Advanced Machine Learning - CS5824/ECE5424 Fall 2019

Department of Electrical and Computer Engineering, Department of Computer Science

Homework 1

Made by:
Amanda Redhouse and Isaiah Herrera

Contents

1 Problem 1

Show what the recursive decision tree learning algorithm would choose for the first split of the following dataset:

ID	**X1**	**X2**	**X3**	**X4**	**Y**
1	0	0	0	0	0
2	0	0	0	1	0
3	0	0	1	0	0
4	0	0	1	1	0
5	0	1	0	0	0
6	0	1	0	1	1
7	0	1	1	0	1
8	0	1	1	1	1
9	1	0	0	0	1
10	1	0	0	1	1

Assume that the criterion for deciding the best split is entropy reduction (i.e., information gain). If there are any ties, choose the first feature to split on tied for the best score. Show your calculations in your response.

1.1 Calculations

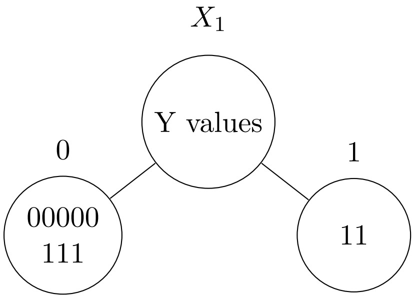
Information Gain:
$infoGain(j) = H(Y) - H(Y|X_j)$

$= -\sum_y Pr(Y = y)log_2(Pr(Y = y)) +$
$\sum_{X_j} Pr(X_j = x_j) \sum_y Pr(Y = y|X_j = x_j)log_2(Pr(Y = y|X_j = x_j))$

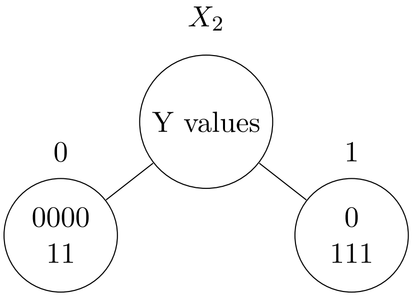
$- \sum_y Pr(Y = Y) \log_2(Pr(Y = y))$

$= -(0.5 * \log_2(0.5) + 0.5 * \log_2(0.5)) = -(-0.5 + -0.5) = 1$

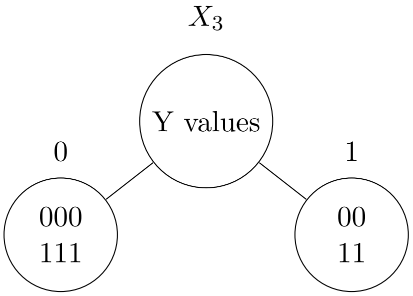
Information Gain for each feature:



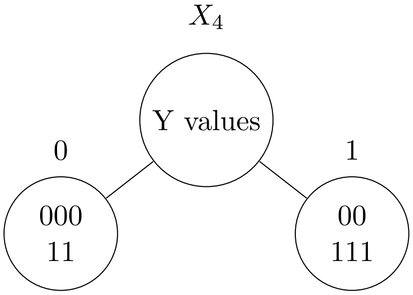
$X_1 = 0: \ \frac{8}{10}[\frac{3}{8} * \log_2(\frac{3}{8}) + \frac{5}{8} * \log_2(\frac{5}{8})] = -0.764$
$X_1 = 1: \ \frac{2}{10}[\frac{2}{2} * \log_2(\frac{2}{2}) + 0 * \log_2(0)] = 0$
$infoGain(X_1) = 1 + (-0.764) + 0 = 0.236$



$X_2 = 0: \ \frac{6}{10}[\frac{4}{6} * \log_2(\frac{4}{6}) + \frac{2}{6} * \log_2(\frac{2}{6})] = -0.551$
$X_2 = 1: \ \frac{4}{10}[\frac{1}{4} * \log_2(\frac{1}{4}) + \frac{3}{4} * \log_2(\frac{3}{4})] = -0.325$
$infoGain(X_2) = 1 + (-0.551) + (-0.325) = 0.124$



$X_3 = 0: \ \frac{6}{10}[\frac{3}{6} * \log_2(\frac{3}{6}) + \frac{3}{6} * \log_2(\frac{3}{6})] = -0.6$
$X_3 = 1: \ \frac{4}{10}[\frac{2}{4} * \log_2(\frac{2}{4}) + \frac{2}{4} * \log_2(\frac{2}{4})] = -0.4$
$infoGain(X_3) = 1 + (-0.6) + (-0.4) = 0$



