CSE 5243 HOMEWORK 3

BRENDAN WHITAKER

Chapter 2 Exercises

19. *We compute several measures of similarity for the given vectors.*
    (a) $x = (1, 1, 1, 1), y = (2, 2, 2, 2)$. We compute cosine similarity, correlation, and Euclidean distance.

$$\begin{aligned}
\text{sim}_{\cos}(x, y) &= \frac{x \cdot y}{||x||||y||} = \frac{2+2+2+2}{\sqrt{4}\sqrt{16}} = \frac{8}{8} = 1. \\
\overline{x} &= \frac{1}{4}(1+1+1+1) = 1. \\
\overline{y} &= \frac{1}{4}(2+2+2+2) = 2. \\
s_{xy} &= \frac{1}{3}(0+0+0+0) = 0. \\
s_x &= \sqrt{\frac{1}{3}(0+0+0+0)} = 0. \\
s_y &= \sqrt{\frac{1}{3}(0+0+0+0)} = 0. \\
\text{corr}(x, y) &= \frac{s_{xy}}{s_x s_y} = \frac{0}{0} = \text{Undef.}
\end{aligned} \tag{1}$$

Note since we are in the odd case where the standard deviations of both vectors are zero since all their components are the same, we get zero in the denominator, but since our covariance is also zero, by convention we take $\text{corr}(x, y) = 0$.

$$d(x, y) = \sqrt{(2-1)^2 + (2-1)^2 + (2-1)^2 + (2-1)^2} = \sqrt{4} = 2. \tag{2}$$

Date: SP18.

(b) $x = (0, 1, 0, 1), y = (1, 0, 1, 0)$. We compute cosine similarity, correlation, Euclidean distance, and Jaccard coefficient.

$$\begin{aligned}
\text{sim}_{\cos}(x, y) &= \frac{x \cdot y}{||x||||y||} = \frac{4(0)}{\sqrt{2}\sqrt{2}} = 0. \\
\overline{x} &= \frac{1}{4}(2) = \frac{1}{2}. \\
\overline{y} &= \frac{1}{4}(2) = \frac{1}{2}. \\
s_{xy} &= \frac{1}{3}(-\frac{1}{4} \cdot 4) = -\frac{1}{3}. \\
s_x &= \sqrt{\frac{1}{3}(4 \cdot \frac{1}{4})} = \sqrt{\frac{1}{3}}. \\
s_y &= \sqrt{\frac{1}{3}(4 \cdot \frac{1}{4})} = \sqrt{\frac{1}{3}}. \\
\text{corr}(x, y) &= \frac{s_{xy}}{s_x s_y} = \frac{-\frac{1}{3}}{\sqrt{\frac{1}{3}}\sqrt{\frac{1}{3}}} = \frac{-\frac{1}{3}}{\frac{1}{3}} = -1. \\
d(x, y) &= \sqrt{(0-1)^2 + (1-0)^2 + (0-1)^2 + (1-0)^2} = \sqrt{4} = 2. \\
J(x, y) &= \frac{f_{11}}{f_{01} + f_{10} + f_{11}} = \frac{0}{2 + 2 + 0} = 0.
\end{aligned} \tag{3}$$

(c) $x = (0, -1, 0, 1), y = (1, 0, -1, 0)$. We compute cosine similarity, correlation, Euclidean distance.

$$\begin{aligned}
\text{sim}_{\cos}(x, y) &= \frac{x \cdot y}{||x||||y||} = \frac{4(0)}{\sqrt{2}\sqrt{2}} = 0. \\
\overline{x} &= \frac{1}{4}(0) = 0. \\
\overline{y} &= \frac{1}{4}(0) = 0. \\
s_{xy} &= \frac{1}{3}(4 \cdot 0) = 0. \\
s_x &= \sqrt{\frac{1}{3}(2 \cdot 1)} = \sqrt{\frac{2}{3}}. \\
s_y &= \sqrt{\frac{1}{3}(2 \cdot 1)} = \sqrt{\frac{2}{3}}. \\
\text{corr}(x, y) &= \frac{s_{xy}}{s_x s_y} = \frac{0}{\sqrt{\frac{2}{3}}\sqrt{\frac{2}{3}}} = 0. \\
d(x, y) &= \sqrt{(0-1)^2 + (-1-0)^2 + (0+1)^2 + (1-0)^2} = \sqrt{4} = 2.
\end{aligned} \tag{4}$$

