CLUSTERING

SIMILARITY BASED CLUSTERING INPUT IS $N \times N$ DISSIMILARITY / DISTANCE MATRIX. $D_{ii} = 0$, $D_{ij} \geq 0$

FEATURE BASED CLUSTERING INPUT IS $N \times D$ FEATURE MATRIX

• **DISTANCES** $\Delta(x_i, x_i') = \sum \Delta_j(x_{ij}, x_{ij}')$
- SQUARED DISTANCE
- CORR COEFF
- L_1 / CITY BLOCK DISTANCE
- HAMMING DISTANCE (CATEGORICAL)

BENCHMARKS

IT'S UNSUPERVISED, HARD 2 EVAL. RELY ON EXTRA DATA

• **PURITY**: EMPIRICAL DISTR. OVER CLASS LABELS FOR CLUSTER $P = \sum_i \frac{N_i}{N} \cdot p_i$ TRIVIALIZED IF EACH OBJECT OWN CLUSTER

• **RAND INDEX**: $R = \frac{TP + TN}{TP + FP + TN + FN}$ REQUIRES REFERENCE SET. COMPARES POINT CLUSTERINGS

• **MUTUAL INFORMATION** $I(U,V) = \sum^{R}\sum^{C} p_{UV}(i,j) \log \frac{p_{UV}(i,j)}{p_U(i)\, p_V(j)}$ $p_{UV} = \frac{|U_i \cap V_j|}{N}$. $0 \leq I(U,V) \leq \min\{H(U), H(V)\}$

- NORMALIZED $NMI = \frac{I(U,V)}{(H(U) + H(V))/2}$

DIRICHLET PROCESS MIXTURE MODELS

• BEST WAY TO CHOOSE K $\rightarrow$ NOT HAVING TO CHOOSE K **INFINITE MIXTURE MODELS** NON PARAMETRIC, DIRICHLET BASED PRIOR

• IS **COMPUTATIONALLY SMART** • DP GIVES A DISTRIBUTION, SAMPLING FROM ~~FOR~~ SAMPLE GIVES US DATAPOINT.

• **DIRICHLET PROCESS** DISTRIBUTION OVER PROBABILITY MEASURES. $G \sim DP(\alpha, H)$ α CONCENTRATION PARAMETER, H BASE MEASURE

MV DISTRIB IS DIRICHLET. ~~BASE IS~~ MARGINALS ARE BETA • **DP** DEFINES CONJUGATE PRIOR FOR ARBITRARILY MEASURABLE SPACES

- **STICK BREAKING**: π INFINITE SEQUENCE OF MIXING WEIGHTS $\beta_k = BETA(1, \alpha)$, $\pi_k = \beta_k\left(1 - \sum_{l=1}^{k-1} \pi_l\right)$. $\pi \sim GEM(\alpha)$

GENERATES DISCRETE NUMBER OF ELEMENTS; INCREASING WITH α. $G(\theta) = \sum \pi_k \delta_{\theta_k}(\theta)$

KEEP SAMPLING $\rightarrow$ MORE REPETITIONS

$\llcorner$ $G \sim DP(\alpha, H)$

• α IS DISPERSION PARAMETER

DATA SAMPLED FROM $\bar{\theta}_i \sim G$ WILL CLUSTER AROUND θ_k

- **POLYA URN / CHINESE RESTAURANT PROCESS**: $\theta_i \sim G$ ARE N OBSERVATIONS FROM $G \sim DP(\alpha, H)$. THEY HAVE K DISTINCT VALUES θ_k.

$\longrightarrow$ NEXT OBSERVATION PREDICTION $p(z_{N+1} = z \mid z_{1:N}; \alpha) = \frac{1}{\alpha + N}\left(\underbrace{\alpha I(z = k^*)}_{\text{NEW CLUSTER}} + \underbrace{\sum^{K} N_k I(z = k)}_{\text{EXISTING CLUSTERS}}\right)$

• POLYA IS SAME WITH θ_k INSTEAD OF INDICATORS, ALSO ASSIGN PARAMS TO GRPS

• **DISTRIBUTION OVER PARTITION OF INTEGERS**

• $N \rightarrow \infty$; $K \rightarrow \alpha \log(N)$ BECAUSE OF PRIOR

• MORE FLEXIBLE PRIOR $\rightarrow$ PITMAN-YOR PROCESS

~~HOW TO FIT?~~ **FULL MODEL**

$\pi \sim GEM(\alpha, H)$ — G DRAWS UNBOUNDED PARAMS θ_k FROM BASE H. EACH HAS WEIGHT π_k.

