CS 383 HW2

Kevin Karnani

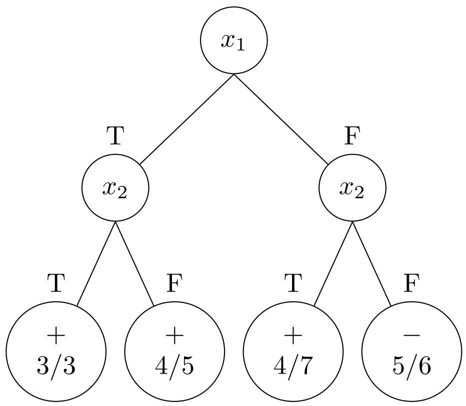
May 2021

1. Table summary:
   count$(+) = 3+4+4+1 = 12$, count$(-) = 0+1+3+5 = 9$, total $= 12+9 = 21$
   $P(+) = \frac{12}{21}$, $P(-) = \frac{9}{21}$

   (a) What is the sample entropy, $H(Y)$, from this training data (using log base 2)?
   $$H(Y) = -\sum_{i}^{n} P(v_i)\log_2(P(v_i))$$
   $$\implies H(+,-) = -\left(\frac{12}{21}\log_2\left(\frac{12}{21}\right) + \frac{9}{21}\log_2\left(\frac{9}{21}\right)\right) = 0.9852$$

   (b) What are the information gains for branching on variables x_1 and x_2?
   $$T1 = \begin{bmatrix} 1 & 1 & 3 \\ 1 & 0 & 4 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \; F1 = \begin{bmatrix} 0 & 1 & 4 \\ 0 & 0 & 1 \\ 0 & 1 & 3 \\ 0 & 0 & 5 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \end{bmatrix}$$
   count$(+_{T1}) = 3+4 = 7$, count$(-_{T1}) = 0+1 = 1$, total $= 7+1 = 8 \implies W_{T1} = \frac{8}{21}$
   count$(+_{F1}) = 4+1 = 5$, count$(-_{F1}) = 3+5 = 8$, total $= 5+8 = 13 \implies W_{F1} = \frac{13}{21}$
   count$(+_{T2}) = 3+4 = 7$, count$(-_{T2}) = 0+3 = 3$, total $= 7+3 = 10 \implies W_{T2} = \frac{10}{21}$
   count$(+_{F2}) = 4+1 = 5$, count$(-_{F2}) = 1+5 = 6$, total $= 5+6 = 11 \implies W_{F2} = \frac{11}{21}$
   $P(+_{T1}) = \frac{7}{8}$, $P(-_{T1}) = \frac{1}{8}$, $P(+_{F1}) = \frac{5}{13}$, $P(-_{T1}) = \frac{8}{13}$
   $P(+_{T2}) = \frac{7}{10}$, $P(-_{T2}) = \frac{3}{10}$, $P(+_{F2}) = \frac{5}{11}$, $P(-_{T2}) = \frac{6}{11}$
   $E_1 = -(W_{T1}E_{T1} + W_{F1}E_{F1})$,
   $E_2 = -(W_{T2}E_{T2} + W_{F2}E_{F2})$
   $$E_1 = -\left[\frac{8}{21}\left(\frac{1}{8}\log_2\left(\frac{1}{8}\right)\right) + \frac{7}{8}\log_2\left(\frac{7}{8}\right) + \frac{13}{21}\left(\frac{5}{13}\log_2\left(\frac{5}{13}\right)\right) + \frac{8}{13}\log_2\left(\frac{8}{13}\right)\right] = .8021$$
   $$E_2 = -\left[\frac{10}{21}\left(\frac{7}{10}\log_2\left(\frac{7}{10}\right)\right) + \frac{3}{10}\log_2\left(\frac{3}{10}\right) + \frac{11}{21}\left(\frac{5}{11}\log_2\left(\frac{5}{11}\right)\right) + \frac{6}{11}\log_2\left(\frac{6}{11}\right)\right] = .9403$$
   $\implies I(x_1) = H(+,-) - E_1 = .1831$
   $\implies I(x_2) = H(+,-) - E_1 = .0449$

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



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.

