A First Course in Network Theory
Comparing Partitionings

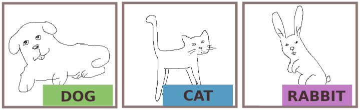
Luce le Gorrec, Philip Knight, Francesca Arrigo

University of Strathclyde, Glasgow

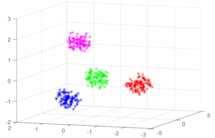
Evaluating Clustering Algorithms

Supervised Learning (Classification) VS Unsupervised Learning (Clustering)

Find the label of a data



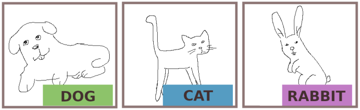
Find groups of similar data



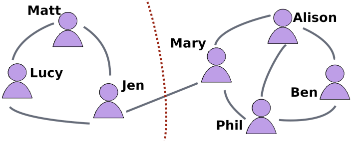
Evaluating Clustering Algorithms

Supervised Learning (Classification) VS Unsupervised Learning (Clustering)

Find the label of a data



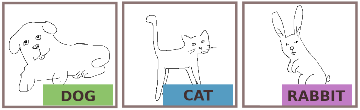
Find groups of similar data



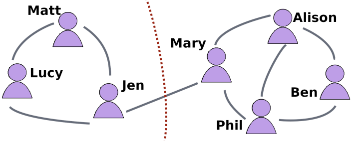
Evaluating Clustering Algorithms

Supervised Learning (Classification) VS Unsupervised Learning (Clustering)

Find the label of a data



Find groups of similar data



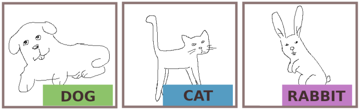
To assess the quality of algorithms...

Counting the errors (bad labels).

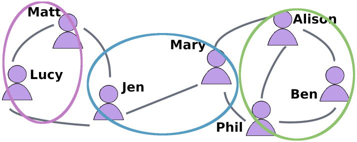
Evaluating Clustering Algorithms

Supervised Learning (Classification) VS Unsupervised Learning (Clustering)

Find the label of a data



Find groups of similar data



To assess the quality of algorithms...

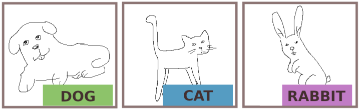
Counting the errors (bad labels).

What is an error ?

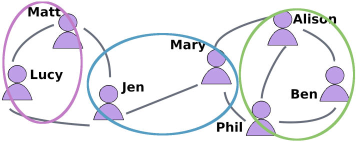
Evaluating Clustering Algorithms

Supervised Learning (Classification) VS Unsupervised Learning (Clustering)

Find the label of a data



Find groups of similar data



To assess the quality of algorithms...

Counting the errors (bad labels).

What is an error ?

$\mathcal{C} = \{C_1, ..., C_p\}$, $\mathcal{K} = \{K_1, ..., K_q\}$ are two partitionings on $\{1, ..., n\} = V$.

Agreement/Disagreement Table

$$\mathbf{N} = \begin{vmatrix} n_{1,1} & n_{1,0} \\ \\ n_{0,1} & n_{0,0} \end{vmatrix}$$

with

$$(TruePositives) \;\; n_{1,1} = \left| \left\{ (u,v) \in V \times V, u \neq v : \exists i, j \;\text{ with } u, v \in C_i \cap K_j \right\} \right|$$

$$(TrueNegatives) \;\; n_{0,0} = \left| \left\{ (u,v) \in V \times V, u \neq v : \exists i \neq i', \exists j \neq j', \;\text{ with } \begin{cases} u \in C_i \cap K_j \\ v \in C_{i'} \cap K_{j'} \end{cases} \right\} \right|$$

$$(FalseNegatives) \;\; n_{1,0} = \left| \left\{ (u,v) \in V \times V, u \neq v : \exists i, \exists j \neq j', \;\text{ with } \begin{cases} u \in C_i \cap K_j \\ v \in C_i \cap K_{j'} \end{cases} \right\} \right|$$

$$(FalsePositives) \;\; n_{0,1} = \left| \left\{ (u,v) \in V \times V, u \neq v : \exists i \neq i', \exists j, \;\text{ with } \begin{cases} u \in C_i \cap K_j \\ v \in C_{i'} \cap K_j \end{cases} \right\} \right|$$

