Mark Klobukov

CS 383: Homework 4

Professor Matthew Burlick

2/19/2018

Part I: Theory

1) Training examples for an unknown target function:

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

The dataset has two classes: + and – . The total number of data with class + is (3+4+4+1) = 12. Total number of data with class – is (0+1+3+5) = 9

(a) What is the sample entropy from this training data?

$$H\big(P(v_1),\dots,P(v_n)\big) = \sum_{i=1}^{n}(-P(v_i)\log_2 P(v_i))$$

$$P(+) = \frac{12}{21}\ , \qquad P(-) = \frac{9}{21}$$

$$H(Y) = -\frac{12}{21}\log_2\frac{12}{21} - \frac{9}{21}\log_2\frac{9}{21} = 0.9852$$

b) What are the IGs for branching on variables x_1 and x_2?

x_1: $p_T = (3+4) = 7,\ n_T = (0+1) = 1$

$p_F = (4+1) = 5, n_F = (3+5) = 8$

$$remainder(x_1) = \frac{8}{21}\left(-\frac{7}{8}\log_2\frac{7}{8} - \frac{1}{8}\log_2\frac{1}{8}\right) + \frac{13}{21}\left(-\frac{5}{13}\log_2\frac{5}{13} - \frac{8}{13}\log_2\frac{8}{13}\right) = 0.80212$$

$$IG(x_1) = H(Y) - remainder(x_1) = 0.9852 - 0.80212 = 0.18308$$

x_2: $p_T = (3+4) = 7,\ n_T = (0+3) = 3$

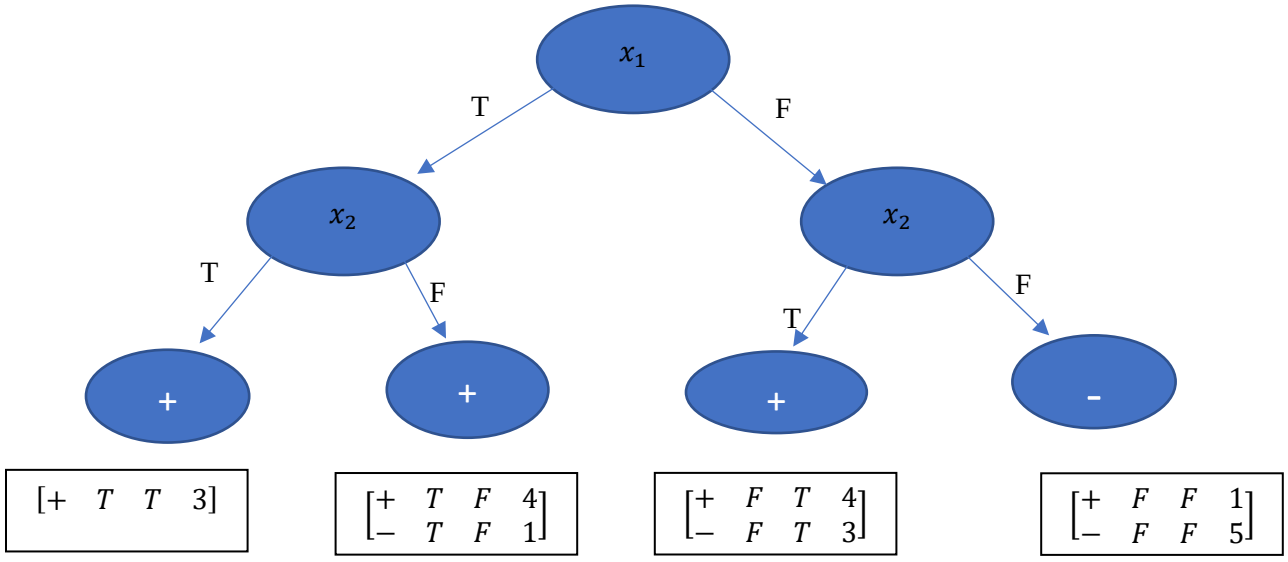
$p_F = (4+1) = 5,\ n_F = (1+5) = 6$

$$remainder(x_2) = \frac{10}{21}\left(-\frac{7}{10}\log_2\frac{7}{10} - \frac{3}{10}\log_2\frac{3}{10}\right) + \frac{11}{21}\left(-\frac{5}{11}\log_2\frac{5}{11} - \frac{6}{11}\log_2\frac{6}{11}\right) = 0.9403$$

$$IG(x_2) = H(Y) - remainder(x_2) = 0.9852 - 0.9403 = 0.0449$$

(c) Draw the decision tree that would be learned by the ID3 algorithm without pruning from this training data.

Feature x_1 has a higher information gain, so place it at the root of the tree. Then split each branch further based on feature x_2. After that, no more splits are possible. Perfect split is not achieved, but we can take the mode class in a given leaf and make that the final decision.



2) Five data samples:

# 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	4.2	No

(a) What are the class priors?

