Formulas and references

A Tour Through the Visualization Zoo – Summary of Graphic Types

Time-Series Data
- Index Charts
- Stacked Graphs
- Small Multiples
- Horizon Graphs

Statistical Distributions
- Stem-and-Leaf Plots
- Q-Q Plots
- SPLOM
- Parallel Coordinates

Maps
- Flow Maps
- Choropleth Maps
- Graduated Symbol Maps
- Cartograms

Hierarchies
- Node-Link diagrams
- Adjacency Diagrams
- Enclosure Diagrams

Networks
- Force-Directed Layouts
- Arc Diagrams
- Matrix Views

Entropy

If S is an arbitrary collection of examples with a binary class attribute, then:

$$Entropy(S) = -P_{C1}log_2(P_{C1}) - P_{C2}log_2(P_{C2})$$

$$= -\frac{N_{C1}}{N} log_2\left(\frac{N_{C1}}{N}\right) - \frac{N_{C2}}{N} log_2\left(\frac{N_{C2}}{N}\right)$$

where $C1$ and $C2$ are the two classes. P_{C1} and P_{C2} are the probability of being in Class 1 or Class 2 respectively. N_{C1} and N_{C2} are the number of examples in each class. N is the total number of examples.

Note: $log_2 x = \frac{log_{10} x}{log_{10} 2} = \frac{log_{10} x}{0.301}$

Information gain

The $Gain(S, A)$ of an attribute A relative to a collection of examples, S, with v groups having $|S_v|$ elements is:

$$Gain(S, A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} * Entropy(S_v)$$

Networking

Closeness Centrality: $C_{CL}(v) = \frac{1}{\sum_{u \in V} dist(u,v)}$

Betweenness Centrality: $C_B(v) = \sum_{s \neq t \neq v \in V} \frac{\sigma(s,t|v)}{\sigma(s,t)}$,

where (s, t) is the number of shortest paths between s and t.
$(s, t|v)$ is the number of shortest paths between s and t passing through v

Density: $den(g) = \frac{|E_g|}{|V_g|(|V_g|-1)/2}$,

where $|E_g|$ is number of edges, $|V_g|$ is number of vertices

Clustering coefficient: $clt(g) = \frac{3\tau_{\Delta}(g)}{\tau 3(g)}$,

where $3\tau_{\Delta}(g)$ is number of triangles, $\tau 3(g)$ is number of connected triples

Naïve Bayes'

For events $A_1, A_2, \dots, A_n$ *and event* C, classification probability is

$$P(C_j | A_1 \cap A_2 \dots \cap A_n) = \frac{P(C_j) \cdot P(A_1 \cap A_2 \dots \cap A_n | C_j)}{P(A_1 \cap A_2 \dots \cap A_n)}$$

For Bayesian classification, a new point is classified to C_j if $P(C_j) * P(A_1|C_j) * P(A_1|C_j) * \dots * P(A_n|C_j)$ is maximised.

Naïve Bayes assumes $P(A \cap B) = P(A) * P(B)$ etc.

Cosine or normalised dot product

For documents i and j with terms w

$$Sim(D_i, D_j) = \frac{\sum_{t=1}^{N} w_{it} * w_{jt}}{\sqrt{\sum_{t=1}^{N}(w_{it})^2 * \sum_{t=1}^{N}(w_{jt})^2}}$$

ROC

$$TPR = \frac{TP}{TP + FN}, \quad FPR = \frac{FP}{FP + TN}$$