

线性回归模型

Maximum likelihood estimation (MLE): Frequentist

$$\text{Loss: } \|y - X\theta\|_2^2 \quad \hat{\theta} = (X^T X)^{-1} X^T y \quad E(Y|X)$$

从投影角度理解: $X^T(y - \hat{X}\theta) = 0$

$$y = X^T \theta + \varepsilon \\ \varepsilon \sim N(0, \sigma^2 I)$$

Maximum a posteriori (MAP) \rightarrow 岭回归

$$\theta \sim N(0, \sigma^2 I) \quad \text{Loss: } \|y - X\theta\|_2^2 + \frac{\lambda^2}{\sigma^2} \|\theta\|_2^2$$

$$\hat{\theta} = (X^T X + \lambda I)^{-1} X^T y$$

复数多元正态分布: $N(\mu, \Sigma)$

$$f(x) = \frac{1}{(2\pi)^{n/2} |\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)}$$

2范数

正则化 $\|x\|_p$ $P=2$ 岭回归 \rightarrow Tikhonov: $\min_{\theta} \|y - X\theta\|_p^2 + \lambda \|\theta - \theta_0\|_2^2$

$$\varepsilon \sim N(0, (P^T P)^{-1})$$

$$\theta \sim N(\theta_0, (\lambda Q^T Q)^{-1})$$

$P=1$ Lasso 回归: $\frac{1}{2} \|y - X\theta\|_2^2 + \lambda \|\theta\|_1$

$$\theta \sim \text{Laplace}(0, b)$$

Iterative soft-thresholding for general situations:

$$\hat{\theta}_d^{t+1} = \arg \min_{\theta} \frac{1}{2} \left(\frac{1}{\|x_d\|_2} \|y - X_d \theta\|_2^2 + \frac{\lambda \|\theta\|_1}{\|x_d\|_2} \right)$$

$$= \underbrace{\frac{\lambda}{\|x_d\|_2}}_{\gamma} \left(\frac{x_d^T (y - X_d \hat{\theta}_d^t)}{\|x_d\|_2} \right)$$

stronger sparsity: $\|y - X\theta\|_2^2 + \lambda \|\theta\|_1$

weaker sparsity: $\|y - X\theta\|_2^2 + \lambda_1 \|\theta\|_1 + \lambda_2 \|\theta\|_2^2$

$$p(\theta) \propto \exp(-\lambda_1 \|\theta\|_1 - \lambda_2 \|\theta\|_2^2)$$

显式特征选择: $\min_{\theta} \|y - X\theta\|_2^2 \quad \text{s.t. } \|\theta\|_0 \leq L$

MP \rightarrow DMP: ① 算与残差角度最小的分量

② 加入该分量计算梯度下降法

更新 θ

③ 更新残差

最小二乘法： $\min_{\theta} \|y - X\theta\|_2^2$

$$\text{IRLS: } \theta^{(t+1)} = \arg \min_{\theta} \sum_{n=1}^N \alpha_n(\theta^{(t)}) |y_n - x_n^\top \theta|^2$$

$$= \arg \min_{\theta} \left\| \text{diag}^{\frac{1}{2}}(\alpha(\theta^{(t)})) (y - X\theta) \right\|_2^2$$

$$= (X^\top A^{(t)} X)^{-1} X^\top A^{(t)} y$$

$$A^{(t)} = \text{diag}(\alpha(\theta^{(t)}))$$

$$\alpha_n^{(t)} = 1 \quad \alpha_n^{(t)} = |y_n - x_n^\top \theta^{(t)}|^2$$

广义线性模型：

- 1. 指数族分布： $P_x(x|\theta) = h(x) \exp(\langle \eta(\theta), T(x) \rangle - A(\theta))$
- 2. 线性预测器 $\eta = X\beta$
- 3. 连接函数 $g \quad E(Y|X) = u = g^{-1}(\eta)$

