总结

Math

1. MLE

$$\theta_{MLE}=\mathop{argmax}_{\theta}\log p(X|\theta)\underset{iid}{=}\mathop{argmax}_{\theta}\sum_{i=1}^{N}\log p(x_i|\theta) \tag{1}$$

2. MAP

$$\theta_{MAP}=\mathop{argmax}_{\theta}p(\theta|X)=\mathop{argmax}_{\theta}p(X|\theta)\cdot p(\theta) \tag{2}$$

3. Gaussian Distribution

$$p(x|\mu,\Sigma)=\frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}}e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)} \tag{3}$$

$$\Delta=(x-\mu)^T\Sigma^{-1}(x-\mu)=\sum_{i=1}^{p}(x-\mu)^Tu_i\frac{1}{\lambda_i}u_i^T(x-\mu)=\sum_{i=1}^{p}\frac{y_i^2}{\lambda_i} \tag{4}$$

4. 已知 $x\sim\mathcal{N}(\mu,\Sigma),y\sim Ax+b$，有：

$$y\sim\mathcal{N}(A\mu+b,A\Sigma A^T) \tag{5}$$

5. 记 $x=(x_1,x_2,\cdots,x_p)^T=(x_{a,m\times1},x_{b,n\times1})^T,\mu=(\mu_{a,m\times1},\mu_{b,n\times1}),\Sigma=\begin{pmatrix}\Sigma_{aa}&\Sigma_{ab}\\\Sigma_{ba}&\Sigma_{bb}\end{pmatrix}$，
   已知 $x\sim\mathcal{N}(\mu,\Sigma)$，则：

$$x_a\sim\mathcal{N}(\mu_a,\Sigma_{aa}) \tag{6}$$

$$x_b|x_a\sim\mathcal{N}(\mu_{b|a},\Sigma_{b|a}) \tag{7}$$

$$\mu_{b|a}=\Sigma_{ba}\Sigma_{aa}^{-1}(x_a-\mu_a)+\mu_b \tag{8}$$

$$\Sigma_{b|a}=\Sigma_{bb}-\Sigma_{ba}\Sigma_{aa}^{-1}\Sigma_{ab} \tag{9}$$

Linear Regression

Model

1. Dataset:

$$\mathcal{D} = \{(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)\} \tag{10}$$

2. Notation:

$$X = (x_1, x_2, \cdots, x_N)^T, Y = (y_1, y_2, \cdots, y_N)^T \tag{11}$$

3. Model:

$$f(w) = w^T x \tag{12}$$

Loss Function

1. 最小二乘误差/高斯噪声的MLE

$$L(w) = \sum_{i=1}^{N} ||w^T x_i - y_i||_2^2 \tag{13}$$

闭式解

$$\hat{w} = (X^T X)^{-1} X^T Y = X^+ Y \tag{14}$$
$$X = U \Sigma V^T \tag{15}$$
$$X^+ = V \Sigma^{-1} U^T \tag{16}$$

正则化

$$L1 - Gaussian\ priori : \underset{w}{argmin}\, L(w) + \lambda ||w||_1, \lambda > 0 \tag{17}$$
$$L2 - Laplasian\ priori - Sparsity : \underset{w}{argmin}\, L(w) + \lambda ||w||_2^2, \lambda > 0 \tag{18}$$

Linear Classification

Hard

PCA

1. Idea: 在线性模型上加入激活函数
2. Loss Function:

$$L(w) = \sum_{x_i \in \mathcal{D}_{wrong}} -y_i w^T x_i \tag{19}$$

3. Parameters:

$$w^{t+1} \leftarrow w^t + \lambda y_i x_i \tag{20}$$

Fisher

1. Idea: 投影，类内小，类间大。
2. Loss Function:

$$J(w) = \frac{w^T S_b w}{w^T S_w w} \tag{21}$$
$$S_b = (\overline{x_{c1}} - \overline{x_{c2}})(\overline{x_{c1}} - \overline{x_{c2}})^T \tag{22}$$
$$S_w = S_1 + S_2 \tag{23}$$

