

1 Basics

- General p-norm: $\|x\|_p = (\sum_{i=1}^n |x_i|^p)^{1/p}$
- Taylor: $f(x) \approx f(a) + \frac{f'(a)}{1!}(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \dots$
 - $f(x) \approx f(a) + \frac{\partial f(x)}{\partial x} \Big|_a - \frac{1}{2}(x-a)^\top \left(\frac{\partial^2 f(x)}{\partial x \partial x^\top} \right) \Big|_a (x-a)$
 - Power series of exp.: $\exp(x) := \sum_{k=0}^{\infty} \frac{x^k}{k!}$
- Entropy: $H(X) \equiv H(p_X) = \mathbb{E}_X[-\log \mathbb{P}(X=x)]$
 - $H(X|Y) = \sum_y \mathbb{P}(Y=y) H(X|Y=y) \leq H(X)$
 - $H(X, Y) = H(X) + H(Y|X)$
 - $H(X|g(X)) \geq 0$ $H(g(X)|X) = 0$
 - $H(5X) \begin{cases} = H(X) & \text{discrete} \\ > H(X) & \text{continuous} \end{cases}$
- MI: $I(X; Y|Z) = H(X|Z) - H(X|Y, Z)$ (symmetric)
 - $I(X; Y) = D_{\text{KL}}(p(x, y) \| p(x)p(y)) \geq 0$
 - $I(X_1, \dots, X_n; Z) = \sum_{i=1}^n I(X_i; Z | X_1, \dots, X_{i-1})$
 - Markov chain: $I(X_1; X_2, X_3, \dots) = I(X_1; X_2)$
 - $I(X, Y; Z) = I(X; Z) + I(Y; Z | X)$
- KL-divergence: $D_{\text{KL}}(p \| q) = \sum_x p(x) \log\left(\frac{p(x)}{q(x)}\right) \geq 0$
- Cauchy-Schwarz: $|\mathbb{E}[X, Y]|^2 \leq \mathbb{E}[X^2] \mathbb{E}[Y^2]$
- $1 - z \leq \exp(-z)$
- Jensen, $f(X)$ convex: $f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)]$

1.1 Probability / Statistics

- Gaussian: $\mathcal{N}(x | \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2} \frac{(x-\mu)^2}{\sigma^2}\right)$

$$\mathcal{N}(x | \mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \exp\left(-\frac{1}{2} (x - \mu)^\top \Sigma^{-1} (x - \mu)\right)$$
 - $X \sim \mathcal{N}(\mu, \Sigma), Y = A + BX \implies Y \sim \mathcal{N}(A + B\mu, B\Sigma B^\top)$
- Binomial: $f(k, n; p) = \mathbb{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$
- $\mathbb{V}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$

$$\mathbb{V}[X + Y] = \mathbb{V}[X] + \mathbb{V}[Y] + 2\text{Cov}(X, Y)$$
- $\text{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$

$$= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$$

$$\text{Cov}(aX, bY) = ab\text{Cov}(X, Y)$$

1.2 Calculus

- Partial: $\int uv' dx = uv - \int u'v dx$ $\frac{\partial}{\partial x} \frac{g}{h} = \frac{g'h - gh'}{h^2}$
- $\frac{\partial}{\partial x} (\|x - b\|_2) = \frac{x-b}{\|x-b\|_2}$ $\frac{d}{dx} |x| = \frac{x}{|x|}$

- $\frac{\partial}{\partial X} \log|X| = X^{-\top}$ $|X^{-1}| = |X|^{-1}$
- $\frac{\partial}{\partial x} (b^\top x) = \frac{\partial}{\partial x} (x^\top b) = b$
- $\frac{\partial}{\partial x} (b^\top Ax) = A^\top b$ $\frac{\partial}{\partial X} (c^\top Xb) = cb^\top$
- $\frac{\partial}{\partial X} (c^\top X^\top b) = bc^\top$ $\frac{\partial}{\partial x} (x^\top x) = 2x$
- $\frac{\partial}{\partial x} (x^\top Ax) = (A^\top + A)x \stackrel{\text{A sym.}}{=} 2Ax$
- $\frac{\partial}{\partial X} \text{Tr}(X^\top A) = A$ Trace trick: $x^\top Ax = \dots$
 - $\dots \stackrel{\text{inn. prod.}}{=} \text{Tr}(x^\top Ax) \stackrel{\text{cycl. permut.}}{=} \text{Tr}(xx^\top A) = \text{Tr}(Axx^\top)$
- $\sigma(x) = \frac{1}{1+\exp(-x)} \implies \nabla \sigma(x) = \sigma(x)(1-\sigma(x))$
- $\tanh(x) = \frac{2\sinh(x)}{2\cosh(x)} = \frac{e^x - e^{-x}}{e^x + e^{-x}}$ $\nabla \tanh(x) = 1 - \tanh^2(x)$

2 Empirical Risk Minimisation (ERM)

- Cost:** $R(c, X, Y) = \sum_{i \leq N} \|y_i - c^\top x_i\|^2$ (regr.)
 or $R(c, X, Y) = \sum_{i \leq N} \max(0, -y_i c^\top x_i)$ (class.)
 or $R(c, \theta, X) = \sum_{i \leq N} \|x_i - \theta_{c(i)}\|^2$ (clust.)

