BIOSTAT 527

5/10/23

Lecture 14

(finish RFF)

Comparing Clusterings

Stability based clustering evaluation

Kernel machine with RFF algorithm

Random Fourier Features

$z \in \mathbb{R}^d \leftrightarrow \omega \in \mathbb{R}^d$

$x \leftrightarrow \omega \quad d=1$

$z = x - x' \Leftrightarrow K(x-x')$

want to approx K in D dimensions $\equiv$ feature map

In Data $x^{1:n}, y^{1:n}$, kernel K

1. Fourier transform $p(\omega) = \frac{1}{2\pi}\int_{\mathbb{R}^d} e^{-i\omega^T z} K(z) dz$.
2. Choose D.
3. Sample $\omega_{1:D}$ i.i.d. from $\boxed{p}$. Sample $\omega_{0,1:D}$ uniformly from $[0, 2\pi]$.
4. Map data to features $\tilde{\phi}(x^i) = \sqrt{\frac{2}{D}}[\cos(\omega_j^T x^i + \omega_{0,j})]_{j=1:D}$ for all $i = 1:n$.
5. Solve SVM Primal problem; obtain $w \in \mathbb{R}^D$ and intercept $b \in \mathbb{R}$.

$x^i \longmapsto \tilde{\varphi}(x^i) \in \mathbb{R}^D$

$\sum_j \tilde{\varphi}_j(x^i)\tilde{\varphi}_j(x^{i'}) = \frac{1}{D}\sum_j e^{-i\omega_j^T(x^i - x^{i'})}$

$$\min_{w \in \mathbb{R}^D, b, \xi_{1:n}} \frac{1}{2}\|w\|^2 + C\sum \xi_i$$

$$y^i(w^T\tilde{\varphi}(x^i) + b) \geq 1 - \xi_i$$

$$\xi_{1:n} \geq 0$$

Why?

Δ, Δ' clusterings of $\mathcal{D}$

$d(\Delta, \Delta')$ $|\mathcal{D}| = n$

- compare with Δ^* true clustering

Lecture IV – ~~Hierarchical clustering~~. Comparing clusterings

- compare methods on $\mathcal{D}$
- estimate variability / robustness of 1 method

Marina Meilă
mmp@stat.washington.edu

Department of Statistics
University of Washington

STAT/BIOST 527

STAT/BIOST 527

Requirements for a distance

metric $d(\ ) \geq 0$ → identity

index $i(\ ) \in [0,1]$ identity

Depend on the application

- Applies to any two partitions of the same data set ← indep of cluster labels
- Makes no assumptions about how the clusterings are obtained
- Values of the distance between two pairs of clusterings comparable under the weakest possible assumptions
- Metric (triangle inequality) desirable
- **understandable, interpretable**

$\boxed{\text{partition} \equiv \text{clustering}\ \Delta}$

$$\Delta = (C_1, \ldots C_K)$$

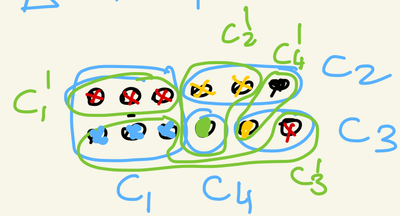
$$\Delta' = (C'_1, \ldots C'_{K'})$$

$$\bigcup_{k=1}^{K} C_k = \bigcup_{k'=1}^{K'} C'_{k'} = \{1, \ldots n\} \equiv \mathcal{D}$$

denote $m_{kk'} = |C_k \cap C'_{k'}| \Rightarrow \boxed{\sum_{k,k'} m_{kk'} = n}$

Marina Meila:

$\Delta \quad K=4$



$\Delta' \quad K'=4$

$\Delta = (C_1, \dots C_K)$

$\Delta' = (C_1', \dots C_{K'}')$

$$\dot{\bigcup_{k=1}} C_k = \dot{\bigcup_{k=1}^{K'}} C_{k'}' = \{1, \dots n\} \equiv \mathcal{D}$$

denote $m_{kk'} = |C_k \cap C_{k'}'| \Rightarrow \boxed{\sum_{k,k'} m_{kk'} = n}$