3. 闭式解，投影方向:

$$S_w^{-1}(\overline{x_{c1}} - \overline{x_{c2}}) \tag{24}$$

Soft

判别模型

Logistic Regression

1. Idea，激活函数:

$$p(C_1|x) = \frac{1}{1 + \exp(-a)} \tag{25}$$
$$a = w^T x \tag{26}$$

2. Loss Function(交叉熵):

$$\hat{w}=\underset{w}{argmax}\,J(w)=\underset{w}{argmax}\sum_{i=1}^N(y_i\log p_1+(1-y_i)\log p_0) \tag{27}$$

3. 解法，SGD

$$J'(w)=\sum_{i=1}^N(y_i-p_1)x_i \tag{28}$$

生成模型

GDA

1. Model
    1. $y\sim Bernoulli(\phi)$
    2. $x|y=1\sim\mathcal{N}(\mu_1,\Sigma)$
    3. $x|y=0\sim\mathcal{N}(\mu_0,\Sigma)$
2. MAP

$$\begin{aligned}&\underset{\phi,\mu_0,\mu_1,\Sigma}{argmax}\log p(X|Y)p(Y)\\&=\underset{\phi,\mu_0,\mu_1,\Sigma}{argmax}\sum_{i=1}^N((1-y_i)\log\mathcal{N}(\mu_0,\Sigma)+y_i\log\mathcal{N}(\mu_1,\Sigma)+y_i\log\phi+(1-y_i)\log(1-\phi))\end{aligned} \tag{29}$$

3. 解

$$\phi=\frac{N_1}{N} \tag{30}$$
$$\mu_1=\frac{\sum\limits_{i=1}^Ny_ix_i}{N_1} \tag{31}$$
$$\mu_0=\frac{\sum\limits_{i=1}^N(1-y_i)x_i}{N_0} \tag{32}$$
$$\Sigma=\frac{N_1S_1+N_2S_2}{N} \tag{33}$$

Naive Bayesian

1. Model, 对单个数据点的各个维度作出限制

$$x_i \perp x_j|y, \forall\, i \neq j \tag{34}$$

   1. x_i 为连续变量：$p(x_i|y)=\mathcal{N}(\mu_i,\sigma_i^2)$
   2. x_i 为离散变量：类别分布（Categorical）：$p(x_i=i|y)=\theta_i,\sum\limits_{i=1}^K\theta_i=1$
   3. $p(y)=\phi^y(1-\phi)^{1-y}$
2. 解：和GDA相同

Dimension Reduction

中心化：

$$\begin{aligned}S&=\frac{1}{N}X^T(E_N-\frac{1}{N}\mathbb{I}_{N1}\mathbb{I}_{1N})(E_N-\frac{1}{N}\mathbb{I}_{N1}\mathbb{I}_{1N})^TX\\&=\frac{1}{N}X^TH^2X=\frac{1}{N}X^THX\end{aligned} \tag{35}$$

PCA

1. Idea: 坐标变换，寻找线性无关的新基矢，取信息损失最小的前几个维度
2. Loss Function:

$$J=\sum_{j=1}^qu_j^TSu_j\ ,\ s.t.\ u_j^Tu_j=1 \tag{36}$$

3. 解：
   1. 特征分解法

$$S=U\Lambda U^T \tag{37}$$

   2. SVD for X/S

$$HX=U\Sigma V^T \tag{38}$$

$$S=\frac{1}{N}V\Sigma^T\Sigma V^T \tag{39}$$

$$new\ co=HX\cdot V \tag{40}$$

   3. SVD for T

$$T = HXX^TH = U\Sigma\Sigma^TU^T \tag{41}$$
$$new\ co = U\Sigma \tag{42}$$

p-PCA

1. Model:

$$z \sim \mathcal{N}(\mathbb{O}_{q1}, \mathbb{I}_{qq}) \tag{43}$$
$$x = Wz + \mu + \varepsilon \tag{44}$$
$$\varepsilon \sim \mathcal{N}(0, \sigma^2\mathbb{I}_{pp}) \tag{45}$$