Goal: $\arg \min_c \mathbb{E}_{\mathcal{X}}[R(c, \mathcal{X})] \approx \arg \min_c \frac{1}{N} R(c, X)$

2.1 Bayesianism / Frequentism

Bayesianism: Define prior $P(\theta)$, define likelihood $P(X | \theta)$, compute posterior $P(\theta | x_{1..n})$.

Bayes: $P(\theta | X) = \frac{P(X|\theta)P(\theta)}{P(X)}$, $P(X) = \sum_{\theta} P(X|\theta_i)P(\theta_i)$

Frequentism: Define param. model $P(Y|X, \theta)$, compute likelihood of data $P((X, Y) | \theta)$ and compute $\hat{\theta}_{\text{MLE}}$ via $\arg \max_{\theta}$ of likelihood.

2.2 Linear Regression

model: $\hat{y} = X\beta$

Ridge: $\epsilon_{\text{RSS}}(\beta, \lambda) = (y - X^\top \beta)^\top (y - X^\top \beta) + \lambda \beta^\top \beta$
 $\hat{\beta} = (X^\top X + \lambda \mathbb{I})^{-1} X^\top y$, prior: $\beta \sim \mathcal{N}(0, \frac{\sigma^2}{\lambda} \mathbb{I})$

Lasso: $\hat{\beta} = \arg \min_{\beta} \sum_{i \leq n} (y_i - x_i^\top \beta)^2 + \lambda \|\beta\|_2$
 (no closed form), prior: $p(\beta_i) = \frac{\lambda}{4\sigma^2} \exp(-|\beta_i| \frac{\lambda}{2\sigma^2})$

3 Maximum Entropy Inference

Sample $c \sim p(\cdot | X)$ s.t. $H[p(\cdot | x)]$ is maximal,
 $\mathbb{E}_{C|X}[R(C, X)] = \mu$ and $\sum_c p(c | X) = 1$.

\implies **Gibbs dist.:** $p(c | X) = \frac{1}{Z(X)} \exp(-\beta R(c, X))$

Free energy: $F(X) := -\frac{1}{\beta} \log Z(X)$

$\iff p(c | X) = \exp(-\beta [R(c, X) - F(X)])$

\implies **entropy:** $H[c | X] = \beta \underbrace{\mathbb{E}_{C|X}[R(C, X)]}_{=\mu} - \beta F(X)$

ME: $\max H[c | X] \iff \max Z(X) \iff \min F(X)$

- Exp. generalisation costs: $\mathbb{E}_{X'} \mathbb{E}_{X''} \mathbb{E}_{C|X'}^{\mathbb{E}_C} [R(c, X'')]$
- Min. out-of-sample descr. length per deg. of freedom

$$\min_{p(\cdot)} \mathbb{E}_{X', X''} \mathbb{E}_{C|X'} \left[-\log \frac{p(c|X'')}{p(c)} \right] \quad p(c) = \mathbb{E}_X [p(c | X)]$$

$$\stackrel{\text{Jensen}}{\geq} \min_{p(\cdot)} \mathbb{E}_{X', X''} \left[-\log \mathbb{E}_{C|X'} [p(c | X'')] \right] - H[c]$$

$$= \max_{p(\cdot)} \mathbb{E}_{X', X''} [e^{H[c]} \cdot \kappa(X', X'')]$$

PA: $T^* = \arg \max_T \kappa(X', X'')$

- PA-kernel: $\kappa(X', X'') := \sum_c p(c | X') p(c | X'')$
- combined: $p(c | X', X'') \propto p(c | X') p(c | X'')$

4 Methods for intractable Gibbs distr.

4.1 Sampling and SA

Well behaving **Markov Chains** are

- irreducible:** can go from/to any state, and
- aperiodic:** chain doesn't go "back & forth" forever.

\implies **Stationary dist.** $p(c') = \sum_c \pi(c | c') p(c)$

\iff **det. balance** $\pi(c' | c) p(c) = \pi(c | c') p(c')$

Metropolis-Hastings: Assume $p(c) \propto f(c)$.

$$\pi(c' | c) := \begin{cases} q(c' | c) A(c, c') & c \neq c' \\ 1 - \sum_{c' \neq c} q(c' | c) A(c, c') & \text{otw.} \end{cases}$$

where $q(c' | c)$: prob. to propose the move $c \rightarrow c'$,
 and $A(c, c') := \min\left\{1, \frac{q(c|c')f(c')/Z}{q(c'|c)f(c)/Z}\right\}$ prob. accept move

Metropolis Algorithm: Assume $p(c) \propto f(c)$ and $q(c' | c) = q(c | c')$, i.e. symmetric.