Agreement/Disagreement Table

$$\mathbf{N} = \left| \begin{array}{cc|c} n_{1,1} & n_{1,0} & n_{1,1} + n_{1,0} = \sum_{i=1}^{p} \binom{|C_i|}{2} \\ n_{0,1} & n_{0,0} & n_{0,1} + n_{0,0} = \sum_{i \neq j} |C_i| \times |C_j| \end{array} \right.$$

with

$$(TruePositives) \quad n_{1,1} = \left| \left\{ (u,v) \in V \times V, u \neq v : \exists i,j \ \text{ with } u,v \in C_i \cap K_j \right\} \right|$$

$$(TrueNegatives) \quad n_{0,0} = \left| \left\{ (u,v) \in V \times V, u \neq v : \exists i \neq i', \exists j \neq j', \ \text{ with } \begin{cases} u \in C_i \cap K_j \\ v \in C_{i'} \cap K_{j'} \end{cases} \right\} \right|$$

$$(FalseNegatives) \quad n_{1,0} = \left| \left\{ (u,v) \in V \times V, u \neq v : \exists i, \exists j \neq j', \ \text{ with } \begin{cases} u \in C_i \cap K_j \\ v \in C_i \cap K_{j'} \end{cases} \right\} \right|$$

$$(FalsePositives) \quad n_{0,1} = \left| \left\{ (u,v) \in V \times V, u \neq v : \exists i \neq i', \exists j, \ \text{ with } \begin{cases} u \in C_i \cap K_j \\ v \in C_{i'} \cap K_j \end{cases} \right\} \right|$$

Agreement/Disagreement Table

$$\mathbf{N} = \left|\begin{array}{cc} n_{1,1} & n_{1,0} \\ n_{0,1} & n_{0,0} \end{array}\right| \begin{array}{l} n_{1,1} + n_{1,0} = \sum_{i=1}^{p} \binom{|C_i|}{2} \\ n_{0,1} + n_{0,0} = \sum_{i \neq j} |C_i| \times |C_j| \end{array}$$

with

$$(TruePositives) \;\; n_{1,1} = \left|\left\{(u,v) \in V \times V, u \neq v : \exists i,j \;\; \text{with } u,v \in C_i \cap K_j \right\}\right|$$

$$(TrueNegatives) \;\; n_{0,0} = \left|\left\{(u,v) \in V \times V, u \neq v : \exists i \neq i', \exists j \neq j', \;\; \text{with } \begin{cases} u \in C_i \cap K_j \\ v \in C_{i'} \cap K_{j'} \end{cases} \right\}\right|$$

$$(FalseNegatives) \;\; n_{1,0} = \left|\left\{(u,v) \in V \times V, u \neq v : \exists i, \exists j \neq j', \;\; \text{with } \begin{cases} u \in C_i \cap K_j \\ v \in C_i \cap K_{j'} \end{cases} \right\}\right|$$

$$(FalsePositives) \;\; n_{0,1} = \left|\left\{(u,v) \in V \times V, u \neq v : \exists i \neq i', \exists j, \;\; \text{with } \begin{cases} u \in C_i \cap K_j \\ v \in C_{i'} \cap K_j \end{cases} \right\}\right|$$

$$\text{Rand Index } RI(\mathcal{C}, \mathcal{K}) = \frac{n_{1,1} + n_{0,0}}{n_{1,1} + n_{0,0} + n_{1,0} + n_{0,1}}$$