$m_{11} = |C_1 \cap C_1'| = 3$

$m_{22} = |C_2 \cap C_2'| = 2$

$|C_2 \cap C_1'| = m_{21} = 0$

$m_{13} = \underline{3}$

$m_{4,2} = 1$

$m_{24} = 1$

$m_{34} = 1$

$m_{33} = \underline{1}$

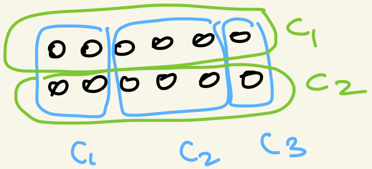
$$M = \begin{array}{c|cccc} & C_1' & C_2' & C_3' & C_4' \\ \hline C_1 & 3 & & 3 & \\ C_2 & & 2 & & 1 \\ C_3 & & & 1 & 1 \\ C_4 & & 1 & & \end{array} \quad \begin{array}{l} \rightarrow |C_1| = n_1 \\ \rightarrow n_2 = 3 \\ \rightarrow n_3 = 2 \\ \rightarrow n_4 = 1 \end{array}$$

$n_1' = 3 \quad n_2' = 3 \quad w$

$\Delta = \Delta'$ as clusterings

$$M = \begin{bmatrix} 0 & 3 & & \\ 3 & 0 & & \\ & & 2 & \\ & & & 1 \end{bmatrix}$$

$$m_{kk'} = \begin{cases} n_k \\ 0 \end{cases}$$



$m_{kk'} > 0$ for all kk'

$\leftarrow$ maximal distance Δ to Δ'

$$M = \begin{array}{cc} C_1' & C_2' \end{array}$$
$$M = \begin{bmatrix} 2 & 2 \\ 3 & 3 \\ 1 & 1 \end{bmatrix} \begin{array}{l} C_1 \\ C_2 \\ C_3 \end{array} \rightarrow P = \begin{bmatrix} \frac{1}{6} & \frac{1}{6} \\ 1/4 & 1/4 \\ 1/12 & 1/12 \end{bmatrix} \rightarrow p_{kk'} = \boxed{p_{xx'} = p_k \cdot p_{k'}}$$

$$p_{kk'} = \frac{m_{kk'}}{n} = \underbrace{p_{xx'}}$$

$\|$

$x \in \{1:K\}$

$x' \in \{1:K'\}$

joint distribution

marginal

$$\boxed{p_x} = Pr[x = k] = \boxed{\frac{n_k}{n}}$$

$$\boxed{p_{x'} = \frac{n'_{k'}}{n}}$$

The confusion matrix

- Let $\Delta = \{C_{1:K}\},\ \Delta' = \{C'_{1:K'}\}$
- Define $n_k = |C_k|,\ n'_{k'} = |C'_{k'}|$
- $m_{kk'} \ = |C_k \cap C'_{k'}|,\ k = 1:K, k' = 1:K'$
- note: $\sum_k m_{kk'} = n'_{k'},\ \sum_{k'} m_{kk'} = n_k,\ \sum_{k,k'} m_{kk'} = n$
- The **confusion matrix** $M \in \mathbb{R}^{K \times K'}$ is

$$M \ = \ [m_{kk'}]_{k=1:K}^{k'=1:K'}$$

- all distances and comparison criteria are based on M
- the **normalized confusion matrix** $P = M/n$

$$p_{kk'} \ = \ \frac{m_{kk'}}{n}$$

- The **normalized cluster sizes** $p_k = n_k/n,\ p'_{k'} = n'_{k'}/n$ are the **marginals** of P

$$p_k \ = \sum_{k'} p_{kk'} \quad p_{k'} \ = \sum_k p_{kk'}$$

Matrix Representations

- matrix reprentations for Δ
  - unnormalized (redundant) representation

$$\tilde{X}_{ik} = \begin{cases} 1 & i \in C_k \\ 0 & i \notin C_k \end{cases} \quad \text{for } i = 1:n, k = 1:K$$

  - normalized (redundant) representation

$$X_{ik} = \begin{cases} 1/\sqrt{|C_k|} & i \in C_k \\ 0 & i \notin C_k \end{cases} \quad \text{for } i = 1:n, k = 1:K$$

    therefore $X_k^T X_{k'} = \delta(k, k')$, X orthogonal matrix
    X_k = column k of X
  - normalized non-redundant reprentation
    - X_K is determined by $X_{1:K-1}$
    - hence we can use $Y \in \mathbb{R}^{n\times(K-1)}$ orthogonal representation
    - intuition: Y represents a subspace (is an orthogonal basis)
    - K centers in $\mathbb{R}^d$, $d \geq K$ determine a $K-1$ dimesional subspace plus a translation

