Supervised Learning: an interactive example

Consider an application that attempts to predict if a user will *read* or *skip* an article posted on a threaded discussion board. The user's action can depend on the following **attributes** or **features** of articles on the discussion board:

- whether the *author* of the article is *known* or *unknown* to the user,
- whether the *thread* is *new* or a *follow up*,
- whether the article's *length* is *long* or *short*,
- whether the user reads the article at *home* or at *work*.

Try to predict, based only on your **prior knowledge** of threaded discussion boards, what the user's action will be (*read* or *skip*) for each of the following **test** examples:

example	*author*	*thread*	*length*	*where read*	*user's action*
t1	unknown	new	long	work	
t2	known	new	short	home	
t3	unknown	follow up	short	work	
t4	unknown	follow up	long	home	
t5	known	follow up	short	home	

DON'T turn the page over until you've finished your predictions!

Now, consider the following set of **training data**, showing a user's actual action for each of a set of actual articles:

example	*author*	*thread*	*length*	*where read*	*user's action*
e1	known	new	long	home	skips
e2	unknown	new	short	work	reads
e3	unknown	follow up	long	work	skips
e4	known	follow up	long	home	skips
e5	known	new	short	home	reads
e6	known	follow up	long	work	skips
e7	unknown	follow up	short	work	skips
e8	unknown	new	short	work	reads
e9	known	follow up	long	home	skips
e10	known	new	long	work	skips
e11	unknown	follow up	short	home	skips
e12	known	new	long	work	skips
e13	known	follow up	short	home	reads
e14	known	new	short	work	reads
e15	known	new	short	home	reads
e16	known	follow up	short	work	reads
e17	known	new	short	home	reads
e18	unknown	new	short	work	reads

Based on the above data (and on your prior knowledge), try to predict what the user's action will be for the same test examples you predicted on the first page:

example	*author*	*thread*	*length*	*where read*	*user's action*
t1	unknown	new	long	work	
t2	known	new	short	home	
t3	unknown	follow up	short	work	
t4	unknown	follow up	long	home	
t5	known	follow up	short	home	

Now you can consider the following questions:

- Did your predictions change after seeing the data? Why or why not?
- How would your predictions change if you only knew the length of each test article (*long* or *short*)?
- How would your predictions change if you only knew where the user read each test article (*home* or *work*)?
- Would your predictions change if the following two new training examples became available?

example	*author*	*thread*	*length*	*where read*	*user's action*
e19	known	follow up	short	work	skips
e20	known	follow up	short	home	skips

To learn a decision tree for this dataset, use information gain (IG) to greedily select nodes to expand.

- initial information content is 9 skips, 9 reads $I(E_0) = -2 * 0.5 \log(0.5) = 1$
- split on author:
  - 12 known: has 6 skips, 6 reads $I(E_1) = 1$
  - 6 unknown: has 3 skips, 3 reads $I(E_2) = 1$
  - so $I(Esplit) = \frac{12}{18} \times 1 + \frac{6}{18} \times 1 = 1$,
  - information gain is 0 (author tells you nothing about reads)
- split on length
  - 7 long: 7 skips, 0 reads $I(E_1) = 0$
  - 11 short: 2 skips, 9 reads
    $I(E_2) = -\frac{2}{11} \log(\frac{2}{11}) - \frac{9}{11} \log(\frac{9}{11}) = 0.447 + 0.23 = 0.684$
  - $I(Esplit) = \frac{7}{18} \times 0 + \frac{11}{18} \times 0.684 = 0.418$,
  - information gain is $1 - 0.418 = 0.582$
- can try other splits but this is the best.
- now split the data into two sets (7 long and 11 short) and start again

Here is the set of short articles:

example	*author*	*thread*	*length*	*where read*	*user's action*
e2	unknown	new	short	work	reads
e5	known	new	short	home	reads
e7	unknown	follow up	short	work	skips
e8	unknown	new	short	work	reads
e11	unknown	follow up	short	home	skips
e13	known	follow up	short	home	reads
e14	known	new	short	work	reads
e15	known	new	short	home	reads
e16	known	follow up	short	work	reads
e17	known	new	short	home	reads
e18	unknown	new	short	work	reads

- initial information is 9 reads, 2 skips
  $I(E_0) = -\frac{9}{11} \log(\frac{9}{11}) - \frac{2}{11} \log(\frac{2}{11}) = 0.684$
- split on author
  - 6 known: 6 reads, 0 skips $I(E_1) = 0$
  - 5 unknown: 3 reads, 2 skips
    $I(E_2) = -\frac{2}{5} \log(\frac{2}{5}) - \frac{3}{5} \log(\frac{3}{5}) = 0.97$

  - $I(Esplit) = \frac{6}{11} \times 0 + \frac{5}{11} \times 0.97 = 0.44$
  - information gain is $0.684 - 0.44 = 0.244$
- split on thread
  - 7 new: 7 reads, 0 skips: $I(E_1) = 0$
  - 4 follow up: 2 reads, 2 skips, $I(E_2) = 1$
  - $I(Esplit) = \frac{4}{11} \times 1 = 0.36$
  - information gain is $0.684 - 0.36 = 0.324$
- split on where read
  - 6 work: 5 reads, 1 skip
    $I(E_1) = -\frac{5}{6}\log(\frac{5}{6}) - \frac{1}{6}\log(\frac{1}{6}) = 0.65$
  - 5 home: 4 reads, 1 skip
    $I(E_2) = -\frac{4}{5}\log(\frac{4}{5}) - \frac{1}{5}\log(\frac{1}{5}) = 0.72$
  - $I(Esplit) = \frac{6}{11} \times 0.65 + \frac{5}{11} \times 0.72 = 0.68$
  - information gain is $0.684 - 0.68 = 0.004$
- so thread is best
- split on thread and continue, then come back and do the long articles