非线性模型。

NWTF估计器： $\hat{f} = \frac{\sum k_n(x-x_n) y_n}{\sum k_n(x-x_n)}$

核

- 从非参数角度：非负、对称、可积、归一化。
- 从贝叶斯角度：去参数的概率密度函数。
- 从泛函分析：再生希尔伯特核空间的再生核。



理论基础：表达定理 eg. 核岭回归： $f^* = \arg \min_{f \in \mathcal{H}_K} \sum_n (y_n - f(x_n))^2 + \lambda \|f\|_K^2$

$$\|f\|_K^2 = \sum_n \alpha_n K(x, x_n)$$

$$\alpha_n = \frac{1}{1 + \lambda K(x, x_n)^2}$$

$$\alpha = \arg \min_{\alpha \in \mathbb{R}^n} \|y - K\alpha\|_2 + \alpha K\alpha$$

选择模型的原则

$AIC: 2k - 2\log L$. 误识别率.	$BIC: k \log N - 2\log L$ 真模型.
-----------------------------	--------------------------------

数据预处理, whitening: $\hat{x} = (x - \mu) \Sigma^{-\frac{1}{2}}$ eg. 类似 $\hat{x} = \frac{x - \mu}{\sigma}$

降维 \rightarrow 原则

$\text{Minimizing reconstruction error.}$ 最小化重构误差	$\text{Maximizing mutual information.}$ KL 散度.
---	--

Isometry (近似保距)

whitening: $\hat{x} = x \Sigma^{-\frac{1}{2}} \sim \frac{1}{\sqrt{n-1}} x V \Lambda^{-\frac{1}{2}}$

迭代找主成分.

角分解法

!无均值.

特征值 & SVD 分解.

$$\tilde{x} = x U_L \quad x^T x = V \Lambda V^T$$

loss: 假设 $x = \hat{x} + e \quad \forall e_{nd} \in E \sim N(0, \sigma^2)$

$$\hat{x} = \arg \min_{x \in \Omega} \|x - \hat{x}\|_F^2 \quad \Omega = \{x \in \mathbb{R}^{N \times P} \mid r(x) \leq L\}$$

PCA

Robust: 假设 $x = \hat{x} + e \quad \forall e_{nd} \in E \sim \text{Laplace}(0, \sigma)$

$$\hat{x} = \arg \min_{x \in \Omega} \|x - \hat{x}\|_1 \quad \Omega = \{x \in \mathbb{R}^{N \times P} \mid r(x) \leq L\}$$

$$x = L + s \quad \|L\|_1 \leq L \quad \|s\|_0 \leq \text{ratio} \cdot N \cdot D.$$

NMF: $\hat{x} = \arg \min_{x \in \Omega} \|x_{noisy} - x\|_F^2$

$x \in \Omega$

$U_{N \times L} V_{P \times L}$ 乘积形式是为了约束低秩.

$$\Omega = \{x = UV^T \mid r(x) = L, U, V \geq 0\}$$

Subspace clustering: Loss, $\min_{W \in \mathbb{R}^{D \times D}} \|X - XW\|_F^2 + \gamma_1 \|W\|_1 + \gamma_2 \|W\|_*$

$W \in \mathbb{R}^{D \times D}$

稀疏 低秩

e.g. 实际上是一种自动表达，用于发现数据的子空间，收敛后的 W 经过合理排列应该为块对角。

dictionary learning \rightarrow 压缩感知: 假设 Ψ 满足 $\|\Psi\|_2 \leq 1/\alpha$, $\|\alpha\|_1 \leq S$, $X \in \mathbb{R}^N$

sensing: $y = \Phi x$, $\Phi: M \times N$, $M \ll N$
且中随机。

recovery: $\min_{\alpha} \frac{1}{2} \|y - \Phi \alpha\|_2^2 + \lambda \|\alpha\|_1$

