HW 4 - Classification

Martin Kraus

March 6, 2020

1 Theory

1.1 Consider the following set of training examples for an unknown target function: $(x1, x2) \rightarrow y$:

Y	x_1	x_2	Count
+	T	T	3
+	T	F	4
+	F	T	4
+	F	F	1
−	T	T	0
−	T	F	1
−	F	T	3
+	F	F	5

1.1.1 What is the sample entropy, $H(Y)$ from this training data (using log base 2)

$$\begin{aligned} H(Y) &= H(P(+), P(-)) \\ &= -\frac{12}{21}\log_2\left(\frac{12}{21}\right) - \frac{9}{21}\log_2\left(\frac{9}{21}\right) \\ &= 0.985 \end{aligned}$$

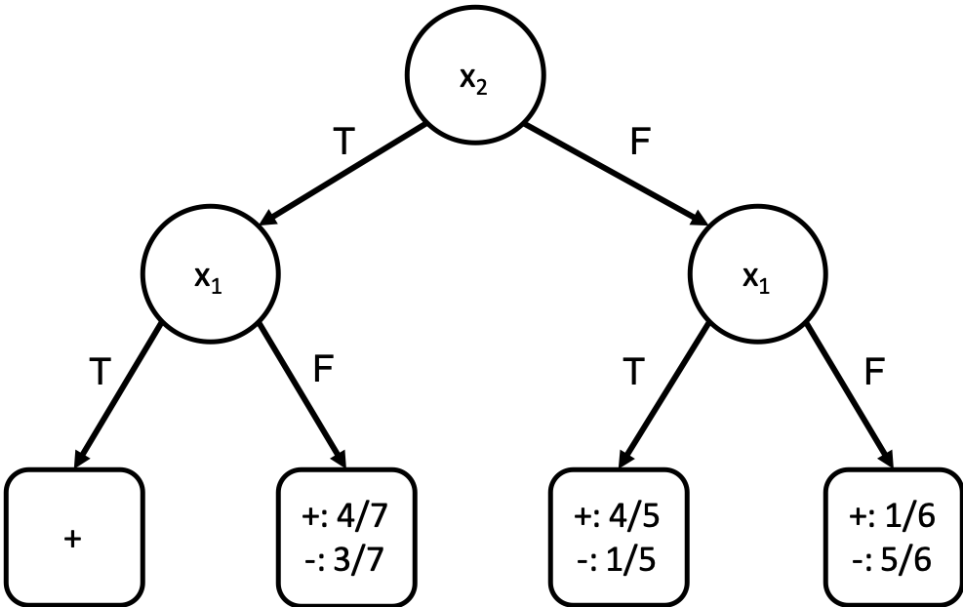
1.1.2 What are the weighted average entropies of the class labels of the subsets created by variables x_1 and x_2

	x_1	x_2
p_T	7	7
n_T	1	3
p_F	5	5
n_F	8	6

$$\begin{aligned}\mathbb{E}(H(x_1)) &= \frac{8}{21}\left(-\frac{7}{8}\log_2\left(\frac{7}{8}\right)-\frac{1}{8}\log_2\left(\frac{1}{8}\right)\right)+\frac{13}{21}\left(-\frac{5}{13}\log_2\left(\frac{5}{13}\right)-\frac{8}{13}\log_2\left(\frac{8}{13}\right)\right)\\ &= 0.802\end{aligned}$$

$$\begin{aligned}\mathbb{E}(H(x_2)) &= \frac{10}{21}\left(-\frac{7}{10}\log_2\left(\frac{7}{10}\right)-\frac{3}{10}\log_2\left(\frac{3}{10}\right)\right)+\frac{11}{21}\left(-\frac{5}{11}\log_2\left(\frac{5}{11}\right)-\frac{6}{11}\log_2\left(\frac{6}{11}\right)\right)\\ &= \mathbf{0.940}\end{aligned}$$

1.1.3 Draw the decision tree that would be learned by the ID3 algorithm without pruning from this training data



1.2 We decided that maybe we can use the number of characters and the average word length an essay to determine if the student should get an A in a class or not. Below are five samples of this data:

of Chars	Average Word Length	Give an A
216	5.68	Yes
69	4.78	Yes
302	2.31	No
60	3.16	Yes
393	4.20	No

1.2.1 What are the class priors, $P(A = Yes)$, $P(A = No)$?

$$P(A = Yes) = \frac{3}{5}$$
$$P(A = No) = \frac{2}{5}$$

1.2.2 Find the parameters of the Gaussians necessary to do Gaussian Naive Bayes classification on this decision to give an A or not

$$mean(chars) = 208$$
$$std(chars) = 145.2154$$

$$mean(avgWrdLen) = 4.026$$
$$std(avgWrdLen) = 1.3256$$

of Chars	Average Word Length	Give an A
0.0551	1.2477	Yes
-0.9572	0.5688	Yes
0.6473	-1.2945	No
-1.0192	-0.6533	Yes
1.2740	-0.1313	No

Table 1: Standardized data

# chars : giveA=Yes	# chars : giveA=No
$\mu = -0.6404$ $\sigma^2 = 0.3638$	$\mu = 0.9606$ $\sigma^2 = 0.1963$
avgWrdLen : giveA=Yes	avgWrdLen : giveA=No
$\mu = 0.3877$ $\sigma^2 = 0.9280$	$\mu = -0.5816$ $\sigma^2 = 1.0164$

Table 2: Gaussian Paramters for each feature and class

1.2.3 Using your response from the prior question, determine if an essay with 242 characters and an average word length of 4.56 should get an A or not.

Gaussian Calculations for # chars:

$$std(242) = 0.2341$$

$$\begin{aligned} gaus_{\#char:yes}(0.2341) &= \frac{1}{\sqrt{0.3638 \times 2\pi}} e^{\frac{-(0.2341-0.6404)^2}{2(0.3638)}} \\ &= 0.2347 \end{aligned}$$

$$\begin{aligned} gaus_{\#char:no}(0.2341) &= \frac{1}{\sqrt{0.1963 \times 2\pi}} e^{\frac{-(0.2341-0.9606)^2}{2(0.1963)}} \\ &= 0.2312 \end{aligned}$$

Gaussian Calculations for avgWrdLen:

$$std(4.56) = 0.4028$$

$$\begin{aligned} gaus_{avgWrdLen:yes}(0.4028) &= \frac{1}{\sqrt{0.9280 \times 2\pi}} e^{\frac{-(0.4028-0.3877)^2}{2(0.9280)}} \\ &= 0.2457 \end{aligned}$$

$$\begin{aligned} gaus_{avgWrdLen:no}(00.4028) &= \frac{1}{\sqrt{1.0164 \times 2\pi}} e^{\frac{-(0.4028+0.5816)^2}{2(1.0164)}} \\ &= 0.4141 \end{aligned}$$

Classification:

$$\begin{aligned} P(yes) &= \frac{3}{5}(0.2312)(0.4141) \\ &= \mathbf{0.057} \\ P(no) &= \frac{2}{4}(0.2347)(0.2457) \\ &= 0.023 \end{aligned}$$

Therefore classification Give and A = **Yes**

2 Naive Bayes Classifier

Precision:	0.64576
Recall:	0.96360
F-Measure:	0.77330
Accuracy:	0.78748

3 Logistic Regression

Precision:	0.91917
Recall:	0.84749
F-Measure:	0.88188
Accuracy:	0.91460