1. Define symmetric $\{q(\cdot | c)\}_{c \in \mathcal{C}}$ s.t. graph G_q is connected and every vertex in G_q has edge to itself.
2. $c_0 \leftarrow \$$ Then, for $t = 1, 2, \dots$, do:
 - $\tilde{c} \leftarrow q(\cdot | c_{t-1})$ // sample
 - $b \leftarrow \text{Bern}\left(\min\left\{1, e^{-\frac{1}{T}[R(\tilde{c}, X) - R(c_{t-1}, X)]}\right\}\right)$
 - If $b = 1$ then $c_t \leftarrow \tilde{c}$ else $c_t \leftarrow c_{t-1}$.

$\pi(c' | c) = \{\dots \leftarrow \text{c.f. scr. (2.7)}$

Simulated annealing: Gradually decrease temp. T to escape bad local minima. \rightarrow MH-sampling from Gibbs (DA does not sample!).

4.2 Laplace's Method (Least angle clust.)

1. Square the cost: $e^{-\frac{1}{T}R(c, X)} = \text{const} \cdot e^{g(c)^T g(c)}$
2. Complete the square: $\int e^{-\frac{1}{T}(y - g(c))^2} dy = (\pi T)^{d/2} \Rightarrow e^{g(c)^T g(c)} = (\pi T)^{-d/2} \int \exp^{-y^T y + 2y^T g(c)} dy$
3. Rewrite normalisation constant: $Z = \sum_c e^{-\frac{1}{T}R(c, X)} = \dots = \text{const} \int e^{-\frac{1}{T}f(y)} dy$
4. Apply Laplace's method: If f has unique min. y_0 and Hessian $H := \frac{\partial^2 f}{\partial y^2}|_{y_0}$

$$\int e^{-\frac{1}{T}f(y)} dy \stackrel{(T \rightarrow 0)}{\approx} e^{-\frac{1}{T}f(y_0)} \left| \frac{H}{2\pi T} \right|^{-1/2}$$

4.3 Mean-field Approximation

Idea: Approximate p_β (Gibbs) with a “simple”, factorisable distribution $p = p_1 \cdots p_N$.

Approach: Minimise $D_{\text{KL}}(p \| p_\beta)$

\iff Minimise **Gibbs free energy:**

$$G(p) = \frac{1}{\beta} D_{\text{KL}}(p \| p_\beta) + F(\beta) = \mathbb{E}_{c \sim p}[R(c)] - \frac{1}{\beta} H[p]$$

Note: $H[p] = \sum_{i=1}^N H[p_i]$ and $F(\beta) \leq G(p)$

Ising model: $R(c | J) = -\frac{1}{2} \sum_{i,j} J_{ij} c_i c_j - \sum_i h_i c_i$ where J_{ij} : interaction between particles,

h_i : noisy image, σ_i : denoised image

Problem: $\frac{\partial G(p)}{\partial p_{i\ell}} = 0$ s.t. $\sum_{\ell'} p_{i\ell'} = 1 \forall i$

Solution: with the mean field $h_i = [\dots h_{i\ell} \dots]^T$

$$h_{i\ell} := \frac{\partial \mathbb{E}[R(c)]}{\partial p_{i\ell}} = \mathbb{E}_{c \sim p_{i \rightarrow \ell}}[R(c)] \leftarrow \text{object } i \text{ chooses class } \ell$$

$$p_{i\ell} = e^{-\beta h_{i\ell}} / Z_i$$

EM-like Algo: Iteratively 1. Pick random i
2. $h_i^{\text{new}} \leftarrow p_j^{\text{old}}$ 3. $p_i^{\text{new}} \leftarrow h_i^{\text{new}}$ until converged.

4.3.1 Smooth k-means scr.20 (p. 39)

$R(c | X) = \sum_i \|x_i - y_{c_i}\|^2 + \frac{\lambda}{2} \sum_i \sum_{j \in N(i)} \mathbb{I}_{\{c_i \neq c_j\}}$ where the second term measures #violations of these neighbourhood constraints.

$$\implies h_{i\ell} = \|x_i - y_\ell\|^2 + \lambda \sum_{j \in N(i)} p_{j\ell} + \text{const}_i$$

5 Deterministic Annealing (Z is tractable)

Lemma: func's \times domain \rightarrow domain \times co-dom.

$$\mathcal{O}(K^N) \rightarrow \sum_c \prod_i \epsilon_{i,c(i)} = \prod_i \sum_k \epsilon_{ik} \leftarrow \mathcal{O}(NK)$$

$$p(c | \theta, X) = \prod_{i \leq N} p_i(c(i) | \theta, X) \text{ where } p_i(k | \theta, X) \propto \exp(-\frac{1}{T} \|x_i - \theta_k\|^2)$$

$$\text{Max. entr.} \implies \frac{\partial \log Z}{\partial \theta_k} = 0 \implies \theta_k^* = \frac{\sum_i p_i(k | \theta^*, X) \cdot x_i}{\sum_i p_i(k | \theta^*, X)}$$

do
E-step: $p_i(k | \theta^{\text{old}}, X) = \frac{\exp(-\frac{1}{T} \|x_i - \theta_k\|^2)}{\sum_{j \leq K} \exp(-\frac{1}{T} \|x_i - \theta_j\|^2)}$
M-step: $\theta_k \leftarrow \dots$
 $\theta^{\text{old}} \leftarrow \theta$

until convergence of θ

$$\theta_k \leftarrow \theta_k + \epsilon \text{ (noise s.t. centroids can separate)}$$

