DSA1101 Final Exam Solution

Nicholas Russell Saerang

Semester 1, 2018/2019

1. (a) Table below is the classified output given thresholds.

i	actual Y_i	$P(Y_i = 1\|X_i)$	$\sigma = 0.3$	$\sigma = 0.6$	$\sigma = 0.8$
1	1	0.9	1 (TP)	1 (TP)	1 (TP)
2	1	0.5	1 (TP)	0 (FN)	0 (FN)
3	0	0.7	1 (FP)	1 (FP)	0 (TN)
4	1	0.4	1 (TP)	0 (FN)	0 (FN)
5	1	0.5	1 (TP)	0 (FN)	0 (FN)
6	0	0.2	0 (TN)	0 (TN)	0 (TN)
7	0	0.7	1 (FP)	1 (FP)	0 (TN)
8	1	0.9	1 (TP)	1 (TP)	1 (TP)
9	0	0.1	0 (TN)	0 (TN)	0 (TN)
10	0	0.1	0 (TN)	0 (TN)	0 (TN)

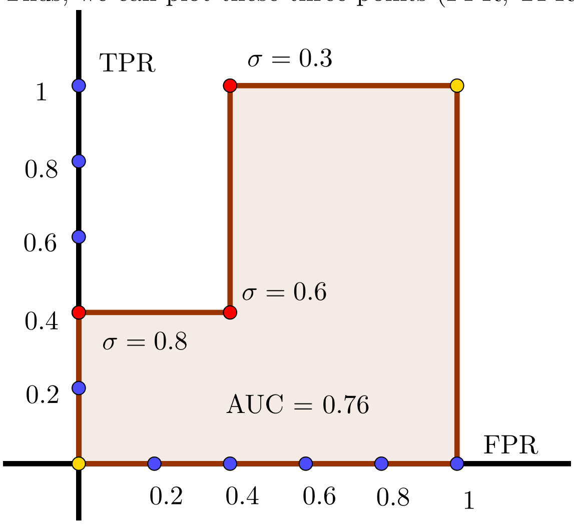
The confusion matrices are shown below.

$\sigma = 0.3$		Predicted y	
TPR = 1, FPR = 0.4		1	0
Actual y	1	5	0
	0	2	3

$\sigma = 0.6$		Predicted y	
TPR = 0.4, FPR = 0.4		1	0
Actual y	1	2	3
	0	2	3

$\sigma = 0.8$		Predicted y	
TPR = 0.4, FPR = 0		1	0
Actual y	1	2	3
	0	0	5

Thus, we can plot these three points (FPR, TPR) as shown below.



(b) If $\sigma > 0.9$, then all test points are predicted as $y = 0$, hence TPR = FPR = 0.

Likewise, if $\sigma < 0.1$, then all test points are predicted as $y = 1$, hence TPR = FPR = 1.

The area under the estimated ROC curve is

$$0.4^2 + 0.6 \times 1 = 0.76$$

(c) We can rewrite the equation as follows.

$$\text{accuracy} = \eta_1 \text{TPR} + \eta_2 \text{TNR}$$
$$\frac{TP + TN}{TP + FN + FP + TN} = \eta_1 \left(\frac{TP}{TP + FN} \right) + \eta_2 \left(\frac{TN}{FP + TN} \right)$$

Since $TP + FN$ and $FP + TN$ are both equal and constant, which is 5, we can multiply both sides by 10 and get

$$TP + TN = 2\eta_1 TP + 2\eta_2 TN$$

Therefore, by comparing coefficients we get $\eta_1 = \eta_2 = 0.5$.

