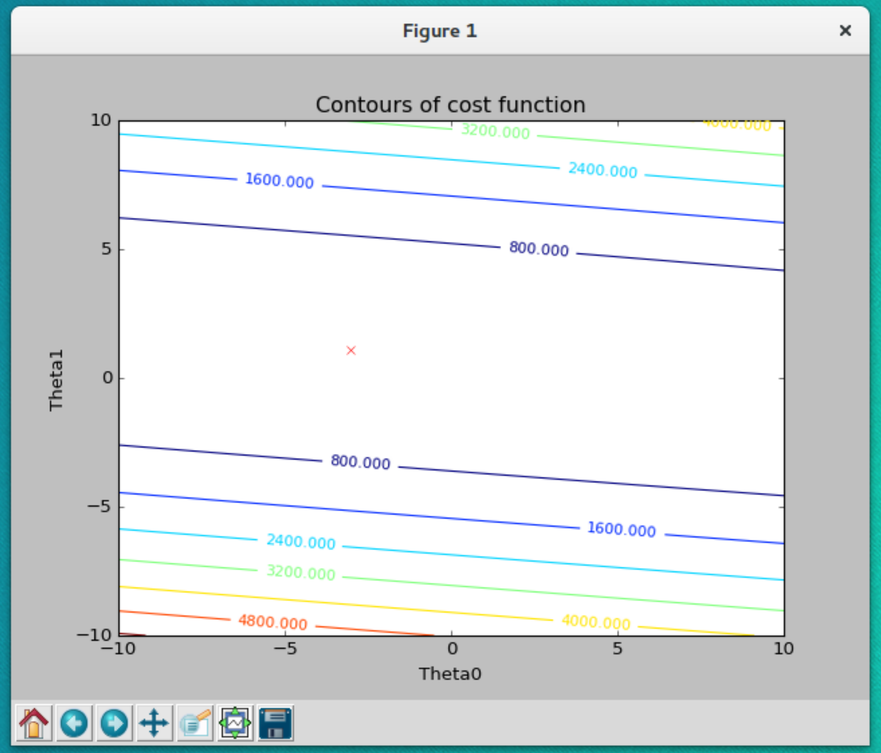
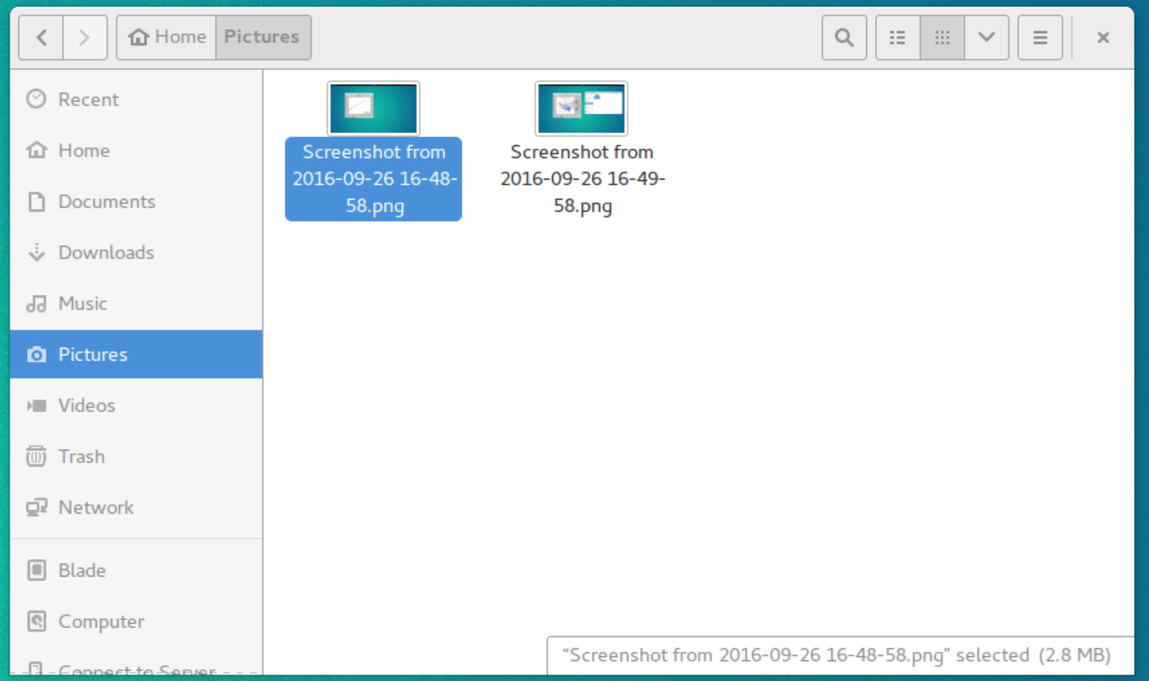


Figure 1
Contours of cost function
3200.000
2400.000
1600.000
800.000
800.000
1600.000
2400.000
3200.000
4800.000
4000.000
10
5
0
−5
−10
Theta1
−10
−5
0
5
10
Theta0


Home
Pictures
Recent
Home
Documents
Downloads
Music
Pictures
Videos
Trash
Network
Blade
Computer
Screenshot from 2016-09-26 16-48-58.png
Screenshot from 2016-09-26 16-49-58.png
"Screenshot from 2016-09-26 16-48-58.png" selected (2.8 MB)