(d) $x = (2, -1, 0, 2, 0, -3), y = (-1, 1, -1, 0, 0, -1)$. We compute cosine similarity, correlation.

$$\begin{aligned}
\text{sim}_{\cos}(x, y) &= \frac{x \cdot y}{||x||||y||} = \frac{-2 - 1 + 0 + 0 + 0 + 3}{\sqrt{4+1+0+4+0+9}\sqrt{1+1+1+1}} \\
&= \frac{0}{\sqrt{18}\sqrt{4}} = 0. \\
\overline{x} &= \frac{1}{6}(2 - 1 + 2 - 3) = 0. \\
\overline{y} &= \frac{1}{6}(-2) = -\frac{1}{3}. \\
s_{xy} &= \frac{1}{5}(-2\frac{2}{3} - 1\frac{4}{3} + 0 + 2\frac{1}{3} + 0 + 3\frac{2}{3}) \\
&= \frac{1}{5}(-\frac{8}{3} + \frac{2}{3} + 2) = 0. \\
s_x &= \sqrt{\frac{1}{5}(4 + 1 + 4 + 9)} = \sqrt{\frac{18}{5}}. \\
s_y &= \sqrt{\frac{1}{5}(4)} = \sqrt{\frac{4}{5}}. \\
\text{corr}(x, y) &= \frac{s_{xy}}{s_x s_y} = \frac{0}{\sqrt{\frac{18}{5}}\sqrt{\frac{4}{5}}} = 0.
\end{aligned} \tag{5}$$

Chapter 4 Exercises

2. *Consider the binary classification problem given by Table 4.7.*

(a) *We compute the Gini index of overall collection.*

Let t_0 be the root node of the entire collection of training examples. Observe:

$$\begin{aligned}
Gini(t_0) &= 1 - \sum_{i=0}^{1} [p(i|t_0)]^2 \\
&= 1 - \left(\left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2 \right) \\
&= 1 - \left(\frac{1}{4} + \frac{1}{4}\right) \\
&= 1 - \frac{1}{2} \\
&= \frac{1}{2}.
\end{aligned} \tag{6}$$

(b) *We compute the Gini index of* `Customer ID`.

Let t_j be the j-th `Customer ID` node. Since each node only has a single data point, the sum will be equal to 1, since one of the probabilities $p(i|t_j)$ will be 1, and the other 0. So $Gini(t_j) = 0$ for all j. Hence the Gini index is 0.

(c) *We compute the Gini index of* **`gender`**.
Let t_M be the male node and t_F be the female node. We have:

$$\begin{aligned}
Gini(t_M) &= 1 - \sum_{i=0}^{1} [p(i|t_M)]^2 \\
&= 1 - \left(\left(\frac{4}{10} \right)^2 + \left(\frac{6}{10} \right)^2 \right) \\
&= 1 - \left(\frac{4}{25} + \frac{9}{25} \right) \\
&= 1 - \frac{13}{25} \\
&= \frac{12}{25}. \\
Gini(t_F) &= 1 - \sum_{i=0}^{1} [p(i|t_F)]^2 \\
&= 1 - \left(\left(\frac{4}{10} \right)^2 + \left(\frac{6}{10} \right)^2 \right) \\
&= 1 - \left(\frac{4}{25} + \frac{9}{25} \right) \\
&= 1 - \frac{13}{25} \\
&= \frac{12}{25}.
\end{aligned} \tag{7}$$

So the Gini index is $\frac{10}{20}\frac{12}{25} + \frac{10}{20}\frac{12}{25} = \frac{12}{25} = 0.48$.

(d) *We compute the Gini index of* **`car type`**.

