Exam 2 Reference Keith G. Williams 800690755

# ROC Curves

| Predicted \ Observed | True | False |
|---|---|---|
| True | True Positive (TP) | False Positive (FP) |
| False | False Negative (FN) | True Negative (TN) |

**Sensitivity (TPF)** = $\frac{\text{TP}}{\text{TP+FN}}$ **FNF** = $\frac{\text{FN}}{\text{TP+FN}} = 1 - \text{TPF}$ **Specificity (TNF)** = $\frac{\text{TN}}{\text{FP+TN}}$ **FPF** = $\frac{\text{FP}}{\text{FP+TN}} = 1 - \text{TNF}$ **Accuracy** = $\frac{\text{TP+TN}}{N}$

*y-axis*: TPF, sensitivity *x-axis*: FPF, 1 – specificity Each point represents TPF/FPF corresponding to a particular decision threshold.

| | | |
|---|---|---|
| Strict threshold | low TPF, low FPF (high specificity) | P(D-)/P(D+) large, rare prevalence |
| Relaxed threshold | high TPF, high FPF (high sensitivity) | P(D-)/P(D+) small, common prevalence |

AUC [0.5, 1.0], AUC = 0.80 means a randomly selected individual from a diseased group has a test value larger than that for a randomly chosen individual from the non-diseased group 80% of the time.

Best operating point minimizes: $\overline{\text{cost}} = C_{\text{overhead}} + C_{\text{TP}} \cdot P(\text{TP}) + C_{\text{TN}} \cdot P(\text{TN}) + C_{\text{FP}} \cdot P(\text{FP}) + C_{\text{FN}} \cdot P(FN)$

| Advantages | Disadvantages |
|---|---|
| Simple, Graphical thus easily appreciated visually, comprehensive representation of pure accuracy, doesn't require selection of particular threshold, independent of prevalence, requires no binning of data, specificity and sensitivity are readily accessible | Thresholds are not displayed on the plot, the number of subjects is not shown, requires help of computer software in plot generation and calculations. |

# Decision Tables

Three Basic parts: (1) Causes, conditions, decision criteria (2) Effects stubs, actions that result from a given set of causes (possible) (3) Effects specify which effects are to appear for a given set of conditions (realized)

Steps to create a decision table:

1. List all causes in the decision table in order of importance, multi-valued causes last
2. Calculate the number of possible combinations ($2^n$ if all binary)
3. Fill columns with all possible combinations
4. Reduce test combinations (Use '-' for redundancies)
5. Check covered combinations (checksum below effects)
6. Add effects to the table

| **Causes** | **Values** | **1** | **2** | **3** | **4** | |
|---|---|---|---|---|---|---|
| Cause 1 | Y, N | Y | Y | Y | N | |
| Cause 2 | Y, N | Y | N | N | - | |
| Cause 3 | Y, N | - | Y | N | - | |
| **Effects** | | | | | | |
| Effect 1 | | X | | X | | |
| Effect 2 | | X | X | | | |
| **Checksum** | | 2 | 1 | 1 | 4 | = $2^3$ |

# Decision Trees

Entropy $= \sum_i -p_i \log_2 p_i$, where $p_i$ is the probability of *class i*. Higher entropy implies more information content. Range [0, 1]

Information Gain = $\text{Entropy}_{\text{parent node}}$ – Weighted Average $\text{Entropy}_{\text{children nodes}}$ Can also think of it as impurity reduced. (min $E_{ch}$)

Example: $N$ = 30, 14+, 16-, Children: $C_1$: $N_1$ = 13, 1+, 12-; $C_2$: $N_2$ = 17, 13+, 4-

$$IG = \left[-\left(\frac{14}{30}\log_2\frac{14}{30}\right) - \left(\frac{16}{30}\log_2\frac{16}{30}\right)\right] - \left[\frac{13}{30}\left(-\left(\frac{1}{13}\log_2\frac{1}{13}\right) - \left(\frac{12}{13}\log_2\frac{12}{13}\right)\right) + \frac{17}{30}\left(-\left(\frac{13}{17}\log_2\frac{13}{17}\right) - \left(\frac{13}{17}\log_2\frac{13}{17}\right)\right)\right]$$

IG prefers attributes with large number of values that split the data into small, pure subsets. Solution IC: *I(C;F)*

$I(C;F) = \sum_{i=1}^{m}\sum_{j=1}^{d} P(C=c_i, F=f_j)\log_2\frac{P(C=c_i,F=f_j)}{P(C=c_i)\cdot P(F=f_j)}$ with *m* classes and *d* features =CF/N * LOG(CF*N/(C*F),2)

# Association Rule Mining

Unsupervised clustering based on categorical inputs

A transaction *t* contains itemsets $X$ and $Y$: $X, Y \subseteq t$; An association rule: $X \rightarrow Y$: $X, Y \subset I$ and $X \cap Y = \emptyset$

Support: SUP in $T$ if SUP% contain X ∪ Y; $SUP = \Pr(X \cup Y) = \frac{(X \cup Y).count}{N_T}$

Confidence: CONF in $T$ if CONF% that contain X also contain Y; $CONF = \Pr(Y|X) = \frac{(X \cup Y).count}{X.count}$ and $LIFT = \frac{\Pr(Y|X)}{\Pr(Y)}$

Does not consider the quantity of each item purchased and the price paid.

Apriori algorithm utilizes downward closure all $i \subseteq freq\ i$ are also frequent and all supersets for infrequent $i$ are also infrequent.

1. Find all frequent itemsets (SUP ≥ minsup), starting with $k = 1$ to $k = \|I\|$
   a. $C_k$: k-itemsets given $F_{k-1}$
   b. $F_k \subseteq C_k$ where SUP ≥ minsup
2. Use frequent itemsets to generate rules, prune based on confidence.
   a. Let $B = X - A$, then A→B is an association rule iff CONF(A→B) ≥ minconf
   b. $2^k - 2$ rules for each k-itemset

Space of all association rules is $O(2^m)$, but b/c of sparseness, typically O(m)

If minsup too high, rules involving rare items will not be found.

If minsup too low, combinatorial explosion causes $O(2^m)$, therefore MIS allows for differential minsup for each item