Disclaimer

This document is an exam summary and follows the given material of the lecture *Statistical Learnint Theory*. Its contribution is a short summary that contains the most important concepts, formulas and algorithms. Due to curriculum content updates, some content may not be relevant to future versions of the course.

I do not guarantee the accuracy or completeness, nor is this document endorsed by the instructors. Any errors that are pointed out to me are welcome. The complete LaTeX source code can be found at `https://github.com/tstreule/eth-cheat-sheets`.

Statistical Learnint Theory

Timo Streule, tstreule@ethz.ch - *18.01.2022*

1 Basics

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

1.1 Probability / Statistics

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

1.2 Calculus

- Partial: $\int uv' \,\mathrm{d}x = uv - \int u'v \,\mathrm{d}x$ • $\frac{\partial}{\partial x}\frac{g}{h} = \frac{g'h}{h^2} - \frac{gh'}{h^2}$
- $\frac{\partial}{\partial \boldsymbol{x}}(\|\boldsymbol{x}-\boldsymbol{b}\|_2) = \frac{\boldsymbol{x}-\boldsymbol{b}}{\|\boldsymbol{x}-\boldsymbol{b}\|_2}$ • $\frac{\mathrm{d}}{\mathrm{d}x}|x| = \frac{x}{|x|}$
- $\frac{\partial}{\partial \boldsymbol{X}}\log|\boldsymbol{X}| = \boldsymbol{X}^{-\top}$ • $|\boldsymbol{X}^{-1}| = |\boldsymbol{X}|^{-1}$
- $\frac{\partial}{\partial \boldsymbol{x}}(\boldsymbol{b}^\top \boldsymbol{x}) = \frac{\partial}{\partial \boldsymbol{x}}(\boldsymbol{x}^\top \boldsymbol{b}) = \boldsymbol{b}$
- $\frac{\partial}{\partial \boldsymbol{x}}(\boldsymbol{b}^\top \boldsymbol{A}\boldsymbol{x}) = \boldsymbol{A}^\top \boldsymbol{b}$ • $\frac{\partial}{\partial \boldsymbol{X}}(\boldsymbol{c}^\top \boldsymbol{X}\boldsymbol{b}) = \boldsymbol{c}\boldsymbol{b}^\top$
- $\frac{\partial}{\partial \boldsymbol{X}}(\boldsymbol{c}^\top \boldsymbol{X}^\top \boldsymbol{b}) = \boldsymbol{b}\boldsymbol{c}^\top$ • $\frac{\partial}{\partial \boldsymbol{x}}(\boldsymbol{x}^\top \boldsymbol{x}) = 2\boldsymbol{x}$
- $\frac{\partial}{\partial \boldsymbol{x}}(\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}) = (\boldsymbol{A}^\top + \boldsymbol{A})\boldsymbol{x} \overset{A\text{ sym.}}{=} 2\boldsymbol{A}\boldsymbol{x}$
- $\frac{\partial}{\partial \boldsymbol{X}} Tr(\boldsymbol{X}^\top \boldsymbol{A}) = \boldsymbol{A}$ • Trace trick: $\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} = \dots$
  $\dots \overset{\text{inn. prod.}}{=} Tr(\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}) \overset{\text{cycl. permut.}}{=} Tr(\boldsymbol{x}\boldsymbol{x}^\top \boldsymbol{A}) = Tr(\boldsymbol{A}\boldsymbol{x}\boldsymbol{x}^\top)$
- $\sigma(x) = \frac{1}{1+\exp(-x)} \implies \nabla\sigma(x) = \sigma(x)(1-\sigma(x))$
- $\tanh(x) = \frac{2\sinh(x)}{2\cosh(x)} = \frac{e^x - e^{-x}}{e^x + e^{-x}}$ • $\nabla\tanh(x) = 1 - \tanh^2(x)$

2 Empirical Risk Minimisation (ERM)

Cost: $R(c,X,Y) = \sum_{i \le N} \|y_i - c^\top \boldsymbol{x}_i\|^2$ (regr.)
or $R(c,X,Y) = \sum_{i \le N} \max(0, -y_i c^\top \boldsymbol{x})$ (class.)
or $R(c,\theta,X) = \sum_{i \le N} \|\boldsymbol{x}_i - \theta_{c(i)}\|^2$ (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 $P(X \mid \theta)$, compute posterior $P(\theta \mid x_{1\dots n})$.

