MTH 9899: Machine Learning
Bo Feng
Spring 2015

Problem Set # 4
Solution

1. Problem 1

   For a point (x, y) in the ROC plane, x represents false positive rate and y true positives rate. The profitable points are such that

   $$(TP - FN) \cdot P > FP \cdot L,$$

   where $TP = \#ProfitableTrades \cdot y, FN = \#ProfitableTrades \cdot (1-y)$ and $FP = \#UnprofitableTrades \cdot x$. Note that

   $$\frac{\#ProfitableTrades}{\#UnprofitableTrades} = H.$$

   Plugging these equations into the above inequality yields

   $$y > (\frac{1}{2H}\frac{L}{P})x + \frac{1}{2}.$$

   Hence, the profitable classifier lies above the line $y = (\frac{1}{2H}\frac{L}{P})x + \frac{1}{2}$. When $\frac{L}{P} \simeq 3$ and $H \simeq 1/2$, we have $y > 3x + 0.5$.
   *The plot is attached at the end of the homework.

2. Problem 2

   For ease of mathematical description, we use $O(XY), O(\bar{X}\bar{Y}), O(\bar{X}Y)$ and $O(X\bar{Y})$ to represent the observed number of occurrences of the minterms.

   a) Since $O(XY) = 200,000$, we have

   $$\begin{aligned} O(\bar{X}Y) &= O(Y) - O(XY) = 400,000 - 200,000 = 200,000 \\ O(X\bar{Y}) &= O(X) - O(XY) = 500,000 - 200,000 = 300,000 \\ O(\bar{X}\bar{Y}) &= 1,000,000 - O(X) - O(Y) + O(XY) = 300,000 \end{aligned}$$

   100% of the minterms are greater than 5. It obviously clear the Minterm Support threshold.

   b) Assuming independence, we have

   $$E(XY) = 200,000 \quad (1)$$
   $$E(\bar{X}Y) = 200,000 \quad (2)$$
   $$E(X\bar{Y}) = 300,000 \quad (3)$$
   $$E(\bar{X}\bar{Y}) = 300,000. \quad (4)$$

Hence the χ^2 test statistic under the null hypothesis is

$$\begin{aligned}\chi^2 &= \frac{(E(XY)-O(XY))^2}{E(XY)} + \frac{(E(\bar{X}Y)-O(\bar{X}Y))^2}{E(\bar{X}Y)} \\ &\quad + \frac{(E(X\bar{Y})-O(X\bar{Y}))^2}{E(X\bar{Y})} + \frac{(E\bar{X}\bar{Y})-O(\bar{X}\bar{Y}))^2}{E(\bar{X}\bar{Y})} \\ &= 0 < \chi_1^2(.95) = 3.84.\end{aligned}$$

It does not clear the significance level at 5%.

c) This is not an interesting Association Rule. However, in last homework, the Support-Confidence Framework mistakenly treated the result as an interesting Association Rule. The dependency rule here is more preferred.

d) After simple computation, we have $O(XY) = 2,000, O(\bar{X}Y) = 398,000, O(X\bar{Y}) = 498,000$ and $O(\bar{X}\bar{Y}) = 102,000$. 100% of the minterms are greater than 5, clearing the threshold as well.

e) Similar computation to b) yields

$$\begin{aligned}\chi^2 &= \frac{(200,000-2,000)^2}{200,000} + \frac{(200,000-398,000)^2}{200,000} \\ &\quad + \frac{(300,000-498,000)^2}{300,000} + \frac{(300,000-102,000)^2}{300,000} \\ &= 653400 \gg \chi_1^2(.95) = 3.84.\end{aligned}$$

f) This is an interesting Association Rule, and is also in contrast with the observation by Support-Confidence Framework.

3. Problem 3

   a) The External Closure of the χ^2 statistic taken together with the Internal Closure of Minterm Support define the Region of Interest. In other words, it's generated by the intersection of Σ and Δ, where Σ is the set of all Minterm Supported Itemsets and Δ the set of all Dependent Itemsets at the $1-\alpha$ confidence Level.

   b) The Dependency Strength Coefficient is not used as it requires evaluation for every Minterms, which would be an impossible task.

   c) There is an indefinite relation regarding the size of the Region of Interest and that of the Support-Confidence Framework. The former has internal closure in terms of Minterm support and external closure in terms of χ^2 statistic. The latter has internal closure in terms of support and external support of confidence. Depending on the specification of the support and confidence, these regions can have different sizes. Generally, the Region of Interest is smaller as it tends to impose strict restrictions on the itemsets.