$X_4 = 0 : \ \frac{5}{10}[\frac{3}{5} * \log_2(\frac{5}{5}) + \frac{2}{5} * \log_2(\frac{2}{5})] = -0.485$
$X_4 = 1 : \ \frac{5}{10}[\frac{2}{5} * \log_2(\frac{2}{5}) + \frac{3}{5} * \log_2(\frac{3}{5})] = -0.485$
$infoGain(X_4) = 1 + (-0.485) + (-0.485) = 0.03$

Therefore, the recursive decision tree algorithm would choose to first split on the feature X_1 because the information gain is the highest of the four features.

2 Problem 2

A Bernoulli distribution has the following likelihood function for a data set $\mathcal{D}$:

$$p(\mathcal{D}|\theta) = \theta^{N_1}(1-\theta)^{N_0} \tag{1}$$

where N_1 is the number of instances in data set $\mathcal{D}$ that have value 1 and N_0 is the number in D that have value 0. The maximum likelihood estimate is

$$\theta = \frac{N_1}{(N_1 + N_0)} \tag{2}$$

2.1 Part (a)

Derive the maximum likelihood estimate above by solving for the maximum of the likelihood. I.e., show the mathematics that get from Equation (1) to Equation (2).

2.1.1 Calculations

$p(\mathcal{D}|\theta) = \theta^{N_1}(1-\theta)^{N_0}$

$\ln p(\mathcal{D}|\theta) = \ln(\theta^{N_1}(1-\theta)^{N_0})$

$\ln p(\mathcal{D}|\theta) = \ln(\theta^{N_1}) + \ln((1-\theta)^{N_0})$

$\ln p(\mathcal{D}|\theta) = N_1 \ln(\theta) + N_0 \ln(1-\theta)$

Take derivative and set equal to 0 to find maximum

$0 = \frac{N_1}{\theta} + \frac{N_0}{\theta - 1}$

$N_1(\theta - 1) = -N_0\theta$

$\theta(N_1 + N_0) = N_1$

$\theta = \frac{N_1}{N_1+N_0}$ Equation 2

2.2 Part (b)

Suppose we now want to maximize a posterior likelihood

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})} \tag{3}$$

where we use the Bernoulli likelihood and a (slight variant[1] of a) symmetric Beta prior over the Bernoulli parameter

$$p(\theta) \propto \theta^\alpha(1-\theta)^\alpha \tag{4}$$

Derive the maximum posterior mean estimate.

2.2.1 Calculations

Replace $p(\mathcal{D}|\theta)$ and $p(\theta)$ with equations (1) and (4):

$p(\theta|\mathcal{D}) = \frac{\theta^{N_1}(1-\theta)^{N_0} * \theta^\alpha(1-\theta)^\alpha}{p(\mathcal{D})}$

$\ln p(\theta|\mathcal{D}) = \ln \frac{\theta^{N_1}(1-\theta)^{N_0} * \theta^\alpha(1-\theta)^\alpha}{p(\mathcal{D})}$

$\ln p(\theta|\mathcal{D}) = N_1 \ln(\theta) + N_0 \ln(1-\theta) + \alpha \ln(\theta) + \alpha \ln(1-\theta) - \ln(p(\mathcal{D}))$

Take derivative w.r.t. θ and set equal to 0 to find maximum: (Note: $p(\mathcal{D})$ does not contain θ, therefore the result can be treated as a constant (which we choose to ignore in this case))

$0 = \frac{N_1}{\theta} + \frac{N_0}{\theta-1} + \frac{\alpha}{\theta} + \frac{\alpha}{\theta-1}$

$\frac{1}{\theta}(N_1 + \alpha) = \frac{-1}{\theta-1}(N_0 + \alpha)$

$(\theta - 1)(N_1 + \alpha) = -(N_0 + \alpha)(\theta)$

$N_1\theta + \alpha\theta - N_1 - \alpha = -N_0\theta - \theta\alpha$

$N_1\theta + \alpha\theta + N_0\theta + \alpha\theta = N_1 + \alpha$

$\theta = \frac{N_1+\alpha}{N_1+2\alpha+N_0}$ maximum posterior mean estimate