Phase transitions: For $T \rightarrow \infty$: $\theta_k^* = \bar{X} \forall k \leq K$
Once $T = 2\lambda_{\max}$, more centroids appear, where $\lambda_{\max} = \text{max. eigenvalue of } \frac{1}{N} X^T X$. (x_i 's row-wise)

6 Histogram Clustering

Least Angle Clust. (LAC): [Idea]

Similarity $S(x_i, x_j) = w_{ij} \cos(\phi_{ij}) = w_{ij} e_i \cdot e_j$ with unit vectors $e_i := x_i / \|x_i\|$, e.g. choice $w_{ij} = \|x_i\| \cdot \|x_j\|$.

Dyadic data: $\mathcal{Z} = \{(x_{i(r)}, y_{j(r)}); 1 \leq r \leq \ell\}$

• prototype / “centroid”: $q(y_j | \alpha)$

• empirical dist.: $\hat{p}(y_j | x_i) = \frac{\hat{p}(x_i, y_j)}{\hat{p}(x_i)} \leftarrow \text{scr. (5.10)}$

Likelihood: $P(\mathcal{Z} | c, q) = \prod_{r \leq \ell} p(x_{i(r)}, y_{j(r)} | c, q) = \text{scr. (5.12)} = \prod_i \prod_j [q(y_j | c(i)) \cdot p(c(i))]^{\ell \hat{p}(x_i, y_i)}$

Assume $p(\alpha) = 1/k$ and $\hat{p}(x_i) = 1/n$

\Rightarrow **Cost:** $R^{\text{hc}}(c, q, \mathcal{Z}) = \frac{\ell}{n} \sum_{i \leq n} D_{\text{KL}}[\hat{p}(\cdot | x_i) \| q(\cdot | c(i))]$

Solving the **Gibbs dist.** $p(c | q, \hat{p}) = \prod_{i \leq n} P_{i,c(i)}$

via Lagrange yields $q^*(y_j | \alpha) = \frac{\sum_{i \leq n} P_{i\alpha} \cdot \hat{p}(y_j | x_i)}{\sum_{i \leq n} P_{i\alpha}}$ Lemma 2 ch.3 p.36

6.1 Information Bottleneck Method

Find efficient code $X \mapsto \hat{X}$ (codebook vector) and preserve relevant info. about context Y .

Criterion: $R^{\text{IB}}(q(\hat{x} | x)) = I(X; \hat{X}) - \beta I(\hat{X}; Y)$

Markov chain: $\hat{X} \xrightarrow{q(\hat{x}|x)} X \xrightarrow{p(y|x)} Y$

Generation process: w/ distortion $d(x, \hat{x}) = D_{\text{KL}}[\cdot]$

$$\begin{cases} q_t(\hat{x}|x) \propto q_t(\hat{x}) \cdot \exp(-\beta D_{\text{KL}}[p(y|x) \| p_t(y|\hat{x})]) \\ q_{t+1}(\hat{x}) = \sum_x p(x) \cdot q_t(\hat{x} | x) \\ p_{t+1}(y|\hat{x}) = \sum_x p(y | x) \cdot p(x) \cdot q_t(\hat{x} | x) / q_t(\hat{x}) \end{cases}$$

6.2 Parametric Distributional Clustering

Idea: Use a mixture of Gaussian prototypes, i.e.

$$p(y_j | v) \equiv p(b | v) = \sum_{\alpha \leq s} p(\alpha | v) G_\alpha(b).$$

$$x_i \xrightarrow{c(i)=v} v \xrightarrow{p(b|v)} \hat{p}(b | i)$$

Note: Feature values y_j (“bins” b) only depend on cluster index v and not explicitly on the site x_i !

Notation: $x_i \leftarrow i$, $y_j \leftarrow b$ (bins), $v \leftarrow$ clusters

Likelihood: (both equivalent if $p(i) = \frac{1}{n}$)

$$P(X | c, \theta) = \prod_{i \leq n} p(c(i)) \prod_{b \leq m} [p(b | c(i))]^{\ell \hat{p}(i,b)},$$

$$P(X, M | \theta) = \prod_{i \leq n} \prod_{v \leq k} [p(v) \cdot \prod_{b \leq m} p(b | v)^{n_{ib}}]^{M_{iv}}$$

where n_{ib} : #occur. an observ. at site i is inside I_b

$M_{iv} = p(v | i) \in \{0, 1\}$ clust. membersh. assign.

