Exam 2 Reference

ROC Curves

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

**Sensitivity (TPF)** = **FNF** = **Specificity (TNF)** = **FPF** = **Accuracy** =

*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:

|  |  |
| --- | --- |
| 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 | = 23 |

Decision Trees

, where *pi* is the probability of *class* *i*. Higher entropy implies more information content. Range [0, 1]

Information Gain = Entropyparent node – Weighted Average Entropychildren nodes Can also think of it as impurity reduced. (min Ech)

Example: *N* = 30, 14+, 16-, Children: C1: *N*1 = 13, 1+, 12-; C2: *N*2 = 17, 13+, 4-

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

with *m* classes and *d* features

Association Rule Mining

Unsupervised clustering based on categorical inputs

A transaction *t* contains itemsets *X* and *Y*: *X, Y* ⊆ *t* *;* An association rule: *X*→*Y*: *X, Y* ⊂*I* and *X* ⋂*Y* = ∅

Support: SUP in *T* if SUP% contain X ∪ Y;

Confidence: CONF in *T* if CONF% that contain X also contain Y; and

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

Apriori algorithm utilizes downward closure all *i* ⊆*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
   1. Ck: k-itemsets given Fk-1
   2. Fk ⊆ Ck where SUP ≥ minsup
2. Use frequent itemsets to generate rules, prune based on confidence.
   1. Let , then A→B is an association rule iff CONF(A→B) ≥ minconf
   2. 2k – 2 rules for each k-itemset

Space of all association rules is O(2m), 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(2m), therefore MIS allows for differential minsup for each item