Let t_F be the family node and t_S be the sports node, and let t_L be the luxury node. We have:

$$\begin{aligned}
Gini(t_S) &= 1 - \sum_{i=0}^{1}[p(i|t_S)]^2 \\
&= 1 - \left(\left(\frac{10}{10}\right)^2 + \left(\frac{0}{10}\right)^2\right) \\
&= 1 - 1 \\
&= 0. \\
Gini(t_F) &= 1 - \sum_{i=0}^{1}[p(i|t_F)]^2 \\
&= 1 - \left(\left(\frac{1}{4}\right)^2 + \left(\frac{3}{4}\right)^2\right) \\
&= 1 - \left(\frac{1}{16} + \frac{9}{16}\right) \\
&= 1 - \frac{5}{8} \\
&= \frac{3}{8}. \\
Gini(t_L) &= 1 - \sum_{i=0}^{1}[p(i|t_L)]^2 \\
&= 1 - \left(\left(\frac{1}{8}\right)^2 + \left(\frac{7}{8}\right)^2\right) \\
&= 1 - \left(\frac{1}{64} + \frac{49}{64}\right) \\
&= 1 - \frac{25}{32} \\
&= \frac{7}{32}.
\end{aligned} \tag{8}$$

So the Gini index is $\frac{8}{20}0 + \frac{4}{20}\frac{3}{8} + \frac{8}{20}\frac{7}{32} = 0.1625$.

(e) *We compute the Gini index of* ***`shirt size`****.*

Let t_S be the small node and t_M be the medium node, and let t_L be the large node, and t_E be the extra large node. We have:

$$\begin{aligned}
Gini(t_S) &= 1 - \sum_{i=0}^{1}[p(i|t_S)]^2 \\
&= 1 - \left(\left(\frac{3}{5}\right)^2 + \left(\frac{2}{5}\right)^2\right) \\
&= 1 - \left(\frac{9}{25} + \frac{4}{25}\right) \\
&= 1 - \frac{13}{25} \\
&= \frac{12}{25}. \\
Gini(t_M) &= 1 - \sum_{i=0}^{1}[p(i|t_M)]^2 \\
&= 1 - \left(\left(\frac{3}{7}\right)^2 + \left(\frac{4}{7}\right)^2\right) \\
&= 1 - \left(\frac{9}{49} + \frac{16}{49}\right) \\
&= 1 - \frac{25}{49} \\
&= \frac{24}{49}. \\
Gini(t_L) &= 1 - \sum_{i=0}^{1}[p(i|t_L)]^2 \\
&= 1 - \left(\left(\frac{2}{4}\right)^2 + \left(\frac{2}{4}\right)^2\right) \\
&= 1 - \left(\frac{1}{4} + \frac{1}{4}\right) \\
&= 1 - \frac{1}{2} \\
&= \frac{1}{2}. \\
Gini(t_E) &= 1 - \sum_{i=0}^{1}[p(i|t_E)]^2 \\
&= 1 - \left(\left(\frac{2}{4}\right)^2 + \left(\frac{2}{4}\right)^2\right) \\
&= 1 - \left(\frac{1}{4} + \frac{1}{4}\right) \\
&= 1 - \frac{1}{2} \\
&= \frac{1}{2}.
\end{aligned} \tag{9}$$

So the Gini index is $\frac{5}{20}\frac{12}{25} + \frac{7}{20}\frac{24}{49} + \frac{4}{20}\frac{1}{2} + \frac{4}{20}\frac{1}{2} = 0.3914$.

(f) *Which of these 3 is preferred?*
Car type is preferred since its Gini index is lowest at 0.1625.

(g) *Explain why* `customer ID` *should not be used even though it's genie index is zero.*
`customer ID` should not be used since it only has a single data point per node, so it's likely that the customer ID's in any test data will not be ones we have already seen, assuming customer ID is unique, so we won't be able to use our model to gain any new information about the test set.

3. *Consider the information in Table 4.8 for a binary classification problem.*

(a) *Find entropy of examples with respect to positive class.* Observe:

$$\begin{aligned} Entropy &= -\sum_{i=0}^{1} p(i)\log_2 p(i) \\ &= -\left(\frac{4}{9}\log_2\frac{4}{9} + \frac{5}{9}\log_2\frac{5}{9}\right) \\ &= -\left(-.52 - .471\right) \\ &= 0.991. \end{aligned} \tag{10}$$