Cost (IB): $R^{\text{PDC}}(c, p_{\cdot|c}) = -\log P(X, M | \theta) = \dots$

$$\dots = -\sum_{i \leq n} \left[\log p_{c(i)} + \frac{\ell}{n} \sum_{b \leq m} \hat{p}(b | i) \log p(b | c(i)) \right]$$

E-step: $h_{iv} = -\log p_v - \sum_b \frac{\ell}{n} \hat{p}(b | i) \log p(b | v)$
 $q_{iv} = \mathbb{E}[\mathbb{I}_{\{c(i)=v\}}] \propto \exp(-h_{iv}/T)$
M-step: $p_v = \frac{1}{n} \sum_{i \leq n} q_{iv}$
 No closed form sol. for $p(\alpha | v)$. Thus, iteratively optimize pairs s.t. $\sum_{\alpha} p(\alpha | v) = 1$.

7 Graph-based Clustering

Non-metric relations: might assume negative values or violate the triangular inequality.

Setting: objects $\mathbf{o}_i, \mathbf{o}_j \in \mathcal{O}$; relations with weights $\mathcal{D} := \{D_{ij}\}$ on the edges (i, j) .

- Cluster α : $\mathcal{G}_{\alpha} \equiv \{\mathbf{o} \in \mathcal{O} : c(\mathbf{o}) = \alpha\}$
- Inter-cluster edges: $\mathcal{E}_{\alpha\beta} = \{(i, j) \in \mathcal{E} : \mathbf{o}_i \in \mathcal{G}_{\alpha} \wedge \mathbf{o}_j \in \mathcal{G}_{\beta}\}$
- cut(A, B) = $\sum_{i \in A, j \in B} W_{ij} \rightarrow$ weight matrix W
- assoc(A, \mathcal{V}) = $\sum_{i \in A, j \in \mathcal{V}} W_{ij} \rightarrow$ total connection strength from nodes in A to all nodes in the graph

Correlation clustering:

Minimise the sum of *pairwise* intracluster distances.

$$R^{cc}(c; \mathcal{D}) = - \sum_{v \leq k} \sum_{(i, j) \in \mathcal{E}_{vv}} S_{ij} + \sum_{v \leq k} \sum_{\substack{\mu \leq k \\ \mu \neq v}} \sum_{(i, j) \in \mathcal{E}_{v\mu}} S_{ij}$$

$$= -2 \sum_{v \leq k} \sum_{(i, j) \in \mathcal{E}_{vv}} S_{ij} + \sum_{\substack{\mu \leq k \\ \mu \neq v}} \sum_{(i, j) \in \mathcal{E}_{v\mu}} S_{ij}$$

\hookrightarrow intra-cluster \hookrightarrow const

up to thresh. $u \quad \stackrel{*}{=} -\frac{1}{2} \sum_{v \leq k} \sum_{(i, j) \in \mathcal{E}_{vv}} (|S_{ij} - u| + S_{ij} - u)$

$$+ \frac{1}{2} \sum_{v \leq k} \sum_{\substack{\mu \leq k \\ \mu \neq v}} \sum_{(i, j) \in \mathcal{E}_{v\mu}} (|S_{ij} + u| - S_{ij} - u)$$

* : altern. def. where $\frac{1}{2}(|X| \pm X) = \max\{0, \pm X\}$

Graph partitioning: $D_{ij} \in \mathbb{R}$

$$R^{gp}(c; \mathcal{D}) = \text{const} - \sum_{v \leq k} \text{cut}(\mathcal{G}_v(\mathcal{D}), \mathcal{V} \setminus \mathcal{G}_v(\mathcal{D}))$$

$$= \text{const} + \sum_{v \leq k} \text{cut}(\mathcal{G}_v(\mathcal{S}), \mathcal{V} \setminus \mathcal{G}_v(\mathcal{S}))$$

Bias in $R(c; \mathcal{D})$: Cost should scale prop. to #objects, i.e. $R(c; \mathcal{D}) = \mathcal{O}(n)$. * : use $D_{ij} = D(1 - \delta_{ij})$

Tip: $\frac{\text{cut}(\mathcal{G}_{\alpha}, \mathcal{V} \setminus \mathcal{G}_{\alpha})}{\text{assoc}(\mathcal{G}_{\alpha}, \mathcal{V})} \stackrel{*}{=} \frac{n \cdot p_{\alpha} \cdot n(1 - p_{\alpha}) \cdot D}{n \cdot p_{\alpha} \cdot n \cdot D} = 1 - p_{\alpha}$

7.1 Pairwise Clustering

Cost: $R^{pc}(c; \mathcal{D}) = \sum_{\alpha} \sum_{(i, j) \in \mathcal{E}_{\alpha\alpha}} \frac{D_{ij}}{|\mathcal{G}_{\alpha}|} = \sum_{\alpha} \sum_{(i, j) \in \mathcal{E}_{\alpha\alpha}} |\mathcal{G}_{\alpha}| \frac{D_{ij}}{|\mathcal{E}_{\alpha\alpha}|}$

Equivariance to k -means: (if $D_{ij} = \|\mathbf{x}_i - \mathbf{x}_j\|^2$)

$$\sum_{i \leq n} \|\mathbf{x}_i - \mathbf{y}_{c(i)}\|^2 = \sum_{i \leq n} \sum_{j \leq n} \sum_{\alpha \leq k} \frac{\mathbb{I}_{\{c(i)=\alpha\}} \mathbb{I}_{\{c(j)=\alpha\}}}{|\mathcal{G}_{\alpha}|} D_{ij}$$

