DSA1101 Final Exam Solution

Nicholas Russell Saerang

Semester 1, 2017/2018

1. Using the Apriori algorithm, we shall find the support of 1-itemsets, 2-itemsets, and so on. As an exception, we compute Support($\{AB\}$) before Support($\{C\}$).

   - 1-itemsets

$$\begin{aligned}
\text{Support}(\{A\}) &= 0.75 > 0.25 \\
\text{Support}(\{B\}) &= 0.75 > 0.25 \\
\text{Support}(\{C\}) &= \frac{\text{Support}(\{ABC\})}{\text{Lift}(\{AB\} \to \{C\}) \times \text{Support}(\{AB\})} \\
&= \frac{0.3}{0.66 \times 0.6} = 0.756 \textbf{ (3 d.p.)} > 0.25 \\
\text{Support}(\{D\}) &= \frac{\text{Confidence}(\{C\} \to \{D\})}{\text{Lift}(\{C\} \to \{D\})} \\
&= \frac{0.4}{0.53} = 0.755 \textbf{ (3 d.p.)} > 0.25 \\
\text{Support}(\{E\}) &= \frac{\text{Support}(\{BCDE\})}{\text{Confidence}(\{E\} \to \{BCD\})} \\
&= \frac{0.1}{0.5} = 0.2 < 0.25
\end{aligned}$$

   - 2-itemsets

$$\begin{aligned}
\text{Support}(\{AB\}) &= \text{Confidence}(\{A\} \to \{B\}) \times \text{Support}(\{A\}) \\
&= 0.8 \times 0.75 = 0.6 > 0.25 \\
\text{Support}(\{AC\}) &= \text{Confidence}(\{A\} \to \{C\}) \times \text{Support}(\{A\}) \\
&= 0.8 \times 0.75 = 0.6 > 0.25 \\
\text{Support}(\{BC\}) &= \text{Lift}(\{B\} \to \{C\}) \times \text{Support}(\{B\}) \times \text{Support}(\{C\}) \\
&= 1.07 \times 0.75 \times 0.756 = 0.607 \textbf{ (3 d.p.)} > 0.25 \\
\text{Support}(\{CD\}) &= \text{Confidence}(\{C\} \to \{D\}) \times \text{Support}(\{C\}) \\
&= 0.4 \times 0.756 = 0.302 \textbf{ (3 d.p.)} > 0.25 \\
\text{Support}(\{BD\}) &= 0.25 \leq 0.25 \\
\text{Support}(\{AD\}) &= 0.3 > 0.25
\end{aligned}$$

   - 3-itemsets

$$\begin{aligned}
\text{Support}(\{ABC\}) &= 0.3 > 0.25 \\
\text{Support}(\{ABD\}) &= \text{Confidence}(\{AD\} \to \{B\}) \times \text{Support}(\{AD\}) \\
&= 0.67 \times 0.3 = 0.201 < 0.25 \\
\text{Support}(\{BCD\}) &= 0.3 > 0.25 \\
\text{Support}(\{ACD\}) &= \text{Lift}(\{AD\} \to \{C\}) \times \text{Support}(\{AD\}) \times \text{Support}(\{C\}) \\
&= 1.33 \times 0.3 \times 0.756 = 0.302 \textbf{ (3 d.p.)} > 0.25
\end{aligned}$$

- 4-itemsets

$$\text{Support}(\{BCDE\}) = 0.1 < 0.25$$

Therefore, the itemsets with frequent support are

$$\{A\}, \{B\}, \{C\}, \{D\}, \{AB\}, \{AC\}, \{BC\}, \{CD\}, \{AD\}, \{ABC\}, \{BCD\}, \{ACD\}$$

Quick Commentary
If you notice, Support($\{BCD\}$) > Support($\{BD\}$), which is illegal. Although the question is still doable in this case, we actually exclude $\{BCD\}$ from the answer since $\{BD\}$ is not in the answer, too.

2. We can use linear regression by expressing $\ln(P)$ as a function of t.
Let

$$T = 1 : 15$$

and

$$N = c(355, 211, 197, 166, 142, 106, 104, 60, 56, 38, 36, 32, 21, 19, 15)$$

Running the following R code has given you the final answer.

```
simple_LS <- function(x,y){
beta_1 <- (sum(x*y)-mean(y)*sum(x))/(sum(x∧2)-mean(x)*sum(x));
beta_0 <- mean(y)-beta_1*mean(x);
return(c(beta_0,beta_1));
}

simple_LS(T,log(N))
```

Which gives you $\ln(n_0) = 5.9731603$ and $\beta = -0.2184253$. Thus, $n_0 = 392.7449 = 393$ (rounded to 1 d.p.)

PS : I know it shouldn't be done with R since it is worth 15 points. However, since R is allowed to be used, the question is a giveaway.