(b) *What are the information gains of a_1 and a_2 relative to these training examples?*
Let I be entropy. Recall:

$$\Delta_{info} = I - \sum_{j=1}^{k} \frac{N(v_j)}{N} I(v_j) \tag{11}$$

So we have:

$$\Delta_{info} = 0.991 - \sum_{j=1}^{2} \frac{N(a_i)}{9} I(a_i) \tag{12}$$

We compute:

$$\begin{aligned}
I(a_1 = T) &= -\sum_{i=0}^{1} p(i|a_1 = F) \log_2 p(i|a_1 = F) \\
&= -(-.5 - .311) \\
&= 0.811. \\
I(a_1 = F) &= -\sum_{i=0}^{1} p(i|a_1 = F) \log_2 p(i|a_1 = F) \\
&= 0.971 \\
I(a_2 = T) &= -\sum_{i=0}^{1} p(i|a_2 = T) \log_2 p(i|a_2 = T) \\
&= -(-.529 - .442) \\
&= 0.971 \\
I(a_2 = F) &= -\sum_{i=0}^{1} p(i|a_2 = F) \log_2 p(i|a_2 = F) \\
&= 1.
\end{aligned} \tag{13}$$

Then:

$$\begin{aligned}
\Delta_{info}(a_1) &= 0.991 - \sum_{j=1}^{2} \frac{N(a_i)}{9} I(a_i) \\
&= 0.991 - \left(\frac{4}{9}0.811 + \frac{5}{9}0.971\right) \\
&= 0.091. \\
\Delta_{info}(a_2) &= 0.991 - \sum_{j=1}^{2} \frac{N(a_i)}{9} I(a_i) \\
&= 0.991 - \left(\frac{4}{9}1 + \frac{5}{9}0.971\right) \\
&= 0.016.
\end{aligned} \tag{14}$$

(c) *We compute info gain for every possible split of* a_3.

Let Δ_i denote the info gain of the split $\leq i$ and $> i$. Then we have:

$$\begin{aligned}
\Delta_{1.5} &= 0.991 \\
\Delta_2 &= 0.991 - \left(\frac{1}{9}0 + \frac{8}{9}(.954)\right) \\
&= 0.143 \\
\Delta_{3.5} &= 0.991 - \left(\frac{2}{9} + \frac{7}{9}(.985)\right) \\
&= 0.0025 \\
\Delta_{4.5} &= 0.991 - \left(\frac{3}{9}(.918) + \frac{6}{9}(.918)\right) \\
&= 0.073 \\
\Delta_{5.5} &= 0.991 - \left(\frac{5}{9}(.971) + \frac{4}{9}\right) \\
&= 0.0071 \\
\Delta_{6.5} &= 0.991 - \left(\frac{6}{9} + \frac{3}{9}(.918)\right) \\
&= 0.018 \\
\Delta_{7.5} &= 0.991 - \left(\frac{8}{9}0 + \frac{1}{9}0\right) \\
&= 0.991 \\
\Delta_{8.5} &= 0.991.
\end{aligned} \tag{15}$$

(d) The best split of all three is Δ_2 on a_3 because it is non trivial, and it has the highest information gain.
(e) a_1 is better out of first two since it has a lower classification rate.
(f) Gini of T for a_1 is 3/8. Gini of F for a_1 is 8/25. So $Gini(a_1) = 4/9 * 3/8 + 5/9 * 8/25 = 0.344$.
Gini of T for a_2 is 12/25, and for F is 1/2. So $Gini(a_2) = 5/9*12/25+4/9*1/2 = 0.489$. And so a_1 is better again.

8. We omit writing out some computations because they are very simple, for brevity.
(a) Optimistic estimate is 1/2.
(b) Pessimistic is 7/10.
(c) Using validation set, gen error is 1/5.

Chapter 5 Exercises

(1) (a) We do $R1$. It is 9.2 since we have 12 pos, 3 neg, and 50 total in the set, with 21 neg and 29 pos. Similarly, we have 0.547 for $R3$ and .886 for $R2$. So $R1$ is best.
(b) Laplace for $R1$ is 0.765, for $R2$ is 2/3. And for R3 is 0.643. So R1 is best.
(c) The mestimate is 0.774, 0.68 and .654 for R 1,2,3 respectively. So R1 is still best.
(d) For the next 3 parts, R1 has accuracy 0.8, and R2 has accuracy 0.7. If we discard none, R3 has accuracy 1/3.

(e) If we discard the positives covered by R1 is has accuracy 6/10.
(f) if we discard both covered by R1, it has accuracy 0.75. R1 is best in all three cases.

(2) *We compute statistics for the given confusion matrix C.*
(a) Accuracy is 0.686. Error rate is 0.313. TPR is 0.742. FPR is 0.339. Precision is 0.504. F measure is 0.6.