CS 446: Machine Learning
Discussion Session

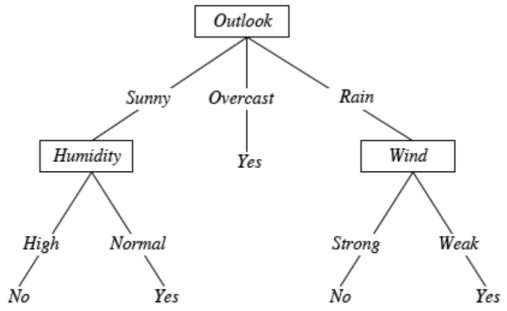
Daniel Khashabi

September 25, 2015

1 Decision Tree:

Consider the following training data and the following decision tree learned from this data using the algorithm described in class. The decision tree predicts the value of the boolean attribute `PlayTennis` using the values of the rest of the attributes.

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No



- Show that the choice of the `Wind` attribute at the second level of the tree is correct, by showing that its information gain is superior to the alternative choices.
- Add one new example to the above data set, so that the learned tree will contain additional nodes.

- Is it possible to add new examples to the above training set, which are consistent with the above tree, to produce a larger training set such that the algorithm will now learn a tree whose root node is not `Outlook`?. (We say an example is consistent with the above tree if the tree classifies the example correctly). Justify your answer by explaining informally why this is impossible, or explaining the new data you would add.

Solution:

- Let's first find the conditional entropies in the target node (when `Outlook = Rain`):

$H(\texttt{PlayTennis}|\texttt{Outlook} = \texttt{Rain}) = -3/5\log(3/5) - 2/5\log(2/5) = 0.29$
$H(\texttt{PlayTennis}|\texttt{Wind} = \texttt{Strong}, \texttt{Outlook} = \texttt{Rain}) = 0$
$H(\texttt{PlayTennis}|\texttt{Wind} = \texttt{Weak}, \texttt{Outlook} = \texttt{Rain}) = 0$
$H(\texttt{PlayTennis}|\texttt{Humidity} = \texttt{High}, \texttt{Outlook} = \texttt{Rain}) = -1/2\log(1/2) - 1/2\log(1/2))$
$H(\texttt{PlayTennis}|\texttt{Humidity} = \texttt{Low}, \texttt{Outlook} = \texttt{Rain}) = -1/3\log(1/3) - 2/3\log(2/3)$
$H(\texttt{PlayTennis}|\texttt{Temperature} = \texttt{Hot}, \texttt{Outlook} = \texttt{Rain}) = 0$
$H(\texttt{PlayTennis}|\texttt{Temperature} = \texttt{Mild}, \texttt{Outlook} = \texttt{Rain}) = -2/3\log(2/3) - 1/3\log(1/3)$
$H(\texttt{PlayTennis}|\texttt{Temperature} = \texttt{Cool}, \texttt{Outlook} = \texttt{Rain}) = -1/2\log(1/2) - 1/2\log(1/2)$

Now we fine the information gains. The information gain, when the branching is done with `Wind`:

$$\begin{aligned} I(\texttt{PlayTennis}|\texttt{Outlook} = \texttt{Rain}; \texttt{Wind}) &= H(\texttt{PlayTennis}|\texttt{Outlook} = \texttt{Rain}) \\ &- P(\texttt{Wind} = \texttt{Strong}|\texttt{Outlook} = \texttt{Rain})H(\texttt{PlayTennis}|\texttt{Wind} = \texttt{Strong}, \texttt{Outlook} = \texttt{Rain}) \\ &+ P(\texttt{Wind} = \texttt{Weak}|\texttt{Outlook} = \texttt{Rain})H(\texttt{PlayTennis}|\texttt{Wind} = \texttt{Weak}, \texttt{Outlook} = \texttt{Rain}) \\ &= 0.29 - (2/50 + 2/50) = 0.29 \end{aligned}$$

The information gain, when the branching is done with `Humidity`:

$$\begin{aligned} I(\texttt{PlayTennis}|\texttt{Outlook} = \texttt{Rain}; \texttt{Humidity}) &= H(\texttt{PlayTennis}|\texttt{Outlook} = \texttt{Rain}) \\ &- P(\texttt{Humidity} = \texttt{Hight}|\texttt{Outlook} = \texttt{Rain})H(\texttt{PlayTennis}|\texttt{Humidity} = \texttt{Hight}, \texttt{Outlook} = \texttt{Rain}) \\ &+ P(\texttt{Humidity} = \texttt{Low}|\texttt{Outlook} = \texttt{Rain})H(\texttt{PlayTennis}|\texttt{Humidity} = \texttt{Low}, \texttt{Outlook} = \texttt{Rain}) \\ &= 0.29 - [2/5 * (-1/2 * \log(1/2) - 1/2 * \log(1/2)) + 3/5 * (-1/3\log(1/3) - 2/3\log(2/3))] \\ &= 0.29 - 0.951 = -0.661 \end{aligned}$$