流形学习 } 核心思想: 近似保距 可以理解为插值拟合
方法 } MDS. strain loss: $\min \left(\frac{\sum_{i,j=1}^n (d_{ij} - \bar{z}_i^T \bar{z}_j)^2}{\sum_{i,j=1}^n d_{ij}^2} \right)^{\frac{1}{2}}$
ISOMAP.: Euclidean distance \rightarrow Geodesic distance
(kNN + Dijkstra)

LLE :
 ① construct kNN
 ② $\min_{W} \|x_i - \tilde{x}^T w_i\|_2$ s.t. $\sum_{k=1}^K w_{ik} = 1$
 ③ $\tilde{w} = \text{scatter}(W)$, $\Phi = (I - \tilde{w}^T)(I - \tilde{w})$
 $\min_{Z} \text{tr}(Z^T \Phi Z) \Rightarrow Z = U_{N-L \times N-L}$

↓.

kernel PCA

$K = k(x_i, x_j)$ e.g. 注意对 K 进行双正定化

t-SNE : $P_{ij} = \frac{P_{ii} + P_{jj}}{2N}$ $P_{j|i} = \frac{\exp(-\|x_i - x_j\|^2 / (2\sigma_i^2))}{\sum_{k \neq i} \exp(-\|x_i - x_k\|^2 / (2\sigma_i^2))}$

$q_{ij} = \frac{(1 + \|z_i - z_j\|^2)^{-1}}{\sum_k \sum_{l \neq i} (1 + \|z_i - z_l\|^2)^{-1}}$

$\min_Z \text{KL}(P || Q)$

聚类

K-means

① 随机初始化中心

② ① ② ③

④ 更新聚类中心

方法

- ① 避免维度诅咒
- ② 让数据线性可分

Spectral Clustering (流形学习 + k-means 聚类)

① 构建相似度矩阵

② Laplacian Eigenmap \Rightarrow Loss: $\min \sum_{m,n=1}^N \|Z_m - Z_n\|^2 \alpha_{mn}$

$$\text{③ } L = \text{diag}(A) - A \Leftrightarrow \text{Tr}(Z^T L Z)$$

$$\text{④ } L = U \Lambda U^T \text{ 这里也可以考虑 } \text{ s.t. } \Gamma(Z) = L$$

⑤ 对 U_L 进行 k-means 聚类 中心化, $\therefore Z = U_L$.

smallest

参数化

评价标准

生成式模型: 高斯混合模型: $w = [w_k] \in \Delta^{K-1}$

$$\{N(\mu_k, \Sigma_k)\}_{k=1}^K$$

EM 算法: E-step: 第 t 步 responsibility.

$$P^{(t)}(k|x_n) = \frac{w_k^{(t)} p(x_n, \mu_k^{(t)}, \Sigma_k^{(t)})}{\sum_{i=1}^K w_i^{(t)} p(x_n, \mu_i^{(t)}, \Sigma_i^{(t)})}$$

M-step:

$$w_k^{(t+1)} = \frac{1}{n} \sum_{i=1}^n P^{(t)}(k|x_i)$$

$$\mu_k^{(t+1)} = \frac{\sum_{i=1}^n P^{(t)}(k|x_i) x_i}{\sum_{i=1}^n P^{(t)}(k|x_i)}$$

$$\Sigma_k^{(t+1)} = \frac{\sum_{i=1}^n P^{(t)}(k|x_i) (x_i - \mu_k^{(t+1)}) (x_i - \mu_k^{(t+1)})^T}{\sum_{i=1}^n P^{(t)}(k|x_i)}$$

Revisit k-means, 一种特殊的高斯混合模型
 权重、方差固定且相等
 Responsibility不被编码.