2. (a) Since we are only interested in itemsets with larger than 0.1 support, the itemsets count must be larger than $250,000 \times 0.1 = 25,000$.
Therefore, the subsets with significant support are

$$\{A\}, \{B\}, \{C\}, \{D\}, \{BC\}, \{BD\}, \{CD\}, \{BCD\}$$

(b) Recall that

$$\text{Confidence}(X \to Y) = \frac{\text{Support}(X \wedge Y)}{\text{Support}(X)}$$

Hence,

$$\text{Support}(X \wedge Y) = \text{Confidence}(X \to Y) \times \text{Support}(X)$$

First of all, note that {E} itself is a satisfying itemset. We include this in our final answer. Next,

$$\begin{aligned}\text{Support}(\{AE\}) &= \text{Confidence}(\{A\} \to \{E\}) \times \text{Support}(\{A\}) \\ &= 0.1 \times \frac{100}{250} \\ &= 0.04 < 0.1\end{aligned}$$

$$\begin{aligned}\text{Support}(\{BCDE\}) &= \text{Confidence}(\{BCD\} \to \{E\}) \times \text{Support}(\{BCD\}) \\ &= 0.95 \times \frac{30}{250} \\ &= 0.114 > 0.1\end{aligned}$$

Since {BCDE} is a satisfying itemset, with the Apriori algorithm, all the subsets of {BCDE} containing E are also satisfying. Hence, {BE},{CE},{DE},{BCE},{BDE},{CDE} are also satisfying.

Finally, all the itemsets containing E with support greater than 0.1 are

$$\{E\}, \{BE\}, \{CE\}, \{DE\}, \{BCE\}, \{BDE\}, \{CDE\}, \{BCDE\}$$

(c) i. First,

$$\begin{aligned}\text{Lift}(F \to G) &= \frac{\text{Support}(FG)}{\text{Support}(F) \times \text{Support}(G)} \\ &= \frac{\text{Confidence}(G \to F)}{\text{Support}(F)} \\ &= \frac{0.8}{0.8} = 1\end{aligned}$$

ii. Since from the previous part Support(FG) = Support(F) $\times$ Support(G), we have

$$\begin{aligned}\text{Leverage}(F \to G) &= \text{Support}(FG) - \text{Support}(F) \times \text{Support}(G) \\ &= 0\end{aligned}$$

iii. The formal way.

$$\begin{aligned}\text{Confidence}(F \to G) &= \frac{\text{Support}(FG)}{\text{Support}(F)} \\ &= \frac{\text{Confidence}(G \to F) \times \text{Support}(G)}{\text{Support}(F)} \\ &= \frac{0.8 \times 0.7}{0.8} = 0.7\end{aligned}$$

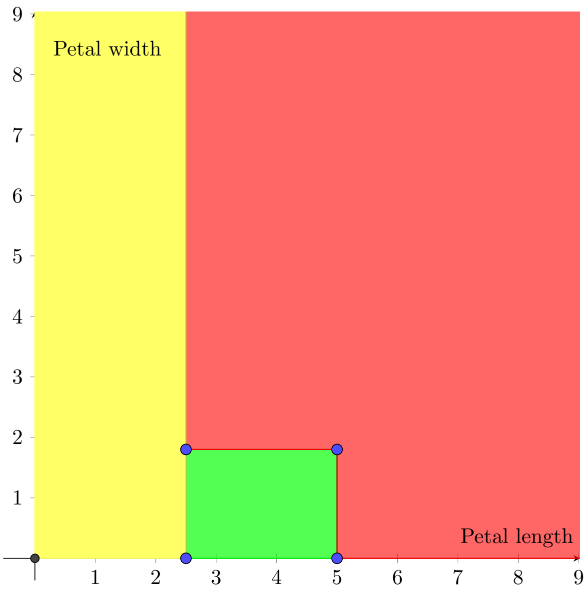
A much quicker way.

$$\begin{aligned}\text{Confidence}(F \to G) &= \text{Lift}(F \to G) \times \text{Support}(G) \\ &= 1 \times 0.7 = 0.7\end{aligned}$$

iv. Finally,

$$\begin{aligned}\text{Lift}(\{F, G\} \to H) &= \frac{\text{Support}(FGH)}{\text{Support}(FG) \times \text{Support}(H)} \\ &= \frac{\text{Confidence}(H \to \{F, G\})}{\text{Support}(FG)} \\ &= \frac{\text{Confidence}(H \to \{F, G\})}{\text{Confidence}(G \to F) \times \text{Support}(G)} \\ &= \frac{0.3}{0.8 \times 0.7} = \frac{15}{28}\end{aligned}$$

3. (a) Yellow for A, green for B, red for C.