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

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

2.2 Linear Regression model: $\hat{\boldsymbol{y}} = \boldsymbol{X}\beta$

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

Lasso: $\hat{\beta} = \arg\min_\beta \sum_{i \le n}(y_i - x_i^\top\beta)^2 + \lambda\|\beta\|_2$
(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 $c \sim p(\cdot \mid X)$ s.t. $H[p(\cdot \mid x)]$ is maximal, $\mathbb{E}_{C|X}[R(C,X)] = \mu$ and $\sum_c p(c \mid X) = 1$.

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

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

$\iff$ $\boxed{p(c \mid X)} = \exp(-\beta[R(c,X) - F(X)])$

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

ME: $\boxed{\max H[c \mid X] \iff \max Z(X) \iff \min F(X)}$

- Exp. generalisation costs: $\mathbb{E}_{X''}\cancel{\mathbb{E}_{X'}}\overset{\mathbb{E}_C}{\mathbb{E}_{C|X'}}[R(c,X'')]$
- Min. out-of-sample descr. length per deg. of freedom
  $\min_{p(\cdot|\cdot)} \mathbb{E}_{X',X''}\mathbb{E}_{C|X'}\left[-\log\frac{p(c|X'')}{p(c)}\right]$ $p(c) = \mathbb{E}_X[p(c \mid X)]$
  $\overset{\text{Jensen}}{\ge} \min_{p(\cdot|\cdot)} \mathbb{E}_{X',X''}\left[-\log \mathbb{E}_{C|X'}[p(c \mid X'')]\right] - H[c]$
  $= \max_{p(\cdot|\cdot)} \mathbb{E}_{X',X''}[e^{H[c]} \cdot \kappa(X',X'')]$

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

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

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:** doesn't go "back&forth" forever.

$\implies$ **Stationary dist.** $\boxed{p(c') = \sum_c \pi(c \mid c')p(c)}$

$\iff$ **det. balance** $\boxed{\pi(c' \mid c)p(c) = \pi(c \mid c')p(c')}$

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

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

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

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

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

$\pi(c' \mid c) = \{\vdots$ ← c.f. scr. (2.7)

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

4.2 Laplace's Method (Least angle clust.)

1. *Square the cost:* $e^{-\frac{1}{T}R(c,X)} = const \cdot e^{g(c)^\top g(c)}$
2. *Complete the square:*
   $\int e^{-\frac{1}{T}(y - g(c))^2} \, dy = (\pi T)^{d/2}$
   $\Rightarrow e^{g(c)^\top g(c)} = (\pi T)^{-d/2} \int \exp^{-y^\top y + 2y^\top g(c)} \, dy$
3. *Rewrite normalisation constant:*
   $Z = \sum_c e^{-\frac{1}{T}R(c,X)} = \ldots = const \int e^{-\frac{1}{T}f(y)} \, dy$
4. *Apply Laplace's method:*
   If f has unique min. y_0 and Hessian $H := \frac{\partial^2 f}{\partial y^2}\big|_{y_0}$
   $\int e^{-\frac{1}{T}f(y)} \, dy \overset{(T \to 0)}{\approx} e^{-\frac{1}{T}f(y_0)} \left|\frac{H}{2\pi T}\right|^{-1/2}$

4.3 Mean-field Approximation

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

Approach: Minimise $D_{\mathrm{KL}}(p \,\|\, p_\beta)$
$\iff$ Minimise **Gibbs free energy:**

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

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

Ising model: $R(c \mid J) = -\frac{1}{2}\sum_{i,j} J_{ij} c_i c_j - \sum_i h_i c_i$
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* $h_i = [\cdots h_{i\ell} \cdots]^\top$

$h_{i\ell} := \frac{\partial \mathbb{E}[R(c)]}{\partial p_{i\ell}} = \mathbb{E}_{c \sim p_{|i \to \ell}}[R(c)]$ ← object i chooses class ℓ

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

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)