Invariance properties:

- Symmetrisation: $R^{pc}(c; \mathcal{D}^s) \equiv R^{pc}(c; \mathcal{D})$
- Off-diagonal shift: $R^{pc}(c; \tilde{\mathcal{D}}) = R^{pc}(c; \mathcal{D}) - \lambda_{\min} \cdot n$

Theorem: If S^c is p.s.d., then D derives from squared Eucl. space. \implies Make S p.s.d.: $\tilde{S} := S - \lambda_{\min} \mathbb{I}$

Constant Shift Embedding:

- Symmetrise** $D \rightarrow D^s$: $D_{ij}^s := \frac{1}{2}(D_{ij} + D_{ji})$
- Centralise** D , then S : $X^c := QX^sQ^T$
 $Q = \mathbb{I} - \frac{1}{n} \mathbf{e}_n \mathbf{e}_n^T$ $S^c = -\frac{1}{2} D^c$
 $X_{ij}^c = X_{ij} - \frac{1}{n} \sum_k X_{ik} - \frac{1}{n} \sum_k X_{kj} + \frac{1}{n^2} \sum_{k, \ell} X_{k\ell}$
 \implies sum over column/rows = 0
- (Off-)Diagonal shift:** Find λ_{\min} of S^c
 $\tilde{S} := S^c - \lambda_{\min} \mathbb{I}$ $\tilde{D} := D - \lambda_{\min}(\mathbf{1} - \mathbb{I})$
 $\tilde{D}_{ij} = \tilde{S}_{ii} + \tilde{S}_{jj} - 2\tilde{S}_{ij} = \|\mathbf{x}_i - \mathbf{x}_j\|^2$

Reconstruction:

- EVD: $\tilde{S} = \mathbf{V} \mathbf{\Lambda} \mathbf{V}^T$ via $(\tilde{S} - \lambda \mathbb{I})\mathbf{v} \stackrel{!}{=} 0$ ($|\mathbf{v}| = 1$)
 where $\mathbf{\Lambda} = \text{diag}(\lambda_1 \dots \lambda_n)$ and $\mathbf{V} = [\mathbf{v}_1 \dots \mathbf{v}_n]$
- Find p s.t. $\lambda_1 \geq \dots \lambda_p > \lambda_{p+1} = \dots = \lambda_n = 0$
- $\implies \mathbf{X}_p = \mathbf{V}_p(\mathbf{\Lambda}_p)^{1/2}$ (each row is a vector)
- $\implies \mathbf{X}_t = \mathbf{V}_t(\mathbf{\Lambda}_t)^{1/2}$ (approx. & denoising)

Cluster membership of new data:

Note: S^{new} is def. by $D_{ij}^{\text{new}} = S_{ii}^{\text{new}} + \tilde{S}_{jj} - 2S_{ij}^{\text{new}}$

- $(S^{\text{new}})^c = -\frac{1}{2} \left[D^{\text{new}}(\mathbb{I}_n - \frac{1}{n} \mathbf{e}_n \mathbf{e}_n^T) - \frac{1}{n} \mathbf{e}_m \mathbf{e}_n^T + \tilde{D}(\mathbb{I}_n - \frac{1}{n} \mathbf{e}_n \mathbf{e}_n^T) \right]$
- Project: $X_p^{\text{new}} = (S^{\text{new}})^c \mathbf{V}_p(\mathbf{\Lambda}_p)^{-1/2}$
- Assign: $\hat{c}_i = \arg \min_c \|(x_p^{\text{new}})_i - \mathbf{y}_{c(i)}\|$

8 Model Selection for Clustering

What is the appropriate #clusters k for my data?

General approach: Measure quality (neg. log-likelihood) for different $k \rightarrow$ **elbow**.

8.1 Complexity-based Model Selection

Strategy: add a complexity term to neg. log-likelihood

Attention: MDL/BIC rely on likelihood optimisation

\rightarrow not generally applicable

Ocam's razor: Choose the model that provides the shortest description of the data.

8.1.1 Min. Description Length (MDL)

Minimise **descr. length:** $-\log p(\mathbf{X} | \theta) - \log p(\theta)$

Approx.: $\hat{k} \in \arg \min_k -\log p(\mathbf{X} | \hat{\theta}) + \frac{k'}{2} \log n$

8.1.2 Bayesian Information Crit. (BIC)

Parametrise likelihood $p(\mathbf{X} | M)$ by θ :

$$p(\mathbf{X} | M) = \int_{\Theta_M} \exp(\log p(\mathbf{X} | M, \theta)) \cdot p(\theta | M) d\theta$$

Assume flat prior $p(\theta | M) \approx \text{const}$ and

expand log-likelihood by ML estimator $\hat{\theta}$:

$$\bar{\ell}(\theta) = \frac{\ell(\theta)}{n} = \frac{1}{n} \log p(\mathbf{X} | M, \theta) \stackrel{\text{i.i.d.}}{=} \frac{1}{n} \sum_i \ell(\theta, X_i) \stackrel{\text{Taylor}}{\approx} \dots$$

$$\implies p(\mathbf{X} | M) = \text{const}_2 \cdot \exp\left(\ell(\hat{\theta}) - \frac{k'}{2} \log n\right)$$

where k' : dimension of (trainable) parameters

9 Model Validation

9.1 Stability-based Validation

Stability: Solutions on two data sets drawn from the same source should be similar.