2. Learning: E-M
3. Inference:

$$p(z|x) = \mathcal{N}(W^T(WW^T + \sigma^2\mathbb{I})^{-1}(x-\mu), \mathbb{I} - W^T(WW^T + \sigma^2\mathbb{I})^{-1}W) \tag{46}$$

SVM

1. 强对偶关系：凸优化+（松弛）Slater 条件->强对偶。
2. 参数求解：KKT条件
   1. 可行域
   2. 互补松弛+梯度为0

Hard-margin

1. Idea: 最大化间隔
2. Model:

$$\underset{w,b}{argmin}\,\frac{1}{2}w^Tw\ s.t.\ y_i(w^Tx_i + b) \geq 1, i = 1, 2, \cdots, N \tag{47}$$

3. 对偶问题

$$\max_{\lambda} -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\lambda_i\lambda_jy_iy_jx_i^Tx_j + \sum_{i=1}^{N}\lambda_i,\ s.t.\ \lambda_i \geq 0 \tag{48}$$

4. 模型参数

$$\hat{w} = \sum_{i=1}^{N} \lambda_i y_i x_i \tag{49}$$

$$\hat{b} = y_k - w^T x_k = y_k - \sum_{i=1}^{N} \lambda_i y_i x_i^T x_k, \exists k, 1 - y_k(w^T x_k + b) = 0$$

Soft-margin

1. Idea:允许少量错误
2. Model:

$$error = \sum_{i=1}^{N} \max\{0, 1 - y_i(w^T x_i + b)\} \tag{50}$$

$$\underset{w,b}{argmin} \frac{1}{2} w^T w + C \sum_{i=1}^{N} \xi_i \ s.t. \ y_i(w^T x_i + b) \geq 1 - \xi_i, \xi_i \geq 0, i = 1, 2, \cdots, N$$

Kernel

对称的正定函数都可以作为正定核。

Exp Family

1. 表达式

$$p(x|\eta) = h(x)\exp(\eta^T \phi(x) - A(\eta)) = \frac{1}{\exp(A(\eta))} h(x)\exp(\eta^T \phi(x)) \tag{51}$$

2. 对数配分函数

$$A'(\eta) = \mathbb{E}_{p(x|\eta)}[\phi(x)] \tag{52}$$
$$A''(\eta) = Var_{p(x|\eta)}[\phi(x)] \tag{53}$$

3. 指数族分布满足最大熵定理

PGM

Representation

1. 有向图

$$p(x_1,x_2,\cdots,x_p)=\prod_{i=1}^{p}p(x_i|x_{parent(i)}) \tag{54}$$

   D-separation

$$p(x_i|x_{-i})=\frac{p(x)}{\int p(x)dx_i}=\frac{\prod_{j=1}^{p}p(x_j|x_{parents(j)})}{\int\prod_{j=1}^{p}p(x_j|x_{parents(j)})dx_i}=\frac{p(x_i|x_{parents(i)})p(x_{child(i)}|x_i)}{\int p(x_i|x_{parents(i)})p(x_{child(i)}|x_i)dx_i} \tag{55}$$

2. 无向图

$$p(x)=\frac{1}{Z}\prod_{i=1}^{K}\phi(x_{ci}) \tag{56}$$

$$Z=\sum_{x\in\mathcal{X}}\prod_{i=1}^{K}\phi(x_{ci}) \tag{57}$$

$$\phi(x_{ci})=\exp(-E(x_{ci})) \tag{58}$$

3. 有向转无向
   1. 将每个节点的父节点两两相连
   2. 将有向边替换为无向边

Learning

参数学习-EM

1. 目的：解决具有隐变量的混合模型的参数估计（极大似然估计）
2. 参数：

$$\theta_{MLE}=\underset{\theta}{argmax}\log p(x|\theta) \tag{59}$$

3. 迭代求解：

$$\theta^{t+1} = \underset{\theta}{argmax} \int_z \log[p(x,z|\theta)]p(z|x,\theta^t)dz = \mathbb{E}_{z|x,\theta^t}[\log p(x,z|\theta)] \tag{60}$$