The information gain, when the branching is done with `Temperature`:

$$\begin{aligned} I(\texttt{PlayTennis}|\texttt{Outlook} = \texttt{Rain}; \texttt{Temperature}) &= H(\texttt{PlayTennis}|\texttt{Outlook} = \texttt{Rain}) \\ &- P(\texttt{Temperature} = \texttt{Hot}|\texttt{Outlook} = \texttt{Rain})H(\texttt{PlayTennis}|\texttt{Temperature} = \texttt{Hot}, \texttt{Outlook} = \texttt{Rain}) \\ &- P(\texttt{Temperature} = \texttt{Mild}|\texttt{Outlook} = \texttt{Rain})H(\texttt{PlayTennis}|\texttt{Temperature} = \texttt{Mild}, \texttt{Outlook} = \texttt{Rain}) \\ &- P(\texttt{Temperature} = \texttt{Low}|\texttt{Outlook} = \texttt{Rain})H(\texttt{PlayTennis}|\texttt{Temperature} = \texttt{Low}, \texttt{Outlook} = \texttt{Rain}) \\ &= 0.29 - [0 + 3/5 * (-2/3log(2/3) - 1/3log(1/3)) + 2/5 * (-1/2log(1/2) - 1/2log(1/2))] \\ &= 0.29 - 0.951 = -0.661 \end{aligned}$$

- Add a new element, to one of the the leaves, but with label different than the label of the that node. For example, to the node which corresponding to `Outlook = Rain, Wind = Weak` , add a new instance with label `No`. Here is an example instance:
`Outlook = Rain, Temperature = Hot, Humidity = Hight, Wind = Weak, PlayTennis = No` .

- Say we intentionally want to make `Humidity` attribute our new root node. We add a lot of new training instances (say 1000 of them), which have the same label for the current root attribute (say al have `Outlook = Sunny`), but distributed half-half based on the labels of the attribute `Humidity`. On this new dataset, the information gain caused by splitting on `Humidity` will now be much higher, for large enough number of new instances.

2 Maximum Likelihood vs Squared Loss:

Consider a series of observations of the form $(x_1, y_1), (x_2, y_2), ...(x_n, y_n)$. Given these observations (training data), our aim is to find a function f of the form $f(x) = w^\top x$ such that $f(x)$ is a good estimate of y for this dataset. We saw this problem in class; this is just the problem of linear regression. In class, we found the w that minimized the square error over the data, i.e. $E = \sum_i (y_i - w^\top x_i)^2$ and considered that to be a good estimate. We will now see why that might be considered a good estimate. Suppose the likelihood of the observation y_i is a Gaussian distribution with mean $w^\top x_i$ and variance σ, i.e.:

$$p(y_i|w, x_i) = \frac{1}{2\pi\sqrt{\sigma}} \exp\left(-(y_i - w^\top x_i)^2/2\sigma^2\right)$$

- Derive the expression for the conditional likelihood of the data assuming that each observation was independently generated. ie. write the expression for

$$P(y_1, y_2, ..., y_n|w, x_1, x_2, ...x_n) = \prod_i p(y_i|w, x_i)$$

- Now, using the expression for the conditional likelihood of the data, write down the expression for the conditional log-likelihood, i.e.

$$L = \log P(y_1, y_2, ..., y_n|w, x_1, x_2, ...x_n) = \sum_i \log p(y_i|w, x_i)$$

- Show that computing the Maximum Likelihood Estimate of w is equivalent to finding the w that minimizes the square error. [1].

Solution

[1] Maximum Likelihood Estimate of a parameter, which in this case w, is the value of the parameter that maximizes the likelihood of the data

- $$\begin{aligned} P(y_1, y_2, ..., y_n | w, x_1, x_2, ...x_n) &= \prod_i p(y_i | w, x_i) \\ &= \prod_i \frac{1}{2\pi\sqrt{\sigma}} \exp\left(-(y_i - w^\top x_i)^2 / 2\sigma^2\right) \\ &= (\frac{1}{2\pi\sqrt{\sigma}})^n \exp\left(-\frac{1}{2\sigma^2} \sum_i (y_i - w^\top x_i)^2\right) \end{aligned}$$

- The log likelihood is:
$$L = n \log(\frac{1}{2\pi\sqrt{\sigma}}) - \frac{1}{2\sigma^2} \sum_i (y_i - w^\top x_i)^2$$

- We want to maximize L with respect to w. The first term $n \log(\frac{1}{2\pi\sqrt{\sigma}})$ is a constant. But we can see that in the second term we have the squared loss. Since there is a also a negative sign, finding the maximum likelihood estimate for w here is equivalent to minimizing the quadratic loss.