$z_i \sim \pi$

$\theta_k \sim H(\lambda)$ — x_i GENERATED BY SAMPLING EACH OWN $\bar{\theta}_i$ FROM G.

$x_i \sim F(\theta_{z_i})$ — MORE DATA $\rightarrow$ MORE LIKELY $\bar{\theta}_i$ CLOSE TO θ_k ALREADY OUT.

HOW TO FIT?

• COLLAPSED GIBBS SAMPLING, MOSTLY. BUT HAS CASE FOR $z_i = k^*$ NEW CLUSTER. • IS EFFICIENT BECAUSE CREATES EXTRA REDUNDANT CLUSTERS EARLY ON

• OTHER METHODS: STAR/BEAM SEARCH, PARTICLE FILTERING.

• **HYPERPARAMS** $\longrightarrow$ PUT $Ga(a, b)$ ~~TO A~~ AS PRIOR FOR α TO CONTROL NO OF CLUSTERS

AFFINITY PROPAGATION

IDEA: EACH POINT MUST CHOOSE ANOTHER DATAPOINT AS ITS EXEMPLAR/CENTROID. SOME WILL CHOOSE THEMSELVES. CLUSTERING VIA MSG PASSING

→ MINIMIZE $S(c) = \underbrace{\sum^{N} s(i, c)}_{\text{SIMILARITY TO CENTROID}} + \underbrace{\sum^{N} \delta_k(c)}_{\text{PENALTY}}$

PENALTY: $-\infty$ IF k DIDN'T CHOOSE ITSELF. ELSE 0.

- OBJECTIVE FCN REPRESENTED AS **FACTOR GRAPH**. N NODES W/ N POSSIBLE VALUES. → USE MAX-PRODUCT BELIEF PROPAGATION FOR FINDING MAXIMUM.
- $c_i \to \delta_k$ VAR → FACTOR RESPONSIBILITY
- $\delta_k \to c_i$ FACTOR → VAR AVAILABILITY.
- NOT STABLE
- $O(N^2)$ BUT $O(E)$ NO EDGES IN SPARSE MATRICES
- CONTROL NO OF CLUSTERS VIA DIAGONAL TERMS $s(i,i)$ HOW MUCH EACH DP WANTS TO BE EXEMPLAR.

SPECTRAL CLUSTERING

USES GRAPHCUTS. WEIGHTED UNDIRECTED GRAPH W FROM SIMILARITY MATRIX **MINIMIZE** $\mathrm{cut}(A_1, \dots, A_k) = \frac{1}{2}\sum^{k} W(A_k, \bar{A}_k)$, $\bar{A}_k$ = COMPLEMENTO DI A_k

→ NORMALIZED CUT $\mathrm{Ncut} = \frac{1}{2}\sum^{k} \frac{\mathrm{cut}(A_k, \bar{A}_k)}{\mathrm{vol}(A_k)}$; $\mathrm{vol}(A_k) = \sum_i d_i$, $d_i = \sum_j^N w_{ij}$ WEIGHTED DEGREE OF NODE.

→ FORMULATION OF SEARCHING FOR BINARY VECTORS SUCKS. → RELAX BINARY MEMBERSHIP CONSTRAINT, REAL VALUES. → EIGENVALUE PROBLEM

GRAPH LAPLACIAN: W IS SYMMETRIC WEIGHT MATRIX. $D = \mathrm{diag}(d_i)$ NODE DEGREES. $L = D - W$

- EACH ROW SUMS TO 0 → $\mathbf{1}$ IS EIGENVECTOR WITH $\lambda_1 = 0$
- SYMMETRIC, POSITIVE DEFINITE

THEOREM: EIGENVECTORS OF L WITH EIGENVALUES 0 IS SPANNED BY INDICATOR VECTORS $\mathbf{1}_{A_1} \dots \mathbf{1}_{A_k}$ WHERE k ARE CONNECTED COMPONENTS OF GRAPH

