Mutual Information feature selection

X. V. Nguyen, J. Chan, S. Romano, and J. Bailey, "Effective global approaches for mutual information based feature selection", https://dl.acm.org/citation.cfm?id=2623611

Summary

1. The general Minimum Redundancy Maximum Relevance framework (MRMR)
2. Quadratic Programming Feature Selection (QPFS)
3. The issues with QPFS
4. Extended QPFS
5. EQPFS as a Global Subset Selection Problem
6. EQPFS via semi-definite programming
7. Implementation with D-Wave

MRMR Family

$$\mathcal{MRMR}: \max_{X_i \in \mathbb{X} \setminus \mathbb{S}} \left\{ I(X_i; C) - \alpha \sum_{X_j \in \mathbb{S}} I(X_i; X_j) \right\}$$

- Greedy scheme
- MIFS → α = 1

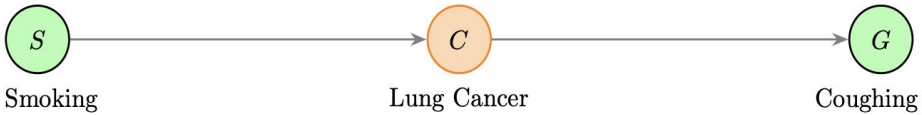
QPFS

$$\mathcal{QPFS} : \min_{\mathbf{x}} \left\{ \alpha \mathbf{x}^T \mathbf{H} \mathbf{x} - \mathbf{x}^T \mathbf{f} \right\} \text{ s.t. } \sum_{i=1}^{n} x_i = 1, x_i \geq 0$$

- The vector of relevancy is $\mathbf{f}_{n \times 1} = [I(X_1; C), \ldots, I(X_n; C)]^T$
- The pairwise matrix of redundancy is $\mathbf{H}_{n \times n} = [I(X_i; X_j)]_{i,j=1 \ldots n}$

Issues of QPFS

- H has to be positive so that QPFS is a convex quadratic program
  - the mutual information function on the space of features must be a proper kernel function.
- H penalizes features for their redundancy, and this should not happen.
  - if we put Hii = 0, then the Hessian matrix H becomes indefinite, violating the previous statement



P(S,C)	C=0	C=1
S=1	0.25	0
S=2	0.25	0
S=3	0	0.25
S=4	0	0.25

P(C,G)	C=0	C=1
G=0	0.475	0.025
G=1	0.025	0.475

P(S,G)	G=0	G=1
S=1	0.2375	0.0125
S=2	0.2375	0.0125
S=3	0.0125	0.2375
S=4	0.0125	0.2375

Extended QPFS

- Problem was: the self redundancy is composed by two terms:
  - Unconditional Redundancy
  - Class Conditional Redundancy

$$\mathcal{EMRMR}: \max_{X_i} \left\{ I(X_i;C) - \alpha \sum_{X_j \in \mathbb{S}} [I(X_i;X_j) - I(X_i;X_j|C)] \right\}$$

Extended QPFS

- The self redundancy is composed by two terms → Conditional Mutual Information:
  - Unconditional Redundancy
  - Class Conditional Redundancy

$$\mathcal{EMRMR}: \max_{X_i} \left\{ I(X_i;C) - \alpha \sum_{X_j \in \mathbb{S}} [I(X_i;X_j) - I(X_i;X_j|C)] \right\}$$

$$\mathcal{EQPFS} \quad : \quad \min_{\mathbf{x}} \left\{ \alpha \mathbf{x}^T [\mathbf{H}_1 - \mathbf{H}_2] \mathbf{x} - \mathbf{x}^T \mathbf{f} \right\}$$

$$\text{s.t.} \qquad \sum_{i=1}^{n} x_i = 1, x_i \geq 0$$

EQPFS as a Global Subset Selection Problem

$$\mathcal{SS}_{\mathcal{CMJ}}: \max_{\substack{\mathbb{S}\subset\mathbb{X} \\ |\mathbb{S}|=k}} \left\{ \sum_{X_i \in \mathbb{S}} I(X_i;C) + \sum_{X_i, X_j \in \mathbb{S}} I(X_i;C|X_j) \right\}$$

$$\downarrow$$

$$\mathcal{QJP}_{\mathcal{CMJ}}: \max_{\mathbf{x}} \left\{ \mathbf{x}^T \mathbf{Q} \mathbf{x} \right\} \text{ s.t. } \mathbf{x} \in \{0,1\}^n, \|\mathbf{x}\| = \sqrt{k}$$

Where $\mathbf{Q}_{ii} = I(X_i;C)$ and $\mathbf{Q}_{ij} = I(X_i;C|X_j), i \neq j$

EQPFS via semi-definite programming

- Semi-definite relaxation can be approximated using the same technique Goemans and Williamson used in approximating the NP-hard max-cut problem in graph theory

$$\mathcal{MAXCUT}: \max_{\mathbf{x}} \{\mathbf{x}^T\mathbf{L}\mathbf{x}\} \quad \text{s.t. } \mathbf{x} \in \{-1,+1\}^n$$

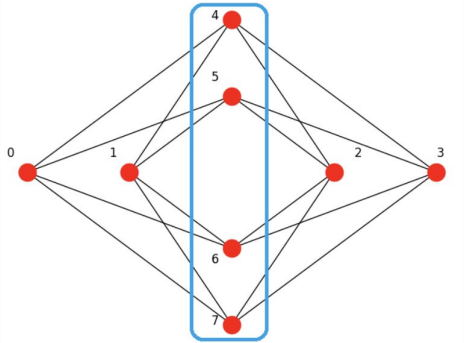
$$\downarrow$$

$$\begin{aligned} \max_{\mathbf{y}} \quad & \left\{\frac{1}{4}(\mathbf{y}+\mathbf{1})^T\mathbf{Q}(\mathbf{y}+\mathbf{1})\right\} \\ \text{s.t.} \quad & \mathbf{y} \in \{-1,1\}^n, (\mathbf{y}+\mathbf{1})^T\mathbf{I}(\mathbf{y}+\mathbf{1}) = 4k \end{aligned}$$

MAXCUT & D-Wave

Maximum Cut

A maximum cut is a subset of a graph's vertices such that the number of edges between this subset and the remaining vertices is as large as possible.



Maximum cut for a Chimera unit cell: the blue line around the subset of nodes {4, 5, 6, 7} cuts 16 edges; adding or removing a node decreases the number of edges between the two complementary subsets of the graph.

References

- https://cloud.dwavesys.com/leap/example-details/222058260
- https://dl.acm.org/doi/10.1145/2623330.2623611