4. 原理

$$\log p(x|\theta^t) \leq \log p(x|\theta^{t+1}) \tag{61}$$

5. 广义EM
   1. E step：

$$\hat{q}^{t+1}(z) = \underset{q}{argmax} \int_z q^t(z) \log \frac{p(x,z|\theta)}{q^t(z)} dz, fixed\ \theta \tag{62}$$

   2. M step：

$$\hat{\theta} = \underset{\theta}{argmax} \int_z q^{t+1}(z) \log \frac{p(x,z|\theta)}{q^{t+1}(z)} dz, fixed\ \hat{q} \tag{63}$$

Inference

1. 精确推断
   1. VE
   2. BP

$$m_{j\to i}(i) = \sum_j \phi_j(j)\phi_{ij}(ij) \prod_{k\in Neighbour(j)-i} m_{k\to j}(j) \tag{64}$$

   3. MP

$$m_{j\to i} = \max_j \phi_j \phi_{ij} \prod_{k\in Neighbour(j)-i} m_{k\to j} \tag{65}$$

2. 近似推断
   1. 确定性近似，VI
      1. 变分表达式

$$\hat{q}(Z)=\underset{q(Z)}{argmax}\,L(q) \tag{66}$$

2. 平均场近似下的 VI-坐标上升

$$\mathbb{E}_{\prod\limits_{i\neq j}q_i(Z_i)}[\log p(X,Z)]=\log\hat{p}(X,Z_j) \tag{67}$$
$$q_j(Z_j)=\hat{p}(X,Z_j)$$

3. SGVI-变成优化问题，重参数法

$$\begin{aligned}\underset{q(Z)}{argmax}\,L(q)=\underset{\phi}{argmax}\,L(\phi)\\ \nabla_\phi L(\phi)=\mathbb{E}_{q_\phi}[(\nabla_\phi\log q_\phi)(\log p_\theta(x^i,z)-\log q_\phi(z))]\\ =\mathbb{E}_{p(\varepsilon)}[\nabla_z[\log p_\theta(x^i,z)-\log q_\phi(z)]\nabla_\phi g_\phi(\varepsilon,x^i)]\\ z=g_\phi(\varepsilon,x^i),\varepsilon\sim p(\varepsilon)\end{aligned} \tag{68}$$

2. 随机性近似
   1. 蒙特卡洛方法采样
      1. CDF 采样
      2. 拒绝采样， $q(z)$，使得 $\forall z_i,Mq(z_i)\geq p(z_i)$，拒绝因子：$\alpha=\frac{p(z^i)}{Mq(z^i)}\leq 1$
      3. 重要性采样

$$\mathbb{E}_{p(z)}[f(z)]=\int p(z)f(z)dz=\int\frac{p(z)}{q(z)}f(z)q(z)dz\simeq\frac{1}{N}\sum_{i=1}^N f(z_i)\frac{p(z_i)}{q(z_i)} \tag{69}$$

      4. 重要性重采样：重要性采样+重采样
   2. MCMC：构建马尔可夫链概率序列，使其收敛到平稳分布 $p(z)$。
      1. 转移矩阵（提议分布）

$$\begin{aligned}p(z)\cdot Q_{z\to z^*}\alpha(z,z^*)=p(z^*)\cdot Q_{z^*\to z}\alpha(z^*,z)\\ \alpha(z,z^*)=\min\{1,\frac{p(z^*)Q_{z^*\to z}}{p(z)Q_{z\to z^*}}\}\end{aligned} \tag{70}$$

      2. 算法（MH）：
         1. 通过在0，1之间均匀分布取点 u

        2. 生成 $z^* \sim Q(z^*|z^{i-1})$
        3. 计算 α 值
        4. 如果 $\alpha \geq u$，则 $z^i = z^*$，否则 $z^i = z^{i-1}$
    3. Gibbs 采样：给定初始值 $z_1^0, z_2^0, \cdots$在 $t+1$ 时刻，采样 $z_i^{t+1} \sim p(z_i|z_{-i})$，从第一个维度一个个采样。