ALGO IDEA: COMPUTE k SMALLEST EIGENVECTORS u_k OF L. U EIGVECTOR COLUMN MATRIX. y_i ARE ROWS. ASSIGN POINT i TO CLUSTER k IFF ROW y_i WAS ASSIGNED TO k.
'RELATED TO KPCA'
└ CLUSTER y_i WITH K-MEANS TO RECOVER CONNECTED COMPONENTS IS $O(N^3)$

IN PRACTICE: NORMALIZE L TO ACCOUNT FOR NODES MORE CONNECTED THAN OTHERS. DIFFERENT WAYS. $L_{norm} = D^{-1}L = I - D^{-1}W$

RELATION TO RANDOM WALKS: CONNECTED, NON BIPARTITE GRAPH HAS UNIQUE STATIONARY DISTRIBUTION $\pi = (\pi_1, \dots, \pi_N)$ $\pi_i = d_i/\mathrm{vol}(V)$ $P = D^{-1}W$

$\mathrm{Ncut}(A, \bar{A}) = P(\bar{A} \mid A) + P(A \mid \bar{A})$ → WE ARE LOOKING FOR CUT RARELY TRANSITIONING $A \longleftrightarrow \bar{A}$

P IS RANDOM WALK TRANSITION MATRIX

ALSO RELATED TO KERNEL PCA

HIERARCHICAL CLUSTERING

- AGGLOMERATIVE / DIVISIVE.
- INPUT: DISSIMILARITY MATRIX
- HEURISTIC METHODS; NO OBJECTIVE FUNCTION → HARD TO EVALUATE

AGGLOMERATIVE: STARTS WITH N GROUPS, EACH STEP MERGES 2 MOST SIMILAR GROUPS. TOTAL $O(N^3)$ BUT $O(N^2 \log N)$ WITH TRICKS. COMMONLY WE RUN KMEANS TO INITIALIZE CLUSTERS

- SINGLE LINK: NEAREST NEIGHBOR. DISTANCE IS CLOSEST DISTANCE BETWEEN MEMBERS. MINIMUM SPANNING TREE OF DATA/EDGE WEIGHTS. CAN DO IN $O(N$
- COMPLETE LINK: FURTHEST NEIGHBOR. " " LARGEST " " ". TENDS TO PRODUCE SMALL CLUSTERS
- AVG LINK: AVERAGES DISTANCES BETWEEN ALL PAIRS OF OBJECTS. INTERMEDIATE BEHAVIOR. PREFERRED

DIVISIVE: STARTS WITH SINGLE CLUSTER. SPLITS IT IN 2, TOP-DOWN. HEURISTICS TO PICK OPTIMAL SPLIT. CAN BE FASTER THAN AGGREGATIVE $O(N)$. SEES ALL DATA.

- PICK k WITH LARGEST DIAMETER → SPLIT IT WITH K-MEANS/MEDOIDS = 2 → BISECTING K-MEANS
- MIN SPANNING TREE → BREAK LARGEST DISSIMILARITY LINK
- DISSIMILARITY ANALYSIS PICK MOST DISSIMILAR OBJECT AND PUT IN OTHER CLUSTER UNTIL WE PICK i^* MAXIMIZING AVG G DISSIMILARITY / MINIMIZING H DISSIMILARITY
  └ FROM G └ H

- TRICKY TO PICK OPTIMAL K, EYBALL GAPS ON DENDROGRAM

BAYESIAN HIERARCHICAL CLUSTERING

SIMILAR TO AGGLOMERATIVE, BAYESIAN HYPOTHESIS TESTING TO PICK CLUSTERS TO MERGE
- INPUT IS DATA MATRIX.

- $P(D_{ij} \mid T_{ij}) = P(D_{ij} \mid M_{ij} = 1)P(M_{ij} = 1) + P(D_{ij} \mid M_{ij} = 0)P(M_{ij} = 0)$ FOR EACH CLUSTER
- PICK PAIR WITH HIGHEST $R_{ij} = \dfrac{P(D_{ij} \mid M_{ij} = 1)P(M_{ij} = 1)}{P(D_{ij} \mid T_{ij})}$
- CONNECTED TO DIRICHLET PROCESS MM → BHC GIVES US LOWER BOUND FOR DPMM MARGINAL LIKELIHOOD
- IS GREEDY SEARCH FOR BEST TREE PARTITION AT EACH STEP.
- THIS MAKES US ABLE TO $\pi_k = P(M_k = 1)$
- HYPERPARAMS α, λ OF DPMM THROUGH 'BACKPROP' THROUGH TREE

- BHC KICKS ASS!