Confusion Table

$$\mathbf{T} = \frac{1}{n}\begin{bmatrix} |C_1 \cap K_1| & \dots & |C_1 \cap K_q| \\ \vdots & \ddots & \vdots \\ |C_p \cap K_1| & \dots & |C_p \cap K_q| \end{bmatrix} \qquad \begin{cases} \sum_{j=1}^{q} \mathbf{T}(i,j) = \dfrac{|C_i|}{n} \\ \sum_{i=1}^{p} \mathbf{T}(i,j) = \dfrac{|K_j|}{n} \end{cases}$$

$$(1) \begin{cases} \text{Probability for a node } u \text{ to lie in a cluster } C_i \in \mathcal{C}: & Pr(u \in C_i) = |C_i|/n \\ \text{Probability for a node } u \text{ to lie in a cluster } K_j \in \mathcal{K}: & Pr(u \in K_j) = |K_j|/n \end{cases}$$

Confusion Table

$$\mathbf{T} = \frac{1}{n}\begin{bmatrix} |C_1 \cap K_1| & \dots & |C_1 \cap K_q| \\ \vdots & \ddots & \vdots \\ |C_p \cap K_1| & \dots & |C_p \cap K_q| \end{bmatrix} \qquad \begin{cases} \sum_{j=1}^{q} \mathbf{T}(i,j) = \frac{|C_i|}{n} \\ \sum_{i=1}^{p} \mathbf{T}(i,j) = \frac{|K_j|}{n} \end{cases}$$

$$(1) \begin{cases} \text{Probability for a node } u \text{ to lie in a cluster } C_i \in \mathcal{C}: & Pr(u \in C_i) = |C_i|/n \\ \text{Probability for a node } u \text{ to lie in a cluster } K_j \in \mathcal{K}: & Pr(u \in K_j) = |K_j|/n \end{cases}$$

Entropy $H(X)$ of a variable X is its uncertainty:

$$H(X) = -\sum_{x \in \mathcal{X}} Pr(X = x) \times log_2(Pr(X = x)).$$

Confusion Table

$$\mathbf{T} = \frac{1}{n}\begin{bmatrix} |C_1 \cap K_1| & \dots & |C_1 \cap K_q| \\ \vdots & \ddots & \vdots \\ |C_p \cap K_1| & \dots & |C_p \cap K_q| \end{bmatrix} \qquad \begin{cases} \sum_{j=1}^{q} \mathbf{T}(i,j) = \dfrac{|C_i|}{n} \\ \sum_{i=1}^{p} \mathbf{T}(i,j) = \dfrac{|K_j|}{n} \end{cases}$$

$$(1) \begin{cases} \text{Probability for a node } u \text{ to lie in a cluster } C_i \in \mathcal{C}: & Pr(u \in C_i) = |C_i|/n \\ \text{Probability for a node } u \text{ to lie in a cluster } K_j \in \mathcal{K}: & Pr(u \in K_j) = |K_j|/n \end{cases}$$

Entropy $H(X)$ of a variable X is its uncertainty:

$$H(X) = -\sum_{x \in \mathcal{X}} Pr(X = x) \times log_2(Pr(X = x)).$$

Mutual Info $MI(X, Y)$ is the reduction in X uncertainty due to knowing Y:

$$MI(X,Y) = H(X) - H(X|Y) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} Pr(x,y) log_2(\frac{Pr(x,y)}{Pr(x)Pr(y)})$$

Confusion Table

$$\mathbf{T} = \frac{1}{n}\begin{bmatrix} |C_1 \cap K_1| & \dots & |C_1 \cap K_q| \\ \vdots & \ddots & \vdots \\ |C_p \cap K_1| & \dots & |C_p \cap K_q| \end{bmatrix} \quad \begin{cases} \sum_{j=1}^{q} \mathbf{T}(i,j) = \dfrac{|C_i|}{n} \\ \sum_{i=1}^{p} \mathbf{T}(i,j) = \dfrac{|K_j|}{n} \end{cases}$$

$$(1) \begin{cases} \text{Probability for a node } u \text{ to lie in a cluster } C_i \in \mathcal{C}: & Pr(u \in C_i) = |C_i|/n \\ \text{Probability for a node } u \text{ to lie in a cluster } K_j \in \mathcal{K}: & Pr(u \in K_j) = |K_j|/n \end{cases}$$