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

$\Longrightarrow h_{i\ell} = \|x_i - y_\ell\|^2 + \lambda \sum_{j \in N(i)} p_{j\ell} + const_i$

5 Deterministic Annealing (Z is tractable)

Lemma: func's × domain → domain × co-dom.
$\mathcal{O}(K^N) \rightarrow \sum_c \prod_i \epsilon_{i,c(i)} = \prod_i \sum_k \epsilon_{ik} \leftarrow \mathcal{O}(NK)$

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

Max. entr. $\Longrightarrow \frac{\partial \log Z}{\partial \theta_k} = 0 \Longrightarrow \theta_k^* = \frac{\sum_i p_i(k \mid \theta^*, X) \cdot \boldsymbol{x}_i}{\sum_i p_i(k \mid \theta^*, X)}$

do
  E-step: $p_i(k \mid \theta^{\text{old}}, X) = \frac{\exp(-\frac{1}{T}\|\boldsymbol{x}_i - \theta_k\|^2)}{\sum_{j \le K} \exp(-\frac{1}{T}\|\boldsymbol{x}_i - \theta_j\|^2)}$
  M-step: $\theta_k \leftarrow \ldots$
  $\theta^{\text{old}} \leftarrow \theta$
until convergence of θ
$\theta_k \leftarrow \theta_k + \epsilon$ (noise s.t. centroids can separate)

Phase transitions: For $T \to \infty$: $\theta_k^* = \overline{\boldsymbol{X}} \;\; \forall k \le K$
Once $T = 2\lambda_{\max}$, more centroids appear, where $\lambda_{\max}$ = max. eigenvalue of $\frac{1}{N}\boldsymbol{X}^\top \boldsymbol{X}$. ($\boldsymbol{x}_i$'s row-wise)

6 Histogram Clustering

Least Angle Clust. (LAC): [Idea]
Similarity $S(\boldsymbol{x}_i, \boldsymbol{x}_j) = w_{ij} \cos(\phi_{ij}) = w_{ij} \boldsymbol{e}_i \cdot \boldsymbol{e}_j$ with unit vectors $\boldsymbol{e}_i := \boldsymbol{x}_i / \|\boldsymbol{x}_i\|$, e.g. choice $w_{ij} = \|\boldsymbol{x}_i\| \cdot \|\boldsymbol{x}_j\|$.

Dyadic data: $\mathcal{Z} = \{(x_{i(r)}, y_{j(r)}); 1 \le r \le \ell\}$
- prototype / "centroid": $q(y_j \mid \alpha)$
- empirical dist.: $\hat{p}(y_j \mid x_i) = \frac{\hat{p}(x_i, y_j)}{\hat{p}(x_i)}$ ←scr. (5.10) ←scr. (5.11)

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

Assume $p(\alpha) = 1/k$ *and* $\hat{p}(x_i) = 1/n$
$\Rightarrow$ **Cost:** $R^{\text{hc}}(c, q, \mathcal{Z}) = \cancel{\tfrac{\ell}{n}} \sum_{i \le n} D_{\mathrm{KL}}[\hat{p}(\cdot \mid x_i) \,\|\, q(\cdot \mid c(i))]$

Solving the **Gibbs dist.** $p(c \mid q, \hat{p}) = \prod_{i \le n} P_{i,c(i)}$
via Lagrange yields $q^*(y_j \mid \alpha) = \frac{\sum_{i \le n} P_{i\alpha} \cdot \hat{p}(y_j | x_i)}{\sum_{i \le n} P_{i\alpha}}$

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

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

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

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

6.2 Parametric Distributional Clustering

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

$$p(y_j \mid \nu) \equiv p(b \mid \nu) = \sum_{\alpha \le s} p(\alpha | \nu) \, G_\alpha(b) \, .$$

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

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

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

Likelihood: (both equivalent if $p(i) = \frac{1}{n}$)
$P(X \mid c, \theta) = \prod_{i \le n} p(c(i)) \prod_{b \le m} [p(b \mid c(i))]^{\ell \hat{p}(i,b)}$,
$P(X, M \mid \theta) = \prod_{i \le n} \prod_{\nu \le k} \left[p(\nu) \cdot \prod_{b \le m} p(b \mid \nu)^{n_{ib}}\right]^{M_{i\nu}}$

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

$M_{i\nu} = p(\nu \mid i) \in \{0,1\}$ clust. membersh. assign.