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

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

   (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.

   $$D = \begin{bmatrix} 216 & 5.68 \\ 69 & 4.78 \\ 302 & 2.31 \\ 60 & 3.16 \\ 393 & 4.2 \end{bmatrix} \implies \mu = \begin{bmatrix} 208 & 4.026 \end{bmatrix}, \sigma = \begin{bmatrix} 145.2154 & 1.3256 \end{bmatrix}$$

   $$\implies D_s = \begin{bmatrix} (216-208)/145.2154 & (5.68-4.026)/1.19 \\ (69-208)/145.2154 & (4.78-4.026)/1.19 \\ (302-208)/145.2154 & (2.31-4.026)/1.19 \\ (60-208)/145.2154 & (3.16-4.026)/1.19 \\ (393-208)/145.2154 & (4.2-4.026)/1.19 \end{bmatrix} = \begin{bmatrix} 0.055091 & 1.247714 \\ -0.957199 & 0.568789 \\ 0.647314 & -1.294484 \\ -1.019176 & -0.653277 \\ 1.273970 & 0.131259 \end{bmatrix}$$

   $\mu(A, C) = (0.055091 - 0.957199 - 1.019176)/3 = -0.6404$
   $\mu(A, W) = (1.247714 + 0.568789 - 0.653277)/3 = 0.3877$
   $\sigma^2(A, C) = ((0.055 + 0.6404)^2 + (-0.957 + 0.6404)^2 + (-1.019 + 0.6404)^2)/3 = 0.2425$
   $\sigma^2(A, W) = ((1.247 - 0.3877)^2 + (0.568 - 0.3877)^2 + (-0.653 - 0.3877)^2)/3 = 0.6187$
   $\mu(\neg A, C) = (0.64731446 + 1.27396994)/2 = 0.9606$
   $\mu(\neg A, W) = (-1.29448434 + 0.1312589)/2 = -0.5816$
   $\sigma^2(\neg A, C) = ((0.64731446 - 0.9606)^2 + (1.27396994 - 0.9606)^2)/2 = 0.098174$
   $\sigma^2(\neg A, C) = ((-1.29448434 + 0.5816)^2 + (0.1312589 + 0.5816)^2)/2 = 0.5082$

   (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.

   $S = \begin{bmatrix} 242 & 4.56 \end{bmatrix} \implies S_s = \begin{bmatrix} (242-208)/145.2154 & (4.56-4.026)/1.3256 \end{bmatrix} = \begin{bmatrix} 0.2341 & 0.4028 \end{bmatrix}$

$P(A \mid C, W) = P(A)P(C \mid A)P(C \mid A)$; $P(x \mid y) = \dfrac{1}{\sqrt{2\pi\sigma_y^2}} \exp\Big(-\dfrac{(x-\mu_y)^2}{2\sigma_y^2}\Big)$
$\implies P(A \,|\, S_s) = \dfrac{3}{10\pi\sigma(A,C)\sigma(A,W)} \exp\Big(-\dfrac{(x_1-\mu(A,C))^2}{2\sigma^2(A,C)}\Big) \exp\Big(-\dfrac{(x_2-\mu(A,W))^2}{2\sigma^2(A,W)}\Big)$
$\implies P(A \mid S_s) = 0.050931$

$P(\neg A \mid C, W) = P(\neg A)P(C \mid \neg A)P(C \mid \neg A)$
$\implies P(\neg A \,|\, S_s) = \dfrac{1}{5\pi\sigma(\neg A,C)\sigma(\neg A,W)} \exp\Big(-\dfrac{(x_1-\mu(\neg A,C))^2}{2\sigma^2(\neg A,C)}\Big) \exp\Big(-\dfrac{(x_2-\mu(\neg A,W))^2}{2\sigma^2(\neg A,W)}\Big)$
$\implies P(\neg A \mid S_s) = 7.4685 * 10^{-3}$
Since $P(A \mid S_s) > P(\neg A \mid S_s)$, an essay with 242 characters and an average word length of 4.56 should get an A.

3. Consider the following questions pertaining to a k-Nearest Neighbors algorithm:

   (a) How could you use a *validation* set to determine the user-defined parameter k?

   The simplest way would be to try various values of k and record the error for each iteration, then pick the k with the least error. A more complex way that may yield better results would be to use cross-folds validation by splitting the data into n partitions (folds) and then iterating through each fold, changing the validation each time, and averaging out the error yielded by each k.