GMM

1. Model

$$p(x) = \sum_{k=1}^{K} p_k \mathcal{N}(x|\mu_k, \Sigma_k) \tag{71}$$

2. 求解-EM

$$\begin{aligned}
Q(\theta, \theta^t) &= \sum_z [\log \prod_{i=1}^{N} p(x_i, z_i|\theta)] \prod_{i=1}^{N} p(z_i|x_i, \theta^t) \\
&= \sum_z [\sum_{i=1}^{N} \log p(x_i, z_i|\theta)] \prod_{i=1}^{N} p(z_i|x_i, \theta^t) \\
&= \sum_{i=1}^{N} \sum_{z_i} \log p(x_i, z_i|\theta) p(z_i|x_i, \theta^t) \\
&= \sum_{i=1}^{N} \sum_{z_i} \log p_{z_i} \mathcal{N}(x_i|\mu_{z_i}, \Sigma_{z_i}) \frac{p_{z_i}^t \mathcal{N}(x_i|\mu_{z_i}^t, \Sigma_{z_i}^t)}{\sum\limits_k p_k^t \mathcal{N}(x_i|\mu_k^t, \Sigma_k^t)}
\end{aligned} \tag{72}$$

$$p_k^{t+1} = \frac{1}{N} \sum_{i=1}^{N} p(z_i = k|x_i, \theta^t) \tag{73}$$

序列模型-HMM，LDS，Particle

1. 假设：
    1. 齐次 Markov 假设（未来只依赖于当前）：

$$p(i_{t+1}|i_t, i_{t-1}, \cdots, i_1, o_t, o_{t-1}, \cdots, o_1) = p(i_{t+1}|i_t) \tag{74}$$

    2. 观测独立假设：

$$p(o_t|i_t, i_{t-1}, \cdots, i_1, o_{t-1}, \cdots, o_1) = p(o_t|i_t) \tag{75}$$

2. 参数

$$\lambda = (\pi, A, B) \tag{76}$$

离散线性隐变量-HMM

1. Evaluation：$p(O|\lambda)$，Forward-Backward 算法

$$p(O|\lambda) = \sum_{i=1}^{N} p(O, i_T = q_i|\lambda) = \sum_{i=1}^{N} \alpha_T(i) = \sum_{i=1}^{N} b_i(o_1)\pi_i\beta_1(i) \tag{77}$$

$$\alpha_{t+1}(j) = \sum_{i=1}^{N} b_j(o_t)a_{ij}\alpha_t(i)$$

$$\beta_t(i) = \sum_{j=1}^{N} b_j(o_{t+1})a_{ij}\beta_{t+1}(j)$$

2. Learning：$\lambda = \underset{\lambda}{argmax}\, p(O|\lambda)$，EM 算法（Baum-Welch）

$$\begin{aligned}\lambda^{t+1} &= \underset{\lambda}{argmax} \sum_I \log p(O, I|\lambda)p(O, I|\lambda^t) \\ &= \sum_I [\log \pi_{i_1} + \sum_{t=2}^{T} \log a_{i_{t-1}, i_t} + \sum_{t=1}^{T} \log b_{i_t}(o_t)]p(O, I|\lambda^t)\end{aligned} \tag{78}$$

3. Decoding：$I = \underset{I}{argmax}\, p(I|O, \lambda)$，Viterbi 算法-动态规划

$$\delta_t(j) = \max_{i_1, \cdots, i_{t-1}} p(o_1, \cdots, o_t, i_1, \cdots, i_{t-1}, i_t = q_i) \tag{79}$$

$$\delta_{t+1}(j) = \max_{1\le i\le N} \delta_t(i)a_{ij}b_j(o_{t+1})$$

$$\psi_{t+1}(j) = \underset{1\le i\le N}{argmax}\, \delta_t(i)a_{ij}$$

连续线性隐变量-LDS

1. Model

$$
\begin{aligned}
p(z_t|z_{t-1}) \sim \mathcal{N}(A\cdot z_{t-1}+B,Q) \quad (80)\\
p(x_t|z_t) \sim \mathcal{N}(C\cdot z_t+D,R) \quad (81)\\
z_1 \sim \mathcal{N}(\mu_1,\Sigma_1) \quad (82)
\end{aligned}
$$