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

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

E-step: $h_{i\nu} = -\log p_\nu - \sum_b \frac{\ell}{n}\hat{p}(b\mid i)\log p(b\mid \nu)$

$q_{i\nu} = \mathbb{E}[\mathbb{I}_{\{c(i)=\nu\}}] \propto \exp(-h_{i\nu}/T)$

M-step: $p_\nu = \frac{1}{n}\sum_{i\le n} q_{i\nu}$

No closed form sol. for $p(\alpha \mid \nu)$.
Thus, iteratively optimize pairs s.t. $\sum_\alpha p(\alpha\mid\nu) = 1$.

7 Graph-based Clustering

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

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

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

Correlation clustering:
Minimise the sum of *pairwise* intracluster distances.

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

$$= -2\underbrace{\sum_{\nu\le k}\sum_{(i,j)\in\mathcal{E}_{\nu\nu}} S_{ij}}_{\hookrightarrow\ \text{intra-cluster}} + \underbrace{\cancel{\sum_{(i,j)} S_{ij}}}_{\hookrightarrow\ \text{const}}$$

up to thresh. u $\overset{*}{=} -\frac{1}{2}\sum_{\nu\le k}\sum_{(i,j)\in\mathcal{E}_{\nu\nu}}(|S_{ij}-u| + S_{ij}-u)$
$+\frac{1}{2}\sum_{\nu\le k}\sum_{\substack{\mu\le k\\ \mu\neq\nu}}\sum_{(i,j)\in\mathcal{E}_{\nu\mu}}(|S_{ij}+u| - S_{ij}-u)$

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

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

$R^{\text{gp}}(c;\mathcal{D}) = const - \sum_{\nu\le k}\text{cut}(\mathcal{G}_\nu(\mathcal{D}), \mathcal{V}\setminus\mathcal{G}_\nu(\mathcal{D}))$
$= const + \sum_{\nu\le k}\text{cut}(\mathcal{G}_\nu(\mathcal{S}), \mathcal{V}\setminus\mathcal{G}_\nu(\mathcal{S}))$

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

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

7.1 Pairwise Clustering

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

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

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

Invariance properties:

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

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:

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

Reconstruction:

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

Cluster membership of new data:

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

1. $(S^{\text{new}})^{\text{c}} = -\frac{1}{2}\Big[D^{\text{new}}(\mathbb{I}_n - \frac{1}{n}\boldsymbol{e}_n\boldsymbol{e}_n^\top)$
   $- \frac{1}{n}\boldsymbol{e}_m\boldsymbol{e}_n^\top + \tilde{D}(\mathbb{I}_n - \frac{1}{n}\boldsymbol{e}_n\boldsymbol{e}_n^\top)\Big]$
2. Project: $X^{\text{new}}_p = (S^{\text{new}})^{\text{c}}\boldsymbol{V}_p(\Lambda_p)^{-1/2}$
3. Assign: $\hat{c}_i = \arg\min_c \left\|(x^{\text{new}}_p)_i - y_{c(i)}\right\|$

8 Model Selection for Clustering

What is the appropriate #clusters k for my data?

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

8.1 Complexity-based Model Selection

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

Attention: MDL/BIC rely on likelihood optimisation $\to$ <u>not generally applicable</u>

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

8.1.1 Min. Description Length (MDL)

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

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

8.1.2 Bayesian Information Crit. (BIC)

Parametrise likelihood $p(\boldsymbol{X}\mid M)$ by θ:
$p(\boldsymbol{X}\mid M) = \int_{\Theta_M}\exp(\log p(\boldsymbol{X}\mid M,\theta))\cdot p(\theta\mid M)\,\mathrm{d}\theta$

Assume flat prior $p(\theta|M) \approx const$ and expand log-likelihood by ML estimator $\hat{\theta}$:

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

$\implies p(\boldsymbol{X}\mid M) = const_2\cdot\exp(\,\ell(\hat{\theta}) - \frac{k'}{2}\log n\,)$
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:** $(\mathcal{S}, \{p(\cdot \mid s)\}_{s\in\mathcal{S}})$, $\mathcal{S}$: alphabet
  - ϵ-noisy binary channel: $p(\hat{s} \mid s) = \begin{cases} 1-\epsilon & \text{if } \hat{s}=s \\ \epsilon & \text{if } \hat{s}\neq s \end{cases}$
- **Capacity:** $\text{cap} = \max_p I(S;\hat{S}) \rightsquigarrow p_S(s)$
- **(M,n)-code:** is a pair (Enc, Dec) ← scr. p.87
  where M: #messages, n: code-length
  - **Rate:** $r = \frac{\log_2 M}{n} \Leftrightarrow M = \lfloor 2^{nr} \rfloor$
  - **Commu. err.:** $p_{\text{err}} := \max_{i\leq M} \mathbb{P}(Dec(\widehat{Enc(i)}) \neq i)$

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

Asymptotic equiparition property (AEP):

- $A_\epsilon^{(n)}$: Typical set of sequences $(s_1,\ldots,s_n) \in \mathcal{S}^n$
  $\left|-\frac{1}{n}\log p_{S^n}(\boldsymbol{s}^n) - H[S]\right| < \epsilon$ ← scr. p.89
- $\mathbb{P}\left((\boldsymbol{S}^n, \hat{\boldsymbol{S}}^n) \in A_\epsilon^{(n)}\right) \overset{n\to\infty}{\to} 1$ ← scr. p.90
- $p_{\text{err}} \leq 2^{-n(\text{cap}-3\epsilon-r)} \overset{n\to\infty}{\to} 0$ 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. $p(\cdot \mid X')$

Ideal variant:

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

Code: $X'_i \xrightarrow{Enc_\mathcal{A}} p(\cdot \mid X'_i) \xrightarrow{\mathcal{C}_\mathcal{A}} p(\cdot \mid X''_i) \xrightarrow{Dec_\mathcal{A}} \hat{X}$

Empirical variant:

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

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

Code: $\tau_i \xrightarrow{Enc} p(\cdot \mid \tau_i \circ X') \xrightarrow{\mathcal{C}_\mathcal{A}} p(\cdot \mid \tau_i \circ X'') \xrightarrow{Dec} \hat{\tau}$

- $Enc_\mathcal{A}$: encodes $\tau_i \in \mathcal{M}$ as $p(\cdot \mid \tau_i \circ X')$
- $Dec_\mathcal{A}$: selects $\hat{\tau} = \arg\max_\tau \kappa(\tau_i \circ X'', \tau \circ X')$

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

Asymptotic Equipartition Property (AEP):

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

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) $\overset{n\to\infty}{\to} 0$ 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.* → quantified by **capacity** (max. # distinguishable messages that can be communicated)

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

Cost functions: Given $R_1(\cdot,\cdot), \ldots, R_s(\cdot,\cdot)$

$\max_{\ell\leq s} \kappa_\ell(X', X'') = \max_{\ell\leq s} \frac{1}{Z_{X'}Z_{X''}} \sum_c e^{-\frac{1}{T}R_\ell(c,X')} e^{-\frac{1}{T}R_\ell(c,X'')}$

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

Approximation Set Coding (ASC):

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

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

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

10 Appendix

10.1 Tips and Tricks

Complete the square:

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

Constrained optimisation:

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

Lagrangian: with each $\alpha_j \geq 0$

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

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

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

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

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

Euler-Lagrange: Find extrema of functional $\mathcal{F}[f] = \int G(x, f(x), f(x))\,\mathrm{d}x$, thus $\frac{\partial\mathcal{F}}{\partial f} \overset{!}{=} 0$.

If G is twice diff'able, then

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

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

10.2 Approximations

Laplace Approximation: $\left.\frac{\mathrm{d}f}{\mathrm{d}x}\right|_{x_0} = 0$

$\Longrightarrow \int_\mathbb{R} e^{Cf(x)}\,\mathrm{d}x \approx \sqrt{2\pi}C \cdot |f''(x_0)| \cdot e^{Cf(x_0)}$