BICLUSTERING

CLUSTERS ROWS AND COLUMNS. WHEN WE WANT TO CLUSTER ON FEATURES. BIOINFORMATICS. COLLABORATIVE FILTERING.

ASSOCIATE ROW/COLS TO LATENT INDICATORS AND ASSUME DATA ARE IID WRT SAMPLES AND FEATURES. $P(X|R,C,\theta) = \prod_i \prod_j P(x_{ij}|R_i, C_j, \theta) = P(x_{ij}|\theta_{R_i, C_j})$

$\theta_{a,b}$ PARAMS FOR ROW AND COL CLUSTER. • WE CAN USE DIRICHLET PROCESS VS FINITE NUMBER OF K!

MULTI-VIEW CLUSTERING

WE WANT TO MODEL DIFFERENT CLUSTERS ON THE BASIS OF DIFFERENT FEATURES. PARTITION COLUMNS INTO V **VIEWS** WHERE WE 'PICK' FEATURES. **DIRICHLET PROCESS FOR P(C)** SO V GROWS AUTOMAGICALLY. THEN FOR EACH VIEW WE PARTITION ROWS $R_{iv} = \{1 \dots K(v)\}$ CLUSTER WHERE R IS IN ROW V

DIRICHLET PROCESS FOR ROWS TOO. ASSUME ALL ROWS/COL IN DATA ARE IID.

• **BINARY DATA** → BETA-BERNOULLI MODEL • **HYPERPARAMS:** M-H WORKS WELL • ROBUST TO IRRELEVANT FEATURES → THEY ARE FILTERED OUT.

DBSCAN

DENSITY-BASED SPATIAL CLUSTERING OF APPLICATION WITH NOISE • **IDEA:** DENSITY IN NEIGHBORHOOD HAS TO EXCEED THRESHOLD

• DIRECT-DENSITY REACHABLE, CORE/BORDER POINT. $P \in N(Q)$, $|N(Q)| \geq \epsilon$. • DENSITY REACHABLE: CHAIN OF DDR POINTS FROM $P_1 \dots P_n$ • DENSITY CONNECTED: P,Q DC IFF ∃O DR P,Q

• **CLUSTER:** MAXIMAL SET OF DC POINTS • **NOISE:** OUTLIER PTS

ALGORITHM: SAME ϵ, MINPTS FOR ALL CLUSTERS. MANUALLY SET OR HEURISTIC: ESTIMATION VIA KNN. - GET N(P), CHECK, EXPAND CLUSTER, RAISE, REPEAT.

- AGGLOMERATIVE IN SUCCESSIVE ITERATIONS - MAY BE UNSTABLE; - NONDETERMINISTIC - RESOLVES NOISE IN SUCCESSIVE ITERATIONS - **WORST CASE:** $O(n^2)$ - **BEST:** $O(n\log m)$
- SUFFERS FROM COD

BIRCH

BALANCED ITERATIVE REDUCING AND CLUSTERING USING HIERARCHIES. IS ONLINE, CAN WORK IN $O(N)$. **CLUSTERING FEATURE:** ALSO MAINTAINS NO OF POINTS, SUM, SQUARED SUM OF CLUSTERS

• **CF-TREE:** A TREE STRUCTURE MAINTAINING THE CFs. PARENT-CHILD RELATIONSHIPS. BINARY TREE. INSERTION/SPLIT FOR ADDITION. LIMITED STRUCTURE. REFINING STEPS TO CONSOLIDATE STRUCTURE. **ALGORITHM:** SCAN DATA, PRUNE/REMOVE OUTLIERS, CONSOLIDATE/GLOBAL CLUSTERING, REFINE. **AGGLOMERATIVE HIERARCHICAL CLUSTERING**

↳ AGGLOMERATE ↳ ANY CLUSTERING ALGO ↳ RECOMPUTE CENTROIDS

• **DIFFERENT DISTANCE METRICS.**

• **IS MORE OF A FRAMEWORK FOR EFFICIENT CLUSTERING UNDER MEMORY CONSTRAINT THAN AN ALGORITHM.**

T-SNE

USED FOR DIM/REDUCTION & VISUALIZATION. RETAINS LOCAL AND GLOBAL STRUCTURE IN LOW-DIM SPACE. T-DISTRIBUTED STOCHASTIC NEIGHBOR EMBEDDING

VANILLA SNE: SIMILARITY OF x_i, x_j IS CONDITIONAL PROB i WOULD PICK j AS NEIGHBOR UNDER GAUSSIAN CENTERED ON x_i.