(b) Supervised : Linear regression, logistic regression, naive Bayes, k-nearest neighbors, decision tree

Unsupervised : Association rules, k-means clustering

4. (a) The purity at the root node is

$$P(\text{buys} = \text{yes}) = \frac{8}{14}$$

Hence, the base entropy D_{buys} is

$$\begin{aligned}D_{buys} &= -\left(\frac{8}{14} \times log_2\left(\frac{8}{14}\right) + \frac{6}{14} \times log_2\left(\frac{6}{14}\right)\right) \\ &= 0.9852281\end{aligned}$$

We also have the following probabilities.

$$\begin{aligned}P(\text{age} = \text{senior, youth}) &= \frac{10}{14} = \frac{5}{7} \\ P(\text{age} = \text{middle aged}) &= \frac{4}{14} = \frac{2}{7} \\ P(\text{buys} = \text{no} \mid \text{age} = \text{senior, youth}) &= \frac{5}{10} = \frac{1}{2} \\ P(\text{buys} = \text{yes} \mid \text{age} = \text{middle aged}) &= \frac{3}{4}\end{aligned}$$

Thus, we can compute the conditional entropy as follows.

$$\begin{aligned}D_{buys|age} &= P(\text{age} = \text{senior, youth}) \times D_{buys|\text{age = senior, youth}} + \\ &\quad P(\text{age} = \text{middle aged}) \times D_{buys|\text{age = middle aged}} \\ &= -\left[\frac{5}{7}\left(\frac{1}{2} \times log_2\left(\frac{1}{2}\right) + \frac{1}{2} \times log_2\left(\frac{1}{2}\right)\right)\right] - \\ &\quad \left[\frac{2}{7}\left(\frac{3}{4} \times log_2\left(\frac{3}{4}\right) + \frac{1}{4} \times log_2\left(\frac{1}{4}\right)\right)\right] \\ &= \frac{5}{7} \times 1 + \frac{2}{7} \times 0.8112781 \\ &= 0.9460795\end{aligned}$$

Therefore, the entropy reduction or equivalently information gain using 'age' as predictor is 0.9852281 - 0.9460795 = 0.0391486 = 0.039 (rounded to 3 decimal places).

(b) From the given decision tree, we can traverse down the tree and predict that the customer **will buy** the computer. Now we shall predict using the Naive Bayes classifier, hence assuming age, student status, and income as independent attributes.

First, we have

$$P(\text{buys} = \text{yes}) = \frac{8}{14}, P(\text{buys} = \text{no}) = \frac{6}{14}$$

Then,

$$\begin{aligned}P(\text{age} = \text{youth} \mid \text{buys} = \text{yes}) = \frac{2}{8} = \frac{1}{4}&, P(\text{age} = \text{youth} \mid \text{buys} = \text{no}) = \frac{3}{6} = \frac{1}{2} \\ P(\text{income} = \text{medium} \mid \text{buys} = \text{yes}) = \frac{4}{8} = \frac{1}{2}&, P(\text{income} = \text{medium} \mid \text{buys} = \text{no}) = \frac{2}{6} = \frac{1}{3} \\ P(\text{student} = \text{yes} \mid \text{buys} = \text{yes}) = \frac{6}{8} = \frac{3}{4}&, P(\text{student} = \text{yes} \mid \text{buys} = \text{no}) = \frac{1}{6}\end{aligned}$$

Then,

$$P(\text{buys} = \text{yes} \mid \text{age, income, student}) = \frac{8}{14} \times \frac{1}{4} \times \frac{1}{2} \times \frac{3}{4} = \frac{3}{56}$$

and

$$P(\text{buys} = \text{no} \mid \text{age, income, student}) = \frac{6}{14} \times \frac{1}{2} \times \frac{1}{3} \times \frac{1}{6} = \frac{1}{84}$$

Since $\frac{3}{56} > \frac{1}{84}$, we classify this data as 'yes', thus the customer **will buy** the computer.