The Misclassification Error (ME) distance

- Define the **Misclassification Error (ME)** distance d_{ME}

$$d_{ME} = 1 - \max_{\pi} \sum_{k=1}^{K} p_{k,\pi(k)} \quad \pi \in \{\text{all } K\text{−permutations}\},\ K \leq K' \text{w.l.o.g}$$

- Interpretation: treat the clusterings as classifications, then minimize the classification error over all possible label matchings
- Or: nd_{ME} is the Hamming distance between the vectors of labels, minimized over all possible label matchings
- can be computed in polynomial time by **Max bipartite matching** algorithm (also known as Hungarian algorithm)
- Is a metric: symmetric, ≥ 0, triangle inequality

$$d_{ME}(\Delta_1, \Delta_2) + d_{ME}(\Delta_1, \Delta_3) \geq d_{ME}(\Delta_2, \Delta_3)$$

- easy to understand (very popular in computer science)
- $d_{ME} \leq 1 - 1/K$
- bad: if clusterings not similar, or K large, d_{ME} is coarse/indiscriminative
- recommended: for small K

I Misclassification Error

$$M = \begin{bmatrix} n_1 & & 0 \\ & n_2 & \\ 0 & & \ddots & n_K \end{bmatrix} \Rightarrow d_{ME}(\Delta, \Delta') = 0$$

$$m_{kk'} = \begin{cases} n_k & k = k' \\ 0 & \text{otherwise} \end{cases}$$

$$d_{ME}(P) = 1 - \max_{\pi} \sum_{k=1}^{K} P_{k\pi(k')}$$

Δ, Δ'

π : permutations of K'

permute columns to maximize diagonal (P)

Max Bipartite matching Alg

$$M = \begin{array}{c|cccc} & C_1' & C_2' & C_3' & C_4' \\ \hline C_1 & 3 & & 4 & \\ C_2 & & 2 & & 1 \\ C_3 & & & 1 & 1 \\ C_4 & & 1 & & \end{array}$$

columns permuted

$$\begin{bmatrix} 4 & & & \\ & 2 & 1 & \\ 1 & & 1 & \\ & 1 & & \end{bmatrix}$$

$$\Rightarrow d_{ME} = 1 - \frac{4+2+1}{13} = \frac{6}{13}$$

d_{ME} = metric interpretable

$1 - d_{ME}$ = index

STAT/BIOST 527

The Variation of Information (VI) distance
Clusterings as random variables

$\mathcal{I}$ Mutual information

- Imagine points in $\mathcal{D}$ are picked randomly, with equal probabilities
- Then $k(i), k'(j)$ are random variables
  with $Pr[k] = p_k$, $Pr[k, k'] = p_{kk'}$

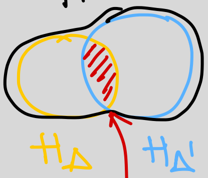
Marina Meila:

Incursion in information theory

$H_{\Delta\Delta'} \leq H_\Delta + H_{\Delta'}$

- **Entropy** of a random variable/clustering $H_\Delta = -\sum_k p_k \ln p_k$
- $0 \leq H_\Delta \leq \ln K$
- Measures uncertainty in a distribution (amount of randomness)
- **Joint entropy** of two clusterings

$$H_{\Delta,\Delta'} = -\sum_{k,k'} p_{kk'} \ln p_{kk'}$$

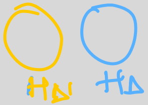


- $H_{\Delta',\Delta} \leq H_\Delta + H_{\Delta'}$ with equality when the two random variables are independent
- **Conditional entropy** of Δ' given Δ

$$H_{\Delta'|\Delta} = -\sum_k p_k \sum_{k'} \frac{p_{kk'}}{p_k} \ln \frac{p_{kk'}}{p_k}$$

- Measures the expected uncertainty about k' when k is known
- $H_{\Delta'|\Delta} \leq H_{\Delta'}$ with equality when the two random variables are independent
- **Mutual information** between two clusterings (or random variables)

independence



$$\begin{aligned} I_{\Delta,\Delta'} &= H_\Delta + H_{\Delta'} - H_{\Delta',\Delta} \\ &= H_{\Delta'} - H_{\Delta'|\Delta} \end{aligned}$$

Incursion in information theory (2)

- Measures the amount of information of one r.v. about the other
- $I_{\Delta,\Delta} \geq 0$, symmetric. Equality iff r.v.'s independent

mutual info

index $\in [0,1]$

$$NMI = \frac{I_{\Delta,\Delta'}}{H_{\Delta,\Delta'}}$$

normalized M.I.