KDE (核密度估计) 用来估计数据的 PDF

$$\hat{P}_h(x) = \frac{1}{n} \sum_{i=1}^N K_h(x, x_i)$$

Mean-shift

$$m(x) = \frac{\sum_{i=1}^N k_h(x, x_i) \cdot x_i}{\sum_{i=1}^N k_h(x, x_i)}$$

eg: 当核函数为 RBF

实际上与梯度下降的方向相同.

$$\frac{\partial \hat{P}(x)}{\partial x} = \frac{1}{n} \sum_{i=1}^n \frac{\partial k_h(x, x_i)}{\partial x} \propto \frac{\sum_{i=1}^n k_h(x, x_i) (x_i - x)}{\sum_{i=1}^n k_h(x, x_i)^2}$$

eg2: 当核函数维数有限时, 可以只考虑计算邻域数据, 采加速度计算.

非参数化

不假设分布

分类

KNN : $\hat{P}(y|x) = \frac{1}{|N_h(x)|} \{d(x_n, x) \leq h\} y_n$
 y_n : one-hot vector.

朴素贝叶斯:

假设: 各特征独立且符合高斯分布.

Modeling: $P(x_d | y=k) = \frac{1}{\sqrt{2\pi} \sigma_{k,d}} e^{-\frac{(x_d - \mu_{k,d})^2}{2\sigma_{k,d}^2}}$

MLE求解 $\mu_{k,d}, \sigma_{k,d}$.

拓展: 多项式贝叶斯

伯努利贝叶斯

只是假设分布不一样, 当处理非高斯分布时,
可用 KDE

LDA:

假设: 两类分布服从正态分布且协方差相等

当不同类
差不等

判据: $y_x = 1 \Leftrightarrow \exists T. \log \frac{P(x|y=1)}{P(x|y=0)} > T$

$$\begin{aligned} \text{由假设 } \log \frac{P(x|y=1)}{P(x|y=0)} &= [(x - u_1)^T \Sigma^{-1} (x - u_1) + (x - u_0)^T \Sigma^{-1} (x - u_0)] \\ &= (u_1 - u_0)^T \Sigma^{-1} x - \frac{1}{2} (u_1 + u_0)^T \Sigma^{-1} (u_1 - u_0) \\ &> 0 \end{aligned}$$

$$\Leftrightarrow \frac{1}{2} w = \Sigma^{-1} (u_1 - u_0) \quad \langle w, x \rangle > \langle w, \frac{u_1 + u_0}{2} \rangle$$

Fisher 判据. $S = \frac{(w^T (u_1 - u_0))^2}{w^T (\Sigma_0 + \Sigma_1) w}$ 类间方差 / 类内方差

$$\Rightarrow w \propto (\Sigma_0 + \Sigma_1)^{-1} (u_1 - u_0) \text{ 知道后即可算.}$$

下推广

Multi-class LDA: } - vs. 其它.
pairwise

LR : | model: $P(y=1|X) = \frac{1}{1+e^{-X\beta}}$

$$MLL: \max_{\beta} \sum_{n=1}^N \underbrace{(y_n \log P_n + (1-y_n) \log (1-P_n))}_L$$

$$\frac{\partial L}{\partial \beta} = -X [y_n \log P_n + (1-y_n) \log (1-P_n)] P_n (1-P_n)$$

Softmax | Model: $P(y_j=1|X) = \frac{e^{(X^\top \beta_j)}}{\sum_{c=1}^C e^{(X^\top \beta_c)}}$
Loss: $\max_{\{\beta_i\}} \sum_{n=1}^N \sum_{c=1}^C y_{nc} \log P(Y_{nc}=1|X_n; \beta_c)$

SVM: | $y \in \{-1, 1\}$

$$\text{Model: } \min_{w,b} \|w\|_2$$

$$\text{s.t. } y_n (w^\top x_n - b) \geq 1 \quad \forall n=1, \dots, N$$

↓ soft-margin

$$\min_{w,b} \lambda \|w\|_2 + \frac{1}{N} \sum_{n=1}^N \xi_n$$

$$\text{s.t. } y_n (w^\top x_n - b) \geq 1 - \xi_n \quad \forall n=1, \dots, N$$

$$\xi_n \geq 0$$

$$\Rightarrow \min_{w,b} \lambda \|w\|_2 + \frac{1}{N} \sum_{n=1}^N \max \{0, 1 - y_n (w^\top x_n - b)\}$$

