Greg Matthews
CS 383
February 19th, 2018

Assignment 4:
Naïve Bayes, Decision Trees and Nearest Neighbor

1 Theory

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

(a.) What is the sample entropy, $H(Y)$ from this training data (using log base 2) (2pts)?

Answer:

Number of positive and negative samples:

$$positive(+) = 3 + 4 + 4 + 1 = \mathbf{12}$$
$$negative(-) = 0 + 1 + 3 + 5 = \mathbf{9}$$

Computing sample Entropy $H(Y)$:

$$H(Y) = H(P(v_i), \ldots (P(v_n)) = \sum_{i=1}^{n} -P(v_i)\log_2 P(v_i) = -\frac{12}{21}\log_2\frac{12}{21} - \frac{9}{21}\log_2\frac{9}{21}$$
$$H(Y) = 0.4613 + 0.5238$$
$$H(Y) = \mathbf{0.985}$$

(b.) What are the information gains for branching on variables $x1$ and $x2$ (4pts)?

Answer:

Branching on variable x_1:

$$IG(x_1) = 0.985 - remainder(x_1)$$
$$IG(x_1) = 0.985 - \sum_{i=1}^{k} \frac{p_i + n_i}{p + n} * H(\frac{p_i}{p_i + n_i}, \frac{n_i}{p_i + n_i})$$
$$IG(x_1) = 0.985 - \left[\frac{7+1}{21} H\left(\frac{7}{8}, \frac{1}{8}\right) + \frac{5+8}{21} H\left(\frac{5}{13}, \frac{8}{13}\right)\right]$$

$$IG(x_1) = 0.985 - \left[\frac{8}{21} * \left[-\frac{7}{8}\log_2\frac{7}{8} - \frac{1}{8}\log_2\frac{1}{8}\right] + \frac{13}{21} * \left[-\frac{5}{13}\log_2\frac{5}{13} - \frac{8}{13}\log_2\frac{8}{13}\right]\right]$$

$$IG(x_1) = 0.985 - \left[\frac{8}{21} * (0.1686 + 0.375) + \frac{13}{21} * (0.5302 + 0.431)\right]$$

$$IG(x_1) = 0.985 - (0.207 + 0.595) = \mathbf{0.183}$$

<u>Branching on variable x_2:</u>

$$IG(x_1) = 0.985 - remainder(x_2)$$

$$IG(x_2) = 0.985 - \sum_{i=1}^{k} \frac{p_i + n_i}{p + n} * H(\frac{p_i}{p_i + n_i}, \frac{n_i}{p_i + n_i})$$

$$IG(x_2) = 0.985 - \left[\frac{7+3}{21} H\left(\frac{7}{10}, \frac{3}{10}\right) + \frac{5+6}{21} H\left(\frac{5}{11}, \frac{6}{11}\right)\right]$$

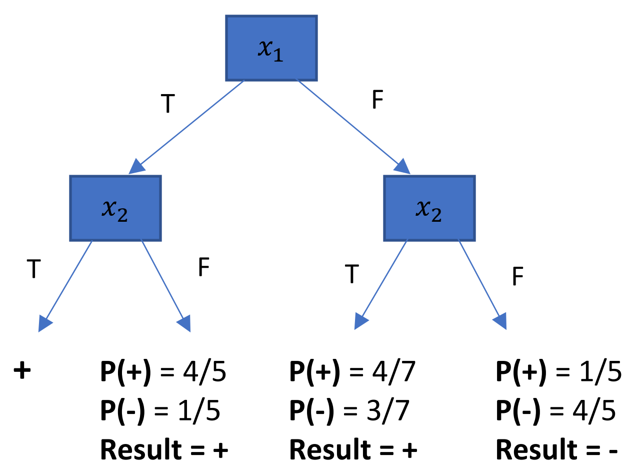
$$IG(x_2) = 0.985 - \left[\frac{10}{21} * \left[-\frac{7}{10}\log_2\frac{7}{10} - \frac{3}{10}\log_2\frac{3}{10}\right] + \frac{11}{21} * \left[-\frac{5}{11}\log_2\frac{5}{11} - \frac{6}{11}\log_2\frac{6}{11}\right]\right]$$

$$IG(x_2) = 0.985 - \left[\frac{10}{21} * (0.3602 + 0.521) + \frac{11}{21} * (0.517 + 0.477)\right]$$

$$IG(x_2) = 0.985 - (0.420 + 0.521) = \mathbf{0.044}$$

(c.) Draw the decision tree that would be learned by the ID3 algorithm without pruning from this training data (5pts)?