(total) variation of information

$$VI = H_{\Delta\Delta'} - I_{\Delta,\Delta'} \geq 0$$

metric !!

The VI distance

- Define the **Variation of Information (VI)** distance

$$
\begin{aligned}
d_{VI}(\Delta, \Delta') &= H_\Delta + H_{\Delta'} - 2I_{\Delta',\Delta} \\
&= H_{\Delta|\Delta'} + H_{\Delta'|\Delta}
\end{aligned}
$$

- Interpretation: d_{VI} is the sum of information gained and information lost when labels are switched from $k()$ to $k'()$
- d_{VI} symmetric, ≥ 0
- d_{VI} obeys triangle inequality (is a metric)

Other properties

- Upper bound
  $d_{VI} \leq 2 \ln K_{max}$ if $K, K' \leq K_{max} \leq \sqrt{n}$
  (asymptotically attained)
- $d_{VI} \leq \ln n$ over all partitions (attained)
- Unbounded! and grows fast for small K

STAT/BIOST 527

Other criteria and desirable properties

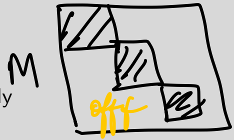
- Comparing clustering by **indices of similarity** $i(\Delta, \Delta')$
  - from statistics (Rand, adjusted Rand, Jaccard, Fowlkes-Mallows ...)
  - Normalized Mutual Information
  - range=[0,1], with $i(\Delta, \Delta') = 1$ for $\Delta = \Delta'$
  - the properties of these indices not so good
  - any index can be transformed into a "distance" by $d(\Delta, \Delta') = 1 - i(\Delta, \Delta')$

but no triangle inequality

- Other desirable properties of indices and distances between clusterings
  - n-invariance
  - locality
  - convex additivity

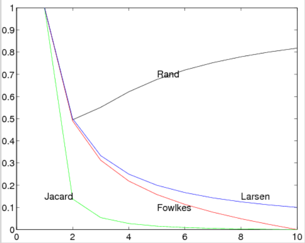
Rand, Jaccard and Fowlkes-Mallows

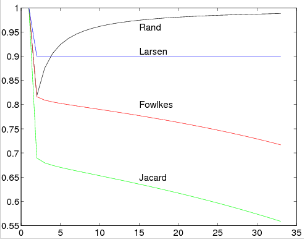
- Define N_{11} = # pairs which are together in both clusterings, N_{12} = # pairs together in Δ, separated in Δ', N_{21} (conversely), N_{22} =#number pairs separated in both clusterings
- Rand index = $\frac{N_{11}+N_{22}}{\#\text{pairs}}$
- Jaccard index = $\frac{N_{11}}{\#\text{pairs}}$
- Fowlkes-Mallows = Precision$\times$ Recall



- all vary strongly with K. Thereforek, **Adjusted** indices used mostly

$$adj(i) = \frac{i - \bar{i}}{\max(i) - \bar{i}}$$





Normalized Mutual Information (NMI)

$$i_{NMI}(\Delta, \Delta') = \frac{I_{\Delta',\Delta}}{H_\Delta + H_{\Delta'}} \leq \frac{1}{2} \quad (1)$$

- Takes values between [0,1]
- No probabilistic interpretaion
- Variant $\frac{I_{\Delta',\Delta}}{H_{\Delta,\Delta'}}$

Evaluation of Δ's in practice

$d(\Delta, \Delta')$ how to use? $\longrightarrow$ measure stability

"Bootstrap"

Idea: perturb – $\mathcal{D}$ by resampling
– algorithm – RF
– MS – seed set
....

$\Rightarrow$ obtain Δ^b, $b = 1:B$ perturbed clusterings

examine [distribution of] $d(\Delta, \Delta^b)$ $b = 1:B$

or $d(\Delta^b, \Delta^{b'})$, $b, b' = 1:B$

"closer to 0" $\Longleftrightarrow$ More stable $\Longleftrightarrow$ "Better" method

$\uparrow$

mean

median

90% quantile

CDF

...