Class priors are the probabilities of encountering the class in a dataset.

$$P(A) = \frac{3}{5}, \qquad P(< A) = \frac{2}{5}$$

(b) Find parameters of the Gaussians.

Standardize data so that there is no unfair bias toward features of different scales. First column of the following matrix represents standardized number of characters, second column is the average word length (also standardized). In the third column, 1/0 correspond to Yes/No.

$$\begin{bmatrix} 0.0551 & 1.2477 & 1 \\ -0.9572 & 0.5688 & 1 \\ 0.6473 & -1.2945 & 0 \\ -1.0192 & -0.6533 & 1 \\ 1.2740 & 0.1313 & 0 \end{bmatrix}$$

Model for "Yes" observations:

characters:

$\mu = -0.6404,\ \sigma = 0.6031 \rightarrow \mathrm{N}(\mu, \sigma) = \mathrm{N}(-0.6404, 0.6031)$

Avg. word length:

$\mu = 0.3877,\ \sigma = 0.9633 \rightarrow \mathrm{N}(\mu, \sigma) = \mathrm{N}(0.3877, 0.9633)$

Model for "No" observations:

#characters:

$\mu = 0.9606,\ \sigma = 0.4431 \rightarrow \mathrm{N}(\mu, \sigma) = \mathrm{N}(0.9606, 0.4431)$

Avg. word length:

$\mu = -0.5816,\ \sigma = 1.0082 \rightarrow \mathrm{N}(\mu, \sigma) = \mathrm{N}(-0.5816, 1.0082)$

(c) Determine if an essay with 242 characters and average word length of 4.56 should get an A.

Standardize 242 characters : $\frac{242-208}{145.2154} = 0.2341\ characters$ (208 is mean, 145.2154 is std)

Standardize 4.56 word length: $\frac{4.56-4.0260}{1.3256} = 0.40283\ letters$ (4.026 is mean, 1.3256 is std)

Compute $P\ (grade = A \mid f = [242\ chars, 4.56\ letters]\ )$

By naïve independence assumption (and discarding the denominator):

$$P\ (grade = A \mid f = [242\ chars, 4.56\ letters]) \rightarrow$$
$$P(grade = A) \times P(chars = 242 \mid grade = A) \times P(avg = 4.56 \mid grade = A)$$

Let *p* stand for the approximation of P with Gaussian probability density function. normpdf() is a MATLAB-syntax Gaussian PDF function with three arguments: value x, mean, std.

$$\rightarrow P(A) \times p(A \mid \mathrm{N}\left(\mu_{A,chars}, \sigma_{A,chars}\right)\ ) \times p(\ A \mid \mathrm{N}\left(\mu_{A,avg}, \sigma_{A,avg}\right)\ ) =$$
$$= \frac{3}{5} \times normpdf(0.2341\ chars, -0.6404, 0.6031)$$
$$\times normpdf(0.40283\ letters, 0.3877, 0.9633) =$$
$$= \ 0.0574$$

This is not a probability. That is, the number above cannot be interpreted as 5.7 % chance of getting an A. It is necessary to perform the same computation for a grade below an A and then compare the two numbers.

$$P\ (\ grade < A) \mid f = [242\ chars, 4.56\ letters]\ )$$

By naïve independence assumption (and discarding the denominator):

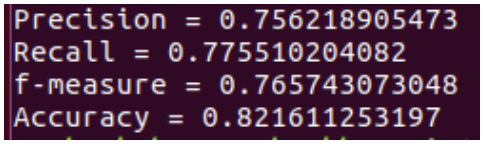
$$\rightarrow P(\ < A\ ) \times p(< A \mid \mathrm{N}\left(\mu_{<A,chars}, \sigma_{<A,chars}\right)\ ) \times p\ (< A\ \Big|\ \mathrm{N}\left(\mu_{<A,avg}, \sigma_{<A,avg}\right)\Big) =$$
$$= \frac{2}{5} \times normpdf(0.2341\ chars, 0.9606, 0.4431)$$
$$\times normpdf(0.40283\ letters, -0.5816, 1.0082) = 0.0231$$

The value of likelihood obtained for class A = 0.0574; for class <A = 0.0231

The value for class A is higher, therefore it is about twice as likely that the essay will get an A than that the essay will get below an A. Answer: yes, by our classifier, the essay should get an A.

Part II: KNNs

Running the code on the provided data with k = 5 produces the following statistics:



Statistic	Value Observed	Value from HW Instructions
Precision	75.622 %	≈ 92%
Recall	77.551 %	≈ 84%
F-Measure	76.574 %	≈ 88%
Accuracy	82.161 %	≈ 91%

Table 1: KNNs statistics for the spam data, k = 5

These values are lower than those provided in the homework instructions by 10-15%. It is possible that the randomization happened differently. I provide the zero seed through the **np.random.seed(0)** command. It is possible that a different method was used by the instructor in the algorithm. There also was not a perfect match between numbers in previous assignments.