↓ 对偶化.

$$\max_{C_n} \sum_{n=1}^N C_n - \frac{1}{2} \sum_{n=1}^N \sum_{m=1}^N y_n C_n (x_n^\top x_m) y_m C_m$$

$$\alpha^T K \alpha$$

$$s.t. \sum_{n=1}^N c_n y_n = 0 \quad 0 \leq c_n \leq \frac{1}{2N\lambda}$$

$$w^* = \sum_{n=1}^N c_n y_n x_n$$

$$\text{核化: } K_{ij} = k(x_i, x_j)$$

信息论

Entropy: $H(X) = - \sum_{x \in X} p(x) \log p(x) = E_{X \sim P_X} [-\log p(x)]$

$$H(X, Y) = E_{X, Y \sim P_{X,Y}} [-\log p(x, y)]$$

$$X, Y \text{ 独立} \Rightarrow H(X, Y) = H(X) + H(Y)$$

Conditional Entropy 条件熵:

$$\begin{aligned} H(X|Y) &= E_{y \sim Y} [H(X|y)] = - \sum_{y \in Y} p(y) \sum_{x \in X} p(x|y) \log p(x|y) \\ &= - \sum_{x, y} p(x, y) \log p(x|y) \end{aligned}$$

$$\Rightarrow H(X|Y) = H(X, Y) - H(Y)$$

互信息 $I(X; Y) = E_{x, y \sim X, Y} [\log \frac{p(x, y)}{p(x)p(y)}]$

$$= H(X) - H(X|Y)$$

$$= H(Y) - H(Y|X)$$

$$= H(X) + H(Y) - H(X, Y)$$

$$KL\text{ 故度: } KL(P_X || Q_X) = \sum_{x \in X} P(x) \log \frac{P(x)}{Q(x)} = \sum_{x \in X} P(x) \log \frac{P(x)}{P(x) + Q(x)}$$

$$\Rightarrow I(X, Y) = KL(P(X, Y) || P(X)P(Y))$$

可利用琴生不等式证得 KL 故度非负

~~决策树~~: 每次分割都选择信息收益最大的方式.

对噪声敏感

↓
集成模型 | Bagging 投票 并行训练
Boosring 串行训练 不断做前面的错题

$$\frac{\partial \mathbf{x}^T \mathbf{a}}{\partial \mathbf{x}} = \frac{\partial \mathbf{a}^T \mathbf{x}}{\partial \mathbf{x}} = \mathbf{a}$$

$$\frac{\partial \mathbf{a}^T \mathbf{X} \mathbf{b}}{\partial \mathbf{X}} = \mathbf{a} \mathbf{b}^T$$

$$\frac{\partial \mathbf{a}^T \mathbf{X}^T \mathbf{b}}{\partial \mathbf{X}} = \mathbf{b} \mathbf{a}^T$$

$$\frac{\partial \mathbf{a}^T \mathbf{X} \mathbf{a}}{\partial \mathbf{X}} = \frac{\partial \mathbf{a}^T \mathbf{X}^T \mathbf{a}}{\partial \mathbf{X}} = \mathbf{a} \mathbf{a}^T$$

$$\frac{\partial \mathbf{x}^T \mathbf{B} \mathbf{x}}{\partial \mathbf{x}} = (\mathbf{B} + \mathbf{B}^T) \mathbf{x}$$

$$\frac{\partial \|\mathbf{y} - \mathbf{Xw}\|_2^2}{\partial \mathbf{w}} = 2 \mathbf{X}^T (\mathbf{Xw} - \mathbf{y})$$