2. 滤波

$$
\begin{aligned}
p(z_t|x_{1:t}) &= p(x_{1:t},z_t)/p(x_{1:t}) \propto p(x_{1:t},z_t) \quad (83)\\
&= p(x_t|z_t)p(z_t|x_{1:t-1})p(x_{1:t-1}) \propto p(x_t|z_t)p(z_t|x_{1:t-1})
\end{aligned}
$$

3. 递推求解-线性高斯模型
   1. Prediction

$$
p(z_t|x_{1:t-1}) = \int_{z_{t-1}} p(z_t|z_{t-1})p(z_{t-1}|x_{1:t-1})dz_{t-1} = \int_{z_{t-1}} \mathcal{N}(Az_{t-1}+B,Q)\mathcal{N}(\mu_{t-1},\Sigma_{t-1})dz_{t-1} \quad (84)
$$

   2. Update:

$$
p(z_t|x_{1:t}) \propto p(x_t|z_t)p(z_t|x_{1:t-1} \quad (85)
$$

连续非线性隐变量-粒子滤波

通过采样(SIR)解决：

$$
\mathbb{E}[f(z)] = \int_z f(z)p(z)dz = \int_z f(z)\frac{p(z)}{q(z)}q(z)dz = \sum_{i=1}^{N} f(z_i)\frac{p(z_i)}{q(z_i)} \quad (86)
$$

1. 采样

$$
\begin{aligned}
w_t^i \propto \frac{p(x_t|z_t)p(z_t|z_{t-1})}{q(z_t|z_{1:t-1},x_{1:t})}w_{t-1}^i \quad (87)\\
q(z_t|z_{1:t-1},x_{1:t}) = p(z_t|z_{t-1})
\end{aligned}
$$

2. 重采样

CRF

1. PDF

$$p(Y=y|X=x)=\frac{1}{Z(x,\theta)}\exp[\theta^T H(y_t,y_{t-1},x)] \tag{88}$$

2. 边缘概率

$$\begin{aligned} p(y_t=i|x) &= \sum_{y_{1:t-1}}\sum_{y_{t+1:T}}\frac{1}{Z}\prod_{t'=1}^{T}\phi_{t'}(y_{t'-1},y_{t'},x) \\ p(y_t=i|x) &= \frac{1}{Z}\Delta_l\Delta_r \\ \Delta_l &= \sum_{y_{1:t-1}}\phi_1(y_0,y_1,x)\phi_2(y_1,y_2,x)\cdots\phi_{t-1}(y_{t-2},y_{t-1},x)\phi_t(y_{t-1},y_t=i,x) \\ \Delta_r &= \sum_{y_{t+1:T}}\phi_{t+1}(y_t=i,y_{t+1},x)\phi_{t+2}(y_{t+1},y_{t+2},x)\cdots\phi_T(y_{T-1},y_T,x) \end{aligned} \tag{89}$$

$$\begin{aligned} \alpha_t(i)=\Delta_l &= \sum_{j\in S}\phi_t(y_{t-1}=j,y_t=i,x)\alpha_{t-1}(j) \\ \Delta_r=\beta_t(i) &= \sum_{j\in S}\phi_{t+1}(y_t=i,y_{t+1}=j,x)\beta_{t+1}(j) \end{aligned} \tag{90}$$

3. 学习

$$\nabla_\lambda L=\sum_{i=1}^{N}\sum_{t=1}^{T}[f(y_{t-1},y_t,x^i)-\sum_{y_{t-1}}\sum_{y_t}p(y_{t-1},y_t|x^i)f(y_{t-1},y_t,x^i)] \tag{91}$$