9.2 Information-theoretic Validation

9.2.1 Shannon's Channel Coding Thm.

- Channel:** $(\mathcal{S}, \{p(\cdot | s)\}_{s \in \mathcal{S}})$, \mathcal{S} : alphabet
 - ϵ -noisy binary channel: $p(\hat{s} | s) = \begin{cases} 1-\epsilon & \text{if } \hat{s}=s \\ \epsilon & \text{if } \hat{s} \neq s \end{cases}$
- Capacity:** $\text{cap} = \max_p I(\mathcal{S}; \hat{\mathcal{S}}) \rightsquigarrow p_s(s)$
- (M, n)-code:** is a pair (Enc, Dec) \leftarrow scr. p.87
 where M : #messages, n : code-length

- **Rate:** $r = \frac{\log_2 M}{n} \Leftrightarrow M = \lfloor 2^{nr} \rfloor$
- **Commu. err.:** $p_{\text{err}} := \max_{i \leq M} \mathbb{P}(\widehat{\text{Dec}(\text{Enc}(i))} \neq i)$

Goal / **Best code:** $\lim_{n \rightarrow \infty} \frac{\log M}{n}$ s.t. $\lim_{n \rightarrow \infty} p_{\text{err}} \rightarrow 0$

Asymptotic equipartition property (AEP):

- $A_\epsilon^{(n)}$: Typical set of sequences $(s_1, \dots, s_n) \in \mathcal{S}^n$
 $\left| -\frac{1}{n} \log p_{\mathcal{S}^n}(s^n) - H[S] \right| < \epsilon \quad \leftarrow \text{scr. p.89}$
- $\mathbb{P}\left((S^n, \hat{S}^n) \in A_\epsilon^{(n)}\right) \xrightarrow{n \rightarrow \infty} 1 \quad \leftarrow \text{scr. p.90}$
- $p_{\text{err}} \leq 2^{-n(\text{cap} - 3\epsilon - r)} \xrightarrow{n \rightarrow \infty} 0$ if $r < \text{cap}$

9.2.2 Algorithm Validation

Assumptions:

- Exponential solution space, i.e. $\log|\mathcal{C}| = \mathcal{O}(n)$
- \mathcal{A} 's output is probabilistic, i.e. $p(\cdot | X')$

Ideal variant:

Messages: $\mathcal{M} = \{X'_1, \dots, X'_m\}$

Code: $X'_i \xrightarrow{\text{Enc}_A} p(\cdot | X'_i) \xrightarrow{\mathcal{C}_A} p(\cdot | X''_i) \xrightarrow{\text{Dec}_A} \hat{X}$

Empirical variant:

Messages: $\mathcal{M} = \{\tau_1, \dots, \tau_m\}$ drawn u.a.r. from \mathbb{T}

- Require $\sum_{\tau} p(c | \tau \circ X') \approx \frac{|\mathbb{T}|}{|\mathcal{C}|} \pm \rho \quad \leftarrow \text{scr. p.95}$

Code: $\tau_i \xrightarrow{\text{Enc}} p(\cdot | \tau_i \circ X') \xrightarrow{\mathcal{C}_A} p(\cdot | \tau_i \circ X'') \xrightarrow{\text{Dec}} \hat{\tau}$

- Enc_A : encodes $\tau_i \in \mathcal{M}$ as $p(\cdot | \tau_i \circ X')$
- Dec_A : selects $\hat{\tau} = \arg \max_{\tau} \kappa(\tau_i \circ X'', \tau \circ X')$

whereby $\kappa(X'', X') := \sum_c p(c | X'') p(c | X')$

Asymptotic Equipartition Property (AEP):

AEP fulfilled if $\log \kappa(X', X'') \xrightarrow{n \rightarrow \infty} \mathcal{E}$

whereby $\mathcal{E} := \mathbb{E}_{X', X''}[\log \kappa(X', X'')]$

- $A_\epsilon^{(n)}$: set of (ϵ, n) -typical pairs X', X''
 $|\log \kappa(X', X'') - \mathcal{E}| < \epsilon$

- $p_{\text{err}} \leq P_{(n)}$ c.f. scr. (6.19) $\xrightarrow{n \rightarrow \infty} 0$ if $\frac{\log m}{\log |\mathcal{C}|} < I$
 where $I := \frac{1}{\log |\mathcal{C}|} \mathbb{E}_{X', X''}[\log(|\mathcal{C}| \kappa(X', X''))]$