Entropy $H(X)$ of a variable X is its uncertainty:

$$H(X) = -\sum_{x \in \mathcal{X}} Pr(X = x) \times log_2(Pr(X = x)).$$

Mutual Info $MI(X, Y)$ is the reduction in X uncertainty due to knowing Y:

$$MI(X,Y) = H(X) - H(X|Y) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} Pr(x,y) log_2(\frac{Pr(x,y)}{Pr(x)Pr(y)})$$

Using (1) and the confusion table $\mathbf{T}$:

$$MI(\mathcal{C}, \mathcal{K}) = \sum_{i=1}^{p} \sum_{j=1}^{q} \mathbf{T}(i,j) log_2 \left(\frac{n^2 \times \mathbf{T}(i,j)}{|C_i| \times |K_j|} \right)$$

Adjusted for Chance

An index $Ind(\mathcal{C}, \mathcal{K})$ can be **adjusted for chance**

$$AInd(\mathcal{C}, \mathcal{K}) = \frac{Ind(\mathcal{C}, \mathcal{K}) - \mathbb{E}[Ind(X, Y)]}{max(Ind(X, Y)) - \mathbb{E}[Ind(X, Y)]}$$

$\implies AInd(\mathcal{C}, \mathcal{K}) \approx 0$ when $\mathcal{C}$, $\mathcal{K}$ are independent.

Adjusted for Chance

An index $Ind(\mathcal{C}, \mathcal{K})$ can be **adjusted for chance**

$$AInd(\mathcal{C}, \mathcal{K}) = \frac{Ind(\mathcal{C}, \mathcal{K}) - \mathbb{E}[Ind(X, Y)]}{max(Ind(X, Y)) - \mathbb{E}[Ind(X, Y)]}$$

$\Longrightarrow$ $AInd(\mathcal{C}, \mathcal{K}) \approx 0$ when $\mathcal{C}$, $\mathcal{K}$ are independent.

⚠ Requires to select random model.

✗ Not always easy to derive, and can be computationally awful.

$$E[\mathrm{MI}(U, V)] = \sum_{i=1}^{|U|} \sum_{j=1}^{|V|} \sum_{n_{ij}=(a_i+b_j-N)^+}^{\min(a_i, b_j)} \frac{n_{ij}}{N} \log \left(\frac{N.\, n_{ij}}{a_i b_j} \right) \frac{a_i! b_j! (N - a_i)! (N - b_j)!}{N! n_{ij}! (a_i - n_{ij})! (b_j - n_{ij})! (N - a_i - b_j + n_{ij})!}$$

Adjusted for Chance

An index $Ind(\mathcal{C}, \mathcal{K})$ can be **adjusted for chance**

$$AInd(\mathcal{C}, \mathcal{K}) = \frac{Ind(\mathcal{C}, \mathcal{K}) - \mathbb{E}[Ind(X, Y)]}{max(Ind(X, Y)) - \mathbb{E}[Ind(X, Y)]}$$

$\implies AInd(\mathcal{C}, \mathcal{K}) \approx 0$ when $\mathcal{C}$, $\mathcal{K}$ are independent.

⚠ Requires to select random model.

✗ Not always easy to derive, and can be computationally awful.

For the Rand Index, choice of the Permutation Model gives:

$$ARI(\mathcal{C}, \mathcal{K}) = \frac{2(n_{0,0}n_{1,1} - n_{0,1}n_{1,0})}{(n_{0,0} + n_{0,1})(n_{1,1} + n_{0,1}) + (n_{0,0} + n_{1,0})(n_{1,1} + n_{1,0})}$$