3. Name the respective test data values A, B, C, D and E. Since A is the only data with NSAL $\geq$ 4419, we shall explore A first. From A, we get TDAYS $\geq$ 283, MCDAYS < 327, FEXP < 6374, and BED ge 112. Hence we predict A as 1. (notice in the tree that the class shown is the minority class)

Next, only B and C that has FEXP $\geq$ 2168. Therefore, we predict both as 0.

Both D and E have PCREV $\geq$ 8636 and FEXP < 1138. Hence, we predict both as 0.

4. (a) Using the least squares method, we have

$$\begin{aligned}
\hat{\beta} &= \frac{\sum_{i=1}^n x_i y_i - \bar{y}\sum_{i=1}^n x_i}{\sum_{i=1}^n x_i^2 - \bar{x}\sum_{i=1}^n x_i} \\
&= \frac{\sum_{i=1}^n x_i(y_i - \bar{y})}{\sum_{i=1}^n x_i(x_i - \bar{x})} \\
\hat{\delta} &= \frac{\sum_{i=1}^n y_i x_i - \bar{x}\sum_{i=1}^n y_i}{\sum_{i=1}^n y_i^2 - \bar{y}\sum_{i=1}^n y_i} \\
&= \frac{\sum_{i=1}^n y_i(x_i - \bar{x})}{\sum_{i=1}^n y_i(y_i - \bar{y})} \\
\hat{\alpha} &= \bar{y} - \hat{\beta}\bar{x} \\
\hat{\gamma} &= \bar{x} - \hat{\delta}\bar{y}
\end{aligned}$$

(simplification for $\hat{\alpha}$ and $\hat{\gamma}$ left to the reader)

(b) Note that

$$\hat{\beta} = r_{xy}\frac{\text{sd}(y)}{\text{sd}(x)}$$

and

$$\hat{\delta} = r_{xy}\frac{\text{sd}(x)}{\text{sd}(y)}$$

Therefore,

$$\hat{\beta}\cdot\hat{\delta} = r_{xy}^2$$

(c) Since

$$\bar{y} = \hat{\alpha} + \hat{\beta}\bar{x}$$

and

$$\bar{x} = \hat{\gamma} + \hat{\delta}\bar{y}$$

and both lines intersect at exactly one point, then the intersection point is $(\bar{x}, \bar{y})$.

5. (a) **Found a quicker solution so just read the red box instead of the whole question (update 25 November 2020).**

Suppose that the fitted linear regression model is denoted with

$$y = \beta_0 + \beta_1 x$$

Then,

$$\beta_1 = r_{xy}\frac{\text{sd}(y)}{\text{sd}(x)} = \frac{\text{cov}(x,y)}{\text{var}(x)}$$

and the residuals are computed as follows

$$e_i = y_i - (\beta_0 + \beta_1 x_i)$$

Therefore, with another fact that

$$\text{var}(a + bx + cy) = b^2\text{var}(x) + c^2\text{var}(y) + 2bc\ \text{cov}(x,y)$$

we have

$$\begin{aligned}
\frac{\frac{1}{n}\sum_{i=1}^n (y_i-\bar{y})^2 - \frac{1}{n}\sum_{i=1}^n (e_i-\bar{e})^2}{\frac{1}{n}\sum_{i=1}^n (y_i-\bar{y})^2} &= \frac{\text{var}(y) - \text{var}(e)}{\text{var}(y)} \\
&= \frac{\text{var}(y) - \text{var}(y - \beta_0 - \beta_1 x)}{\text{var}(y)} \\
&= \frac{\text{var}(y) - (\text{var}(y) + \beta_1^2\text{var}(x) - 2\beta_1\text{cov}(x,y))}{\text{var}(y)} \\
&= \frac{2\beta_1\text{cov}(x,y) - \beta_1^2\text{var}(x)}{\text{var}(y)} \\
&= \frac{2\frac{\text{cov}(x,y)}{\text{var}(x)}\text{cov}(x,y) - (\frac{\text{cov}(x,y)}{\text{var}(x)})^2\text{var}(x)}{\text{var}(y)} \\
&= \frac{\text{cov}(x,y)^2}{\text{var}(x)\text{var}(y)} = r_{xy}^2
\end{aligned}$$

PS : Claimed the answer to be r_{xy}^2 so that the whole calculation is simply backtracking. This is Question 5(b) from the 2016/2017 Semester 2 Exam.

Suppose that the fitted linear regression model is denoted with

$$y = \beta_0 + \beta_1 x$$

where

$$\beta_1 = \frac{\sum_{i=1}^n x_i y_i - \bar{y}\sum_{i=1}^n x_i}{\sum_{i=1}^n x_i^2 - \bar{x}\sum_{i=1}^n x_i} = \frac{\sum_{i=1}^n x_i(y_i - \bar{y})}{\sum_{i=1}^n x_i(x_i - \bar{x})}$$

and

$$\beta_0 = \bar{y} - \beta_1\bar{x}$$

Then, the residuals are computed as follows

$$e_i = y_i - (\beta_0 + \beta_1 x_i)$$

and thus

$$\bar{e} = \bar{y} - (\beta_0 + \beta_1\bar{x}) = 0 \to e_i - \bar{e} = y_i - (\beta_0 + \beta_1 x_i)$$