4. Problem 4

   Applying the Chernoff Bound, we have

$$m_s \geq \frac{1}{2\epsilon^2}\ln(\frac{2}{\delta}) = \frac{1}{2\times 0.1\%^2}\ln(\frac{2}{0.01}) = 2649159.$$

5. Problem 5

   Suppose we are given a k-itemset $X = X_1 \cap ... \cap X_k$ that has support s, i.e., $P(X) \geq s$. Consider an l-itemset Y ($l < k$), without loss of generality, let $Y = X_1 \cap ... \cap X_l$. We then have $P(Y) \geq P(X) \geq s$.

Hence Support is Internally Closed. If we find an Itemset that's not supported, we can ignore all its super-itemsets by using this property. As the number of transactions that contain a certain itemset increases, the percentage of such transactions also increases, so Support is a Type I Property.

6. Problem 6

   To make the Framework profitable, we have

$$TP \cdot P > FP \cdot L + FN \cdot P.$$

   Here $TP = 1 - \alpha, FP = \alpha$ and $FN = \beta$, so

$$(1 - \alpha)P > \alpha L + \beta P.$$

   As a result, the power $1 - \beta$ satisfies

$$1 - \beta > \frac{P + L}{P}\alpha$$

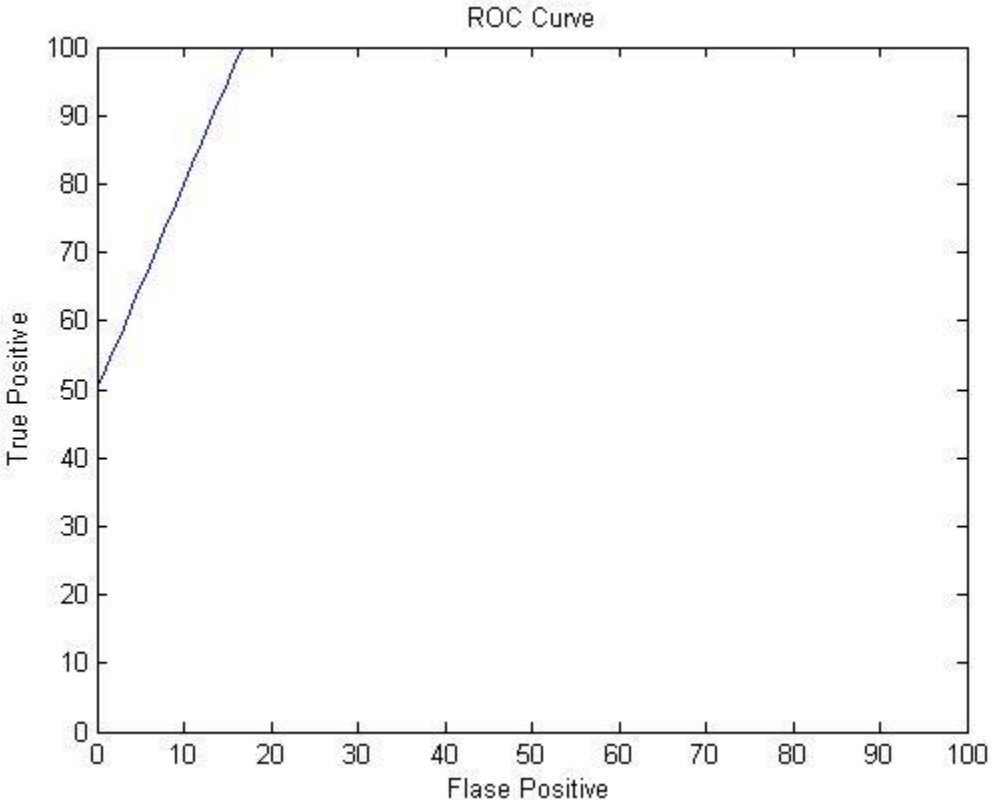
   The minimum power for our Framework to be profitable is $\frac{P+L}{P}\alpha$.

7. Problem 7

   With the Autoregression model, we could build a linear time series model for y_t.

   However, using network with TDL allows us to build a more general model for y_t. The model could be in non-linear form, and more regressors (such as x_t) could be included in the model.

Q1 ROC Plot


ROC Curve
True Positive
Flase Positive
0
10
20
30
40
50
60
70
80
90
100