Answer:



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	4.2	No

(a.) What are the class priors, $P(A = Yes)$, $P(A = No)$? (1pt)

Answer:

$$P(A = Yes) = \frac{3}{5} = \mathbf{60}\%$$

$$P(A = No) = \frac{2}{5} = \mathbf{40}\%$$

(b.) Find the parameters of the Gaussians necessary to do Gaussian Naive Bayes classification on this decision to give an A or not. Standardize the features first over all the data together so that there is no unfair bias towards the features of different scales (5pts).

Answer:

<u>Standardizing Data:</u>

$$\mu_{\#chars} = 208.0, \quad \sigma_{\#chars} = 145.2$$

$$\mu_{avg.len} = 4.026, \quad \sigma_{avg.len} = 1.326$$

$$\begin{bmatrix} 0.0551 & 1.25 \\ -0.957 & 0.569 \\ 0.647 & -1.29 \\ -1.02 & -0.653 \\ 1.27 & 0.131 \end{bmatrix}$$

$$where\ 0th\ column\ is\ \#chars, 1st\ column\ is\ avg.len$$

<u>Computing # of Chars Gaussians:</u>

$$\mu_{yes,\#chars} = \frac{1}{3} * \sum_{i:y_i=yes} \left(x_{i,\#chars}\right) = \frac{1}{3} * (0.0551 - 0.957 - 1.02) = \mathbf{-0.640}$$

$$\mu_{no,\#chars} = \frac{1}{2} * \sum_{i:y_i=no} \left(x_{i,\#chars}\right) = \frac{1}{2} * (0.647 + 1.27) = \mathbf{0.959}$$

$$\sigma^2_{yes,\#chars} = \frac{1}{3} * \sum_{i:y_i=yes} \left(x_{i,\#chars} - \mu_{yes,\#chars}\right)^2$$

$$= \frac{1}{3} * [(0.0551 + 0.640)^2 + (-0.957 + 0.640)^2 + (-1.02 + 0.640)^2] = \mathbf{0.243}$$

$$\sigma^2_{no,\#chars} = \frac{1}{3} * \sum_{i:y_i=no} \left(x_{i,\#chars} - \mu_{no,\#chars}\right)^2$$

$$= \frac{1}{2} * [(0.647 - 0.959)^2 + (1.27 - 0.959)^2] = \mathbf{0.0647}$$

Normal Distribution: $\mathcal{N}\left(\mu_{yes,\#chars}, \ \sigma_{yes,\#chars}\right) = \boldsymbol{\mathcal{N}}(-\mathbf{0.640}, \sqrt{\mathbf{0.243}})$

Normal Distribution: $\mathcal{N}\left(\mu_{no,\#chars}, \ \sigma_{no,\#chars}\right) = \boldsymbol{\mathcal{N}}(\mathbf{0.959}, \sqrt{\mathbf{0.0647}})$

<u>Computing Average Word Length Gaussians:</u>

$$\mu_{yes,avg.len} = \frac{1}{3} * \sum_{i:y_i=yes} \left(x_{i,avg.len}\right) = \frac{1}{3} * (0.0551 - 0.957 - 1.02) = \mathbf{0.389}$$

$$\mu_{no,avg.len} = \frac{1}{2} * \sum_{i:y_i=no} \left(x_{i,avg.len}\right) = \frac{1}{2} * (0.647 + 1.27) = -\mathbf{0.580}$$

$$\sigma^2_{yes,avg.len} = \frac{1}{3} * \sum_{i:y_i=yes} \left(x_{i,avg.len} - \mu_{yes,avg.len}\right)^2$$

$$= \frac{1}{3} * [(0.0551 + 0.640)^2 + (-0.957 + 0.640)^2 + (-1.02 + 0.640)^2] = \mathbf{1.68}$$

$$\sigma^2_{no,avg.len} = \frac{1}{3} * \sum_{i:y_i=no} \left(x_{i,avg.len} - \mu_{no,avg.len}\right)^2$$

$$= \frac{1}{2} * [(0.647 - 0.959)^2 + (1.27 - 0.959)^2] = \mathbf{1.91}$$

Normal Distribution: $\mathcal{N}\left(\mu_{yes,avg.len}, \ \sigma_{yes,avg.len}\right) = \boldsymbol{\mathcal{N}}(\mathbf{0.389}, \sqrt{\mathbf{1.68}})$

Normal Distribution: $\mathcal{N}\left(\mu_{no,avg.len}, \ \sigma_{no,avg.len}\right) = \boldsymbol{\mathcal{N}}(-\mathbf{0.580}, \sqrt{\mathbf{1.91}})$

(c.) 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 (5pts).