- $P_{j|i} = \dfrac{\exp(-\|x_i - x_j\|^2 / 2\sigma_i^2)}{\sum_{k \neq i} \exp(-\|x_i - x_k\|^2 / 2\sigma_i^2)}$ H-DIM
- $Q_{j|i} = \dfrac{\exp(-\|y_i - y_j\|^2)}{\sum_{k \neq i} \exp(-\|y_i - y_k\|^2)}$ L-DIM

IDEA: MINIMIZE MISMATCH BETWEEN $P_{j|i}$ AND $Q_{j|i}$ OVER ALL DATAPOINTS VIA GRADIENT DESCENT

COST FCN: $C = \sum_i KL(P_i \| Q_i) = \sum \sum P_{j|i} \cdot \log \dfrac{P_{j|i}}{Q_{j|i}}$

FIND: σ TO MINIMIZE P_i WITH FIXED, SPECIFIED PERPLEXITY $P(i) = 2^{H(P)}$, ENTROPY

GRADIENT: $\dfrac{\delta C}{\delta y_i} = 2 \sum_j (P_{j|i} - Q_{j|i} + P_{i|j} + Q_{i|j})(y_i - y_j)$

INIT: RANDOM SAMPLES AROUND ORIGIN. GRADDESC W/ MOMENTUM. ANNEALING W/ ARTIFICIAL NOISE REDUCED OVER TIME.

SYMMETRIC SNE: ASSUME $P_{i|j} = P_{j|i}$, $Q_{i|j} = Q_{j|i}$ $\forall i,j$. → $C = KL(P\|Q) = \sum_i \sum_j P_{ij} \cdot \log \dfrac{P_{ij}}{Q_{ij}}$ → $\dfrac{\delta C}{\delta y_i} = 4 \sum_j (P_{ij} - Q_{ij})(y_i - y_j)$

ISSUE: CROWDING PROBLEM: AREA FOR DISTANT POINTS NOT ENOUGH COMPARED TO THAT FOR CLOSE POINTS IN LOW DIM. → **USE A STUDENT'S T** IN LOW DIM INSTEAD OF GAUSSIAN. HEAVIER TAILS

T-SNE: • $Q_{ij} = \dfrac{(1 + \|y_i - y_j\|^2)^{-1}}{\sum_{k \neq l} (1 + \|y_k - y_l\|^2)^{-1}}$ • T IS 1-DOF CAUCHY. NUMERATOR ≈ INVERSE SQUARE LAW. • INVARIANT OF CHANGE OF SCALE. • LARGE CLUSTERS INTERACT LIKE INDIVIDUAL POINTS. LONG-RANGE FORCES

• $\dfrac{\delta C}{\delta y_i} = 4 \sum_j (P_{ij} - Q_{ij})(y_i - y_j)(1 + \|y_i - y_j\|^2)^{-1}$. EFFICIENT TO COMPUTE

• DISSIMILAR POINTS LARGE DISTANCES
• SIMILAR POINTS SMALL PAIRWISE DISTANCES

OPTIMIZATION: SMALL MOMENTUM IN FIRST ITERATIONS. L2 PENALTY TO COST FCN; PROPORTIONAL TO SSQUARES, MAP INITIALLY CLOSE. OR EXAGGERATE HIGH-DIM DISTANCES AT START. MAKES EMPTY SPACE IN LOW-DIM. BETTER VIZ WELL. • INITIALLY REDUCE DATA WITH PCA TO, LIKE, 30 TO AVOID TOO MUCH OVERHEAD

• **> 10K POINTS:** REDUCE CPLXITY VIA NEIGHBOURHOOD GRAPH FOR ALL POINTS. RANDOM WALKS TO PICK P_{ij} • LANDMARK POINTS.