5. (a) The error term ϵ_i can be positive or negative, but they all sum up to 0. In order not to cancel each other, we do the least squares estimates by taking the minimum value of $\sum \epsilon_i^2$. In order to achieve this, we must have $\hat{\beta}_0$ and $\hat{\beta}_1$ to be the solution of

$$\frac{d\sum \epsilon_i^2}{d\beta_0} = \frac{d\sum \epsilon_i^2}{d\beta_1} = 0$$

Note that

$$\sum \epsilon_i^2 = \sum (y_i - \beta_0 - \beta_1 x_i)^2$$

Thus,

$$\begin{aligned}
\frac{d \sum \epsilon_i^2}{d\beta_0} &= \frac{d \sum (y_i - \beta_0 - \beta_1 x_i)^2}{d\beta_0} \\
&= -2 \sum (y_i - \beta_0 - \beta_1 x_i) = 0
\end{aligned}$$

$$\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \qquad \Rightarrow (1)$$

$$\begin{aligned}
\frac{d \sum \epsilon_i^2}{d\beta_1} &= \frac{d \sum (y_i - \beta_0 - \beta_1 x_i)^2}{d\beta_1} \\
&= -2 \sum x_i (y_i - \beta_0 - \beta_1 x_i) = 0
\end{aligned}$$

$$\sum_{i=1}^{n} x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \qquad \Rightarrow (2)$$

(b) Backtracking.
Note that from (1),

$$\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = \sum_{i=1}^{n} (y_i - \hat{y}_i) = 0$$

Also, $\bar{y}$ and $\hat{\beta}_0$ are constants, so

$$\sum_{i=1}^{n} \bar{y}(y_i - \hat{y}_i) = \sum_{i=1}^{n} \hat{\beta}_0 (y_i - \hat{y}_i) = 0 \qquad \Rightarrow (3)$$

From (2), we have

$$\sum_{i=1}^{n} x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

Therefore, since $\hat{\beta}_1$ is constant,

$$\sum_{i=1}^{n} \hat{\beta}_1 x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \qquad \Rightarrow (4)$$

Combine (3) and (4), we have

$$\sum_{i=1}^{n} \bar{y}(y_i - \hat{y}_i) = \sum_{i=1}^{n} (\hat{\beta}_0 + \hat{\beta}_1 x_i)(y_i - \hat{y}_i) = \sum_{i=1}^{n} \hat{y}_i (y_i - \hat{y}_i) = 0$$

We take the first and last sum expressions.

$$\sum_{i=1}^{n} \bar{y}(y_i - \hat{y}_i) = \sum_{i=1}^{n} \hat{y}_i(y_i - \hat{y}_i)$$

$$\sum_{i=1}^{n} \bar{y}y_i = \sum_{i=1}^{n}(\hat{y}_i y_i - \hat{y}_i^2 + \bar{y}\hat{y}_i)$$

$$\sum_{i=1}^{n} -2\bar{y}y_i = \sum_{i=1}^{n} -2(\hat{y}_i y_i - \hat{y}_i^2 + \bar{y}\hat{y}_i)$$

$$\sum_{i=1}^{n}(y_i^2 - 2\bar{y}y_i + \bar{y}^2) = \sum_{i=1}^{n}[y_i^2 - 2(\hat{y}_i y_i - \hat{y}_i^2 + \bar{y}\hat{y}_i) + \bar{y}^2]$$

$$\sum_{i=1}^{n}(y_i^2 - 2\bar{y}y_i + \bar{y}^2) = \sum_{i=1}^{n}(y_i^2 - 2\hat{y}_i y_i + \hat{y}_i^2) + (\hat{y}_i^2 - 2\bar{y}\hat{y}_i + \bar{y}^2)$$

$$\sum_{i=1}^{n}(y_i^2 - 2\bar{y}y_i + \bar{y}^2) = \sum_{i=1}^{n}(y_i^2 - 2\hat{y}_i y_i + \hat{y}_i^2) + \sum_{i=1}^{n}(\hat{y}_i^2 - 2\bar{y}\hat{y}_i + \bar{y}^2)$$