Thus,

$$\frac{\frac{1}{n}\sum_{i=1}^n (y_i - \bar{y})^2 - \frac{1}{n}\sum_{i=1}^n (e_i - \bar{e})^2}{\frac{1}{n}\sum_{i=1}^n (y_i - \bar{y})^2}$$
$$= \frac{\sum_{i=1}^n (y_i-\bar{y})^2 - \sum_{i=1}^n (y_i - (\beta_0+\beta_1 x_i))^2}{\sum_{i=1}^n (y_i-\bar{y})^2}$$
$$= \frac{\sum_{i=1}^n (y_i - \bar{y})^2 - \sum_{i=1}^n (y_i - ((\bar{y} - \beta_1\bar{x}) + \beta_1 x_i))^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$
$$= \frac{\sum_{i=1}^n (y_i - \bar{y})^2 - \sum_{i=1}^n (y_i - ((\bar{y} - \beta_1\bar{x}) + \beta_1 x_i))^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$
$$= \frac{\sum_{i=1}^n (y_i - \bar{y})^2 - \sum_{i=1}^n ((y_i - \bar{y}) - \beta_1(x_i - \bar{x}))^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$
$$= \frac{2\beta_1\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) - \beta_1^2\sum_{i=1}^n (x_i - \bar{x})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

Now we want to prove that

$$\begin{aligned}
\operatorname{cov}(x,y) &= \beta_1 \operatorname{var}(x) \\
\sum (x_i - \bar{x})(y_i - \bar{y}) &= \frac{\sum x_i(y_i - \bar{y})}{\sum x_i(x_i - \bar{x})}\sum (x_i - \bar{x})^2 \\
\sum (x_i - \bar{x})(y_i - \bar{y})\sum x_i(x_i - \bar{x}) &= \sum x_i(y_i - \bar{y})\sum (x_i - \bar{x})^2 \\
\left[\sum y_i(x_i - \bar{x}) - \bar{y}\sum (x_i - \bar{x})\right]\sum x_i(x_i - \bar{x}) &= \sum x_i(y_i - \bar{y})\left[\sum x_i(x_i - \bar{x}) - \bar{x}\sum (x_i - \bar{x})\right] \\
\sum x_i(x_i - \bar{x})\sum (y_i - \bar{y}) &= \sum x_i(y_i - \bar{y})\sum (x_i - \bar{x})
\end{aligned}$$

Which is true since

$$\sum (y_i - \bar{y}) = \sum (x_i - \bar{x}) = 0$$

Thus, we can continue with our counting.

$$\frac{2\beta_1\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) - \beta_1^2\sum_{i=1}^n (x_i - \bar{x})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$
$$= \frac{2\frac{\operatorname{cov}(x,y)}{\operatorname{var}(x)}\operatorname{cov}(x,y) - (\frac{\operatorname{cov}(x,y)}{\operatorname{var}(x)})^2\operatorname{var}(x)}{\operatorname{var}(y)}$$
$$= \frac{2(\operatorname{cov}(x,y))^2}{\operatorname{var}(x)\operatorname{var}(y)}$$
$$= r_{xy}^2.$$

(b) Since

$$\hat{y}_i = \hat{\alpha} + \hat{\beta}\cdot x_i$$

and

$$e_i = y_i - \hat{y}_i = y_i - \hat{\alpha} - \hat{\beta}\cdot x_i$$

Then,

$$
\begin{aligned}
\operatorname{cov}(\hat{y}, e) &= \operatorname{cov}(\hat{\alpha} + \hat{\beta}x, y - \hat{\alpha} - \hat{\beta}x) \\
&= \operatorname{cov}(\hat{\beta}x, y - \hat{\beta}x) \\
&= \hat{\beta}\operatorname{cov}(x, y) - \hat{\beta}^2\operatorname{cov}(x, x) \\
&= \hat{\beta}\operatorname{cov}(x, y) - \hat{\beta}^2\operatorname{var}(x)
\end{aligned}
$$

Recall the fact that

$$
\hat{\beta} = r_{xy}\frac{\operatorname{sd}(y)}{\operatorname{sd}(x)} = \frac{\operatorname{cov}(x, y)}{\operatorname{var}(x)}
$$

Finally,

$$
\begin{aligned}
\operatorname{cov}(\hat{y}, e) &= \hat{\beta}\operatorname{cov}(x, y) - \hat{\beta}^2\operatorname{var}(x) && \text{(rewrite)} \\
&= \frac{\operatorname{cov}(x, y)}{\operatorname{var}(x)}\operatorname{cov}(x, y) - \left(\frac{\operatorname{cov}(x, y)}{\operatorname{var}(x)}\right)^2 \operatorname{var}(x) \\
&= \frac{\operatorname{cov}(x, y)^2}{\operatorname{var}(x)} - \frac{\operatorname{cov}(x, y)^2}{\operatorname{var}(x)} = 0
\end{aligned}
$$