9.3 Applications of PA

PA: quantifies the amount of information that algorithms extract from phenomena. \rightarrow quantified by **capacity** (max. # distinguishable messages that can be communicated)

Temperature: $T^* = \arg \max_T \kappa(X', X'')$

Cost functions: Given $R_1(\cdot, \cdot), \dots, R_s(\cdot, \cdot)$
 $\max_{\ell \leq s} \kappa_\ell(X', X'') = \max_{\ell \leq s} \frac{1}{Z_{X'} Z_{X''}} \sum_c e^{-\frac{1}{T} R_\ell(c, X')} e^{-\frac{1}{T} R_\ell(c, X'')}$

Algorithms: Many MST (min. spanning tree) algo's are **contractive** (\rightarrow sequence of candidate sol's).

Approximation Set Coding (ASC):

$$p^{\text{ASC}}(c | X') = \begin{cases} 1/|G_\gamma(X')| & \text{if } c \in G_\gamma(X') \\ 0 & \text{otw.} \end{cases}$$

$$G_\gamma(X') := \left\{ c \in \mathcal{C} : R(c, X') - \min_{c \in \mathcal{C}} R(c, X') \leq \gamma \right\}$$

1. Run \mathcal{A} to compute $G_t^A(X')$ and $G_t^A(X'')$, for all t
2. $t^* = \arg \max_t \kappa(X', X'') = \arg \max_t \frac{|G_t^A(X') \cap G_t^A(X'')|}{|G_t^A(X')| \cdot |G_t^A(X'')|}$
3. $c^* \xleftarrow{\$ \text{sample}} \text{Unif}(G_{t^*}^A(X') \cap G_{t^*}^A(X''))$

10 Appendix

10.1 Tips and Tricks

Complete the square:

If $p(\mathbf{x}) \propto \exp(-\frac{1}{2} \mathbf{x}^\top \mathbf{A} \mathbf{x} + \mathbf{x}^\top \mathbf{b})$,
 then $p(\mathbf{x}) = \mathcal{N}(\mathbf{x} | \mathbf{A}^{-1} \mathbf{b}, \mathbf{A}^{-1})$

Constrained optimisation:

primal: $\min_{\mathbf{x}} f(\mathbf{x})$ s.t. $g_i(\mathbf{x}) = 0; h_j(\mathbf{x}) \leq 0$

Lagrangian: with each $\alpha_j \geq 0$

$$\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\alpha}) = f(\mathbf{x}) + \sum_i \lambda_i g_i(\mathbf{x}) + \sum_j \alpha_j h_j(\mathbf{x})$$

Solve: $\frac{\partial \mathcal{L}}{\partial \mathbf{x}} = 0; g_i(\mathbf{x}) = 0; \alpha_j \geq 0; h_j(\mathbf{x}) \leq 0$

If **Slater's cond.** holds, $\exists \mathbf{x} : g_i(\mathbf{x}) = 0, h_j(\mathbf{x}) < 0$, then we can solve the **dual** instead:

$$\max_{\boldsymbol{\lambda}, \boldsymbol{\alpha}} \{ \min_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\alpha}) \} \text{ s.t. } \alpha_j \geq 0$$

Solve: $\frac{\partial \mathcal{L}}{\partial \mathbf{x}} = 0; \frac{\partial \mathcal{L}}{\partial \boldsymbol{\lambda}} = 0; \alpha_j h_j(\mathbf{x}) = 0; \alpha_j \geq 0$

Euler-Lagrange: Find extrema of functional $\mathcal{F}[f] =$

$\int G(x, f(x), f'(x)) dx$, thus $\frac{\partial \mathcal{F}}{\partial f} \stackrel{!}{=} 0$.

If G is twice diff'able, then

$$\frac{\partial \mathcal{F}}{\partial f} = \frac{\partial G}{\partial f(x)} - \frac{d}{dx} \left(\frac{\partial G}{\partial f'(x)} \right) \stackrel{(*)}{=} \frac{\partial G}{\partial f(x)}.$$

(*) : when G does not depend on f' .

Hyperbolic Functions:

$$\sinh(x) = \frac{e^x - e^{-x}}{2}, \quad \frac{d}{dx} \sinh(x) = \cosh(x)$$

$$\cosh(x) = \frac{e^x + e^{-x}}{2}, \quad \frac{d}{dx} \cosh(x) = \sinh(x)$$

$$\tanh(x) = \frac{\sinh(x)}{\cosh(x)} = \frac{e^x - e^{-x}}{e^x + e^{-x}}, \quad \cosh^2(x) + \sinh^2(x) = 1$$

$$\frac{d}{dx} \tanh(x) = 1 - \tanh^2(x) = \frac{1}{\cosh^2(x)} = \text{sech}^2(x)$$

10.2 Approximations

Laplace Approximation: $\frac{df}{dx} \Big|_{x_0} = 0$

$$\Rightarrow \int_{\mathbb{R}} e^{Cf(x)} dx \approx \sqrt{2\pi C} \cdot |f''(x_0)| \cdot e^{Cf(x_0)}$$