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$

6. (a) Since $P(Y=1|x) = \frac{\exp(\beta_0+\beta_1 x)}{1+\exp(\beta_0+\beta_1 x)}$, then $P(Y=0|x) = \frac{1}{1+\exp(\beta_0+\beta_1 x)}$. Therefore,

$$\begin{aligned} L(\beta_0,\beta_1) &= \prod_{i=1}^{n} P(Y=1|x_i)^{y_i} P(Y=0|x_i)^{1-y_i} \\ &= \prod_{i=1}^{n} \left(\frac{\exp(\beta_0+\beta_1 x_i)}{1+\exp(\beta_0+\beta_1 x_i)}\right)^{y_i} \left(\frac{1}{1+\exp(\beta_0+\beta_1 x_i)}\right)^{1-y_i} \\ &= \prod_{i=1}^{n} \frac{\exp(y_i(\beta_0+\beta_1 x_i))}{1+\exp(\beta_0+\beta_1 x_i)} \end{aligned}$$

(b) Since the odds of getting a head with a dollar coin is $\frac{n_{11}}{n_{01}}$ and the odds of getting a head with a non-dollar coin is $\frac{n_{10}}{n_{00}}$, then the odds ratio is

$$OR = \frac{\frac{n_{11}}{n_{01}}}{\frac{n_{10}}{n_{00}}} = \frac{n_{11}n_{00}}{n_{10}n_{01}}$$

(c) Note that

$$L(\beta_0,\beta_1) = \left(\frac{\exp(\beta_0+\beta_1)}{1+\exp(\beta_0+\beta_1)}\right)^{n_{11}} \left(\frac{\exp(\beta_0)}{1+\exp(\beta_0)}\right)^{n_{10}} \left(\frac{1}{1+\exp(\beta_0+\beta_1)}\right)^{n_{01}} \left(\frac{1}{1+\exp(\beta_0)}\right)^{n_{00}}$$

Let $K = \ln L(\beta_0,\beta_1)$. To find the maximum likelihood estimate, we simply need to find the maximum of K instead of $L(\beta_0,\beta_1)$.

$$K = n_{11}(\beta_0+\beta_1) + n_{10}(\beta_0) - (n_{11}+n_{01})\ln(1+\exp(\beta_0+\beta_1)) - (n_{10}+n_{00})\ln(1+\exp(\beta_0))$$

By taking the derivative of K with respect to both parameters β_0 and β_1, we have

$$
\begin{aligned}
\frac{dK}{d\beta_1} &= n_{11} - \frac{n_{11}+n_{01}}{1+\exp(\beta_0+\beta_1)}\exp(\beta_0+\beta_1) = 0 \\
&\Rightarrow \frac{n_{11}}{n_{11}+n_{01}} = \frac{\exp(\hat{\beta}_0+\hat{\beta}_1)}{1+\exp(\hat{\beta}_0+\hat{\beta}_1)} \qquad\qquad \Rightarrow (1) \\
\frac{dK}{d\beta_0} &= n_{11}+n_{10} - \frac{n_{11}+n_{01}}{1+\exp(\beta_0+\beta_1)}\exp(\beta_0+\beta_1) - \frac{n_{10}+n_{00}}{1+\exp(\beta_0)}\exp(\beta_0) \\
&= n_{11}+n_{10} - (n_{11}+n_{01})\frac{n_{11}}{n_{11}+n_{01}} - \frac{n_{10}+n_{00}}{1+\exp(\beta_0)}\exp(\beta_0) \\
&= n_{10} - \frac{n_{10}+n_{00}}{1+\exp(\beta_0)}\exp(\beta_0) = 0 \\
&\Rightarrow \frac{n_{10}}{n_{10}+n_{00}} = \frac{\exp(\hat{\beta}_0)}{1+\exp(\hat{\beta}_0)} \qquad\qquad \Rightarrow (2)
\end{aligned}
$$

From (1) and (2), we have

$$
\begin{aligned}
\exp(\hat{\beta}_0+\hat{\beta}_1) &= \frac{n_{11}}{n_{01}} \\
\exp(\hat{\beta}_0) &= \frac{n_{10}}{n_{00}}
\end{aligned}
$$

$$
\begin{aligned}
\Rightarrow \exp(\hat{\beta}_1) &= \frac{\exp(\hat{\beta}_0+\hat{\beta}_1)}{\exp(\hat{\beta}_0)} \\
e^{\hat{\beta}_1} &= \frac{\frac{n_{11}}{n_{01}}}{\frac{n_{10}}{n_{00}}} = \frac{n_{11}n_{00}}{n_{10}n_{01}}
\end{aligned}
$$

which is the odds ratio that we have calculated in part (b).