Statistical Learning Theory Cheat Sheet

Basics

1. $\mathcal{N} \sim (2\pi)^{-\frac{d}{2}} \det(\mathbf{\Sigma})^{-\frac{1}{2}} \exp(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\mathsf{T}\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}))$
2. $p(z) \propto exp\big(-\dfrac{1}{2}z^T\Lambda z + z^T m\big)$ and Λ p.d. $\implies p(z) \sim \mathcal{N}(\Lambda^{-1}m, \Lambda^{-1})$
3. $\int_{-\infty}^{+\infty} e^{-a(x+b)^2}dx = \frac{\sqrt{\pi}}{\sqrt{a}}$
4. $p_\beta(c) = exp[-\beta(R(c) - \mathcal{F}(\beta))]$
5. $\mathcal{F}(\beta) = -\dfrac{1}{\beta}\log \mathcal{Z}$

Matrix Identities

1. $Tr(A) = \sum_i A_{ii} = \sum_i \lambda_i$, $Tr(AB) = Tr(BA)$
2. $x^T Ax = Tr(x^T Ax) = Tr(Axx^T)$

Vector Calculus

1. $\frac{\partial}{\partial x}x^T Ax = (A + A^T)x = 2Ax$ for A sym.
2. $\frac{\partial}{\partial A}|A| = |A|A^{-T}$
3. $\frac{\partial}{\partial A}Tr(A^T B) = B$

Euler-Lagrange Equation

1. $F[y] = \int G(y(x), y'(x), x)dx$
2. $\dfrac{\delta F}{\delta y(x)} = \dfrac{\partial G}{\partial y} - \dfrac{d}{dx}\dfrac{\partial L}{\partial y'}$

Hyperbolic Functions

1. $sinh(x) = \dfrac{e^x - e^{-x}}{2} = \dfrac{e^{2x}-1}{2e^x} = \dfrac{1-e^{-2x}}{2e^{-x}}$
2. $cosh(x) = \dfrac{e^x + e^{-x}}{2} = \dfrac{e^{2x}+1}{2e^x} = \dfrac{1+e^{-2x}}{2e^{-x}}$
3. $tanh(x) = \dfrac{e^x - e^{-x}}{e^x + e^{-x}} = \dfrac{e^{2x}-1}{e^{2x}+1}$

Information Theory Inequalities

1. $H(X, Y) = H(X) + H(Y|X)$
2. $I(X; Y) = H(X) - H(X|Y)$
3. $I(X; Y|Z) = H(X|Z) - H(X|Y, Z)$
4. $I(X, Y; Z) = I(X; Z) + I(Y; Z|X)$
5. $I(X; Y, Z) = I(X; Z) + I(X; Y|Z)$
6. $I(X; Y, Z) = I(X; Y) + I(X; Z|Y)$
7. $H(g(X)) \leq H(X)$ (X discrete, $g : \mathbb{R} \to \mathbb{R}$)
8. ψ convex $\implies \psi(\mathbb{E}[X]) \leq \mathbb{E}[\psi(X)]$

Maximum Entropy Distributions

1. non-negative r.v. with mean $\mu \implies$ **Exponential**
2. mean μ and variance $\sigma \implies$ **Gaussian**
3. r.v $\sim \mathcal{N}$ with random variance $\implies$ **Laplace**

Markov Chain Montecarlo

1. **Irreducibility**: for all $c, c' \in C$ there is a path $c_0, \ldots, c_n$ of length n, connecting c to c' with non-zero probability.
2. **Aperiodicity**: for all $c, c' \in C$ one of the following holds: (i) there is an integer $n(c, c')$ such that, for any $n > n(c, c')$ there is a path of length n connecting c to c' with non-zero probability, or (ii) there is no path connecting c to c' with non-zero probability. [(ii) can be dropped if we assume irreducibility]
3. **Stationarity** with respect to distribution π: $\sum_{c\in C}\pi(c)P(c', c) = \pi(c')$
4. **Mixing Time**: $t \propto \dfrac{1}{\lambda_1 - \lambda_2}$ where $\lambda_1 = 1$ and $\lambda_2 \leq \lambda_1$ are the eigenvalues of P. ($\lambda_1 = 1$ is always the biggest eigenvalue of P)
5. **Detailed Balance**: $\forall c, c' \quad P(c'|c)\pi(c) = P(c|c')\pi(c')$
6. **Detailed Balance $\implies$ Stationarity**
7. $lim_{t\to\infty}\dfrac{1}{t}\sum_{s=1}^{t} f(X_s) = \sum_c \pi(c)f(c)$

Constant Shift Embedding

1. Symmetrize: $D' \leftarrow \dfrac{1}{2}(D + D^T)$
2. Centralize: $D^C \leftarrow QD'Q$
3. Similarities: $S^C \leftarrow -\dfrac{1}{2}D^C$
4. Off-diagonal Shift: $\tilde{S}_{ii} \leftarrow S^C_{ii} - \lambda_{min}(S^C)$
5. $\tilde{D}_{ij} \leftarrow \tilde{S}_{ii} + \tilde{S}_{jj} - 2\tilde{S}_{ij}$
6. $\tilde{S}^C \leftarrow -\dfrac{1}{2}\tilde{D}^C$
7. $\tilde{S}^C = V\Lambda V^T$
8. $X_p = V_p\Lambda_p^{1/2}$

Parametric Distributional Clustering

1. Replace the non-parametric density estimation via histograms by a continuos mixture model
2. Gaussian prototype $G_\alpha(j) = \int_{I_j} g_\alpha(x)dx$ are used to create mixture densities : $p(y|\nu) = \sum_\alpha p(\alpha|\nu)G_\alpha(y)$
3. $\Theta = \{p_\nu, p_{\alpha|\nu}, \mu_\alpha \mid \alpha = 1, \ldots l; \nu = 1, \ldots, k\}$
4. $P(X, M|\Theta) = \prod_{i=1}^{n}\prod_{\nu=1}^{k}[p_\nu \prod_{j=1}^{m} p(y_j|\nu)^{n_{ij}}]^{M_{iv}}$
5. $h_{i\nu} = -\log p_\nu - \sum_j n_{ij}\log\big(\sum_\alpha p_{\alpha|\nu}G_\alpha(j)\big)$
6. $q_{i\nu} \propto \exp\big(-\dfrac{1}{T}h_{i\nu}\big)$
7. $p_\nu = \dfrac{1}{n}\sum_{i=1}^{n} q_{i\nu}$, $\nu = 1, \ldots k$
8. No closed form solution for $p_{\alpha|\nu}, \mu_\alpha \implies$ numerical methods

Information Bottleneck

1. Minimize w.r.t $q(c|x)$: $L = I(X, C) - \beta I(C, Y)$
2. Minimize w.r.t $q(c|x)$: $I(X, C)$ s.t. $\mathbb{E}[d(x, c)] \leq D$

Mean Field Approximation

1. $G(p_0) = \dfrac{1}{\beta}D_{KL}(p_0||p_\beta) + \mathcal{F}(\beta)$
2. $G(p_0) = \mathbb{E}_{c\sim p_0}[\mathcal{R}(c)] - \dfrac{1}{\beta}H[p_0]$
3. $G(p_0) = \mathbb{E}_{c\sim p_0}[\mathcal{R}(c)] + \mathcal{F}_{p_0} - \mathbb{E}_{c\sim p_0}[\mathcal{R}_0(c)]$
4. **Ising Model**: $E(\sigma) = -\lambda\sum_i \sigma_i h_i - \sum_{i,j} J_{i,j}\sigma_i\sigma_j$