Answer:

Standardize Data:

$$x = \left\{\frac{242 - 208.0}{145.2}, \frac{4.56 - 4.026}{1.326}\right\} = \{0.234, 0.403\}$$

$$estimating\ \ p(f_k = x_k \mid y = i)\ as\ p(x_k|\mathcal{N}(u_{ik}, \sigma_{ik}),$$

$$where\ i = classification, k = attribute, then$$

Computing probability of "yes":

$$p(y = yes \mid f = x) \propto p(y = i) * \prod_{k=1}^{D} p(x_k|\mathcal{N}(u_{ik}, \sigma_{ik})$$

Likelihood side-computation:

$$\prod_{k=1}^{D} p(x_k|\mathcal{N}(u_{ik}, \sigma_{ik})$$

$$= p(x_{\#chars}|\ \mathcal{N}\left(\mu_{yes,\#chars}, \sigma_{yes,\#chars}\right)) +$$

$$p(x_{avg.len}|\mathcal{N}\left(\mu_{yes,avg.len}, \sigma_{yes,avg.len}\right)$$

$$= p(0.234|\mathcal{N}(-0.640, \sqrt{0.243})) + \ p(0.403|\mathcal{N}(0.389, \sqrt{1.68}))$$

$$= \frac{1}{\sigma_{\#chars}\sqrt{2\pi}} e^{-\frac{(x - u_{\#chars})^2}{2\sigma^2_{\#chars}}} + \frac{1}{\sigma_{avg.len}\sqrt{2\pi}} e^{-\frac{\left(x - u_{avg.len}\right)^2}{2\sigma^2_{avg.len}}}$$

$$= \frac{1}{\sqrt{0.243}\sqrt{2\pi}} e^{-\frac{(0.234+0.640)^2}{2*0.234}} + \frac{1}{\sqrt{1.68}\sqrt{2\pi}} e^{-\frac{(0.403-0.389)^2}{2*1.68}}$$

$$= \frac{1}{1.236} e^{-1.632} + \frac{1}{3.249} e^{-5.83*10^{-5}} = 0.1582 + 0.3078 = 0.4659$$

Computing posterior probability of "yes":

$$p(y = yes \mid f = x) = p(y = i) * \prod_{k=1}^{D} p(x_k|\mathcal{N}(u_{ik}, \sigma_{ik}) =$$

$$0.6 * 0.4659 = \mathbf{0.2796}$$

Computing probability of "no":

$$p(y = no \mid f = x) \propto p(y = i) * \prod_{k=1}^{D} p(x_k|\mathcal{N}(u_{ik}, \sigma_{ik})$$

Likelihood side-computation:

$\prod_{k=1}^{D} p(x_k|\mathcal{N}(u_{ik}, \sigma_{ik})$

$$= p(x_{\#chars}| \mathcal{N}\left(\mu_{no,\#chars}, \sigma_{no,\#chars}\right)) +$$

$$p(x_{avg.len}|\mathcal{N}\left(\mu_{no,avg.len}, \sigma_{no,avg.len}\right)$$

$$= p(0.234|\mathcal{N}(0.959, \sqrt{0.0647})) + \ p(0.403|\mathcal{N}(-0.580, \sqrt{1.91}))$$

$$= \frac{1}{\sigma_{\#chars}\sqrt{2\pi}} e^{-\frac{(x-u_{\#chars})^2}{2\sigma_{\#chars}^2}} + \frac{1}{\sigma_{avg.len}\sqrt{2\pi}} e^{-\frac{\left(x-u_{avg.len}\right)^2}{2\sigma_{avg.len}^2}}$$

$$= \frac{1}{\sqrt{0.0647}\sqrt{2\pi}} e^{-\frac{(0.234-0.959)^2}{2*0.0647}} + \frac{1}{\sqrt{1.91}\sqrt{2\pi}} e^{-\frac{(0.403+0.580)^2}{2*1.91}}$$

$$= \frac{1}{0.637} e^{-4.062} + \frac{1}{3.464} e^{-0.2529} = 0.027 + 0.224 = 0.2512$$

Computing posterior probability of "no":

$$p(y = no \mid f = x) = p(y = i) * \prod_{k=1}^{D} p(x_k|\mathcal{N}(u_{ik}, \sigma_{ik}) =$$

$$0.4 * 0.2512 = \mathbf{0.1004}$$

Putting it all together:

$$p(y = yes \mid f = x) > \ p(y = no \mid f = x)$$

$$\mathbf{0.2796} > \ \mathbf{0.1004}$$

$$\boldsymbol{Therefore\ the\ essay\ is\ predicted\ to\ get\ an\ A.}$$

2 k-Nearest Neighbors (KNN)

Classification Statistics:

$$Precision = \frac{t_p}{t_p+f_p} = 0.8994 = \mathbf{89.4}\%$$

$$Recall = \frac{t_p}{t_p+f_n} = 0.8385 = \mathbf{83.6}\%$$

$$F1\ Measure = 2 * \frac{precision*recall}{precison+recall} = 0.8679 = \mathbf{86.8}\%$$

$$Accuracy = \frac{t_p+t_n}{t_p+t_n+f_p+f_n} = 0.9042 = \mathbf{90.4}\%$$