CS 4980/6980: Introduction to Data Science © Spring 2018

Lecture 14: Decision Tree Example and Covariance

INSTRUCTOR: DANIEL L. PIMENTEL-ALARCÓN Scribed by: Nekhena Campbell & Benjamin Crom

This is preliminary work and has not been reviewed by instructor. If you have comments about typos, errors, notation inconsistencies, etc., please email the scribers.

14.1 Last Time

(a) Random Forests

(b) Decision Trees

14.2 Is my GF/BF cheating?

Age	23	20	33	23	25	32	27	21	25	27	26	29	47	$\begin{cases}1 & \text{if Age} > 28\\ 0 & \text{if Age} < 28\end{cases}$
Length	$1/2$	2	7	5	6	10	3	4	$3/2$	3	4	$1/3$	1	$\begin{cases}1 & \text{if Length} > 1\\ 0 & \text{if Length} < 1\end{cases}$
Children	0	0	2	1	0	3	0	0	0	0	1	0	6	$\begin{cases}1 & \text{if} > 1\\ 0 & \text{if} < 1\end{cases}$
Hours away/day	20	7	10	7	13	2	8	7	13	11	8	4	12	$\begin{cases}1 & \text{if} > 10\\ 0 & \text{if} < 10\end{cases}$
Trips/year	2	3	5	2	2	5	10	4	3	2	0	2	6	$\begin{cases}1 & \text{if} > 3\\ 0 & \text{if} < 3\end{cases}$
Previous offenses	2	1	0	2	0	0	2	3	0	4	0	0	13	$\begin{cases}1 & \text{if} > 1\\ 0 & \text{if} < 1\end{cases}$
Gender	M	M	M	F	F	F	F	M	M	F	F	M	M	
Is Cheating	1	0	1	1	0	1	0	0	0	0	1	1	1	

														$H(x_i)$	$\hat{P}_1$
x_1 — AGE	0	0	1	0	0	1	0	0	0	0	0	1	1	.89	4/13
x_2 — LENGTH	0	1	1	1	1	1	1	1	1	1	1	0	1	.619	11/13
x_3 — CHILDREN	0	0	1	1	0	1	0	0	0	0	1	0	1	.9612	5/13
x_4 — HOURS	1	0	0	0	1	0	0	0	1	1	0	0	1	.96	5/13
x_5 — TRIPS	0	0	1	0	0	1	1	1	0	0	0	0	1	.9612	5/13
x_6 — PREV	1	1	0	1	0	0	1	1	0	1	0	0	1	.995	7/13 $\leftarrow \star$
x_6 — GENDER	0	0	0	1	1	1	1	0	0	1	1	0	0	.995	6/13 $\leftarrow$

Compute the entropy for each of these features:

(a) Computer count number of 1 out of all variables

(b) Find Bernoulli random variable

$$\begin{aligned} H(x_3) &= \hat{P}_0 \log_2 \frac{1}{\hat{P}_0} + \hat{P}_1 \log_2 \frac{1}{\hat{P}_1} \\ &= \frac{8}{13} \log_2 \frac{13}{8} + \frac{5}{13} \log_2 \frac{13}{5} \\ &= 0.962 \end{aligned}$$

Thus, previous offenses and gender are both about equally informative.

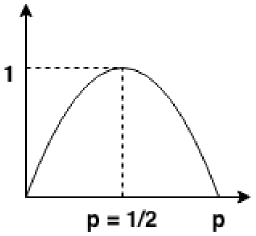
All <u>zeros</u> in previous offenses:

							$H(x_i)$	$\hat{P}_1$
$\star$ Age	1	0	1	0	0	1	1	3/6
Length	1	1	1	1	1	0	< 1	5/6
Children	1	0	1	0	1	0	1	3/6
$\rightarrow$ Hours	0	1	0	1	0	0	< 1	2/6
Trips	1	0	1	0	0	0	< 1	2/6
Gender	0	1	1	0	1	0	1	3/6
Cheating	1	0	1	0	1	1		

All <u>ones</u> in previous offenses:

								$H(x_i)$	$\hat{P}_1$
$\star$ Age	0	0	0	0	0	0	1		1/7
Length	0	1	1	1	1	1	1		6/7
Children	0	0	1	0	0	0	1		2/7
$\rightarrow$ Hours	1	0	0	0	0	1	1	0.98	3/7$\star$
Trips	0	0	0	1	1	0	1	0.98	3/7$\star$
Gender	0	0	1	1	0	1	0	0.98	3/7$\star$
Cheating	1	0	1	0	0	0	1		

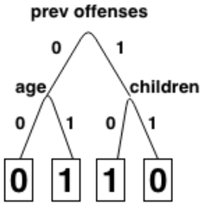
$\star$ = informative variable
The closer to $1/2$ has a large entropy; the maximum you can have is attained with $1/2$.



We use the information above to determine the next question in our decision tree by selecting the most informative variable. For our data, we can choose age, children or hours:

(a) Hours can be overfitting

(b) Age is a good choice but we may need to add another layer of questions.



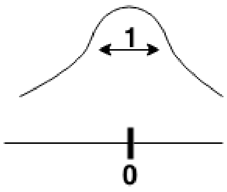
14.3 Covariance [matrices]

Main Goal: Find that variables are related

14.3.1 Random Vectors

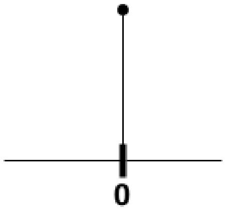
Let x be a random variable distributed gaussian$(0,1)$:

$$\text{random} \to x \sim \underset{y \sim \mathcal{N}(0,1)}{\underset{||}{\mathcal{N}(0,1)}}$$



Variance being 0 makes it deterministic and no longer random:

$$\begin{aligned}\text{deterministic} \rightarrow & x \sim \mathcal{N}(0,0)\\ & x = 0\end{aligned}$$



Which is better?
$y = 0.05$ vs. $y = 0$

if y and x are equal:
$y|x \sim \mathcal{N}(x, 0)$

if y and x are independent:
$y|x \sim \mathcal{N}(0, 1)$

$\exists$ a strong correlation between x, y:

$$\begin{aligned} x_1 &= 0.05 & y_1 &= 0.05\\ x_2 &= 0.07 & y_2 &= 0.07\\ x_3 &= -0.03 & y_3 &= -0.03\\ x_4 &= -0.09 & y_4 &= -0.09 \end{aligned}$$

not so strongly correlated:

$$\begin{aligned} x_1 &= 0.05 & y_1 &= -0.1\\ x_2 &= 0.07 & y_1 &= 0.09\\ x_3 &= -0.03 & y_1 &= 0.02\\ x_4 &= -0.09 & y_1 &= 0.05 \end{aligned}$$

$$\boldsymbol{Z} = \begin{bmatrix} x \\ y \end{bmatrix}$$

14.4 Wrap up

14.4.1 Today

(a) Example of Decision Trees

(b) Example of Random Forests

14.4.2 Next

(a) Correlation