

# CS7304H 统计学习理论与方法

2025 Fall, UNikeEN

内容来源：15、16 周期末复习课

## 0. 数学知识复习

### 向量 / 矩阵求导

- $\nabla_x(a^T x) = a$
- $\nabla_x(x^T a) = a$
- $\nabla_x(x^T x) = 2x$
- $\nabla_x(x^T Ax) = (A + A^T)x$
- 若  $A = A^T$ :  $\nabla_x(x^T Ax) = 2Ax$
- $\nabla_x \|Ax - b\|^2 = 2A^T(Ax - b)$
- $\nabla_X \text{tr}(A^T X) = A$
- $\nabla_X \text{tr}(X^T AX) = (A + A^T)X$
- 若  $A = A^T$ :  $\nabla_X \text{tr}(X^T AX) = 2AX$
- $\nabla_X \|X\|_F^2 = 2X$
- $\nabla_X \|AX - B\|_F^2 = 2A^T(AX - B)$

### 迹 (Trace) 常用恒等式

- $\text{tr}(A) = \sum_i A_{ii}$
- $\text{tr}(A) = \text{tr}(A^T)$
- $\text{tr}(AB) = \text{tr}(BA)$
- $\text{tr}(ABC) = \text{tr}(BCA)$
- $x^T x = \text{tr}(xx^T)$
- $\|AX\|_F^2 = \text{tr}(X^T A^T AX)$

### 期望、方差、协方差 (定义 + 常用公式)

#### 期望

- $E[X] = \sum_x x p(x) / \int x p(x) dx$
- $E[aX + b] = aE[X] + b$

#### 方差 (Variance)

- $\text{Var}(X) = E[(X - E[X])^2]$
- $\text{Var}(X) = E[X^2] - E[X]^2$
- $\text{Var}(aX + b) = a^2 \text{Var}(X)$

#### 协方差 (Covariance)

- $\text{Cov}(X, Y) = E[(X - E[X])(Y - E[Y])]$
- $\text{Cov}(X, X) = \text{Var}(X)$
- $\text{Cov}(aX, bY) = ab \text{Cov}(X, Y)$
- 若  $X, Y$  独立:  $\text{Cov}(X, Y) = 0$

#### 向量形式

- $\text{Cov}(X) = E[(X - E[X])(X - E[X])^T]$
- $\text{Cov}(AX) = A \text{Cov}(X) A^T$

#### 高斯分布常用

- $X \sim \mathcal{N}(\mu, \Sigma)$
- $E[X] = \mu$

- $\text{Cov}(X) = \Sigma$
- $E[(X - \mu)^T A(X - \mu)] = \text{tr}(A\Sigma)$  ( $A$  对称)

## 1. 基本概念

### 损失与风险函数

- loss 函数是给定单个（单批）样本上的误差；risk 函数是 loss 在真实数据分布上的期望（泛化误差）。
- 最小化 risk 是真正的目标，但无法直接算出来。引入 **Empirical Risk Minimization (ERM, 经验风险最小化)**，也就是训练损失

$$R_{\text{emp}}(\alpha) = \frac{1}{k} \sum_{i=1}^k Q(z_i, \alpha) \quad (1)$$

- 如果训练集是从真实分布 i.i.d 采样的，那么经验风险最小值可以近似真实风险最小值

### 维数灾难

- 高维空间，样本在空间中极度稀疏；需要样本数随维数呈指数级增长
- 高维几何现象：**伪正交性**——任意两随机高维向量的内积趋近于 0，距离/角度分布高度集中（高维空间任意两点距离几乎相等），导致“相似度”与“距离邻近性”失效。同时样本空间稀疏

NOTES: k 越大，KNN 的复杂度越低；kNN 不适合高维问题：存在维数灾难（见上）

### 回归函数

- 在平方损失下，预测误差（期望预测误差）定义为

$$EPE(f) = \mathbb{E}[Y - f(X)]^2 \quad (2)$$

- 在所有可测函数中，使  $EPE(f)$  取得最小值的函数是条件期望：

$$f(x) = \mathbb{E}(Y | X = x) \quad (3)$$

输入是  $x$ 、平方损失下最好的预测值就是此时  $Y$  的加权平均，推导：固定  $x_0$ ，预测值  $f(x_0)$ （常数），则需要最小化

$$\begin{aligned} E[(Y - f(x_0))^2 | X = x_0] &= E[Y^2 | X = x_0] - 2f(x_0) * E[Y | X = x_0] + f(x_0)^2 \\ \frac{\partial E[(Y - f(x_0))^2 | X = x_0]}{\partial f(x_0)} &= -2E[Y | X = x_0] + 2f(x_0) = 0 \\ f(x_0) &= E[Y | X = x_0] \end{aligned} \quad (4)$$

### 分类问题与贝叶斯分类器

分类可视为在给定输入  $x$  时，从有限类别集合  $\mathcal{G}$  中选择一个标签

- 损失通常是 0-1 损失，预测错误为 1。在 0-1 损失下，使风险最小的最优分类规则为：

$$\hat{G}(x) = G_k \quad \text{if} \quad P(G_k | x) = \max_{g \in \mathcal{G}} P(g | X = x) \quad (5)$$

- 含义：对每个输入  $x$ ，选择后验概率最大的类别。分类问题的理论最优解，性能上界称为 Bayes error.

$$\begin{aligned} P(G_k | x) &= \frac{P(x | G_k) P(G_k)}{P(x)} \\ &= \frac{P(x | G_k) P(G_k)}{\sum_j P(x | G_j) P(G_j)} \end{aligned} \quad (6)$$

### 误差分解

固定  $x_0$  输入，真实函数为  $f(x_0)$ 、基于训练集  $T$  的预测值为  $\hat{y} = f_t(x_0)$ ，误差分解：

$$\begin{aligned}
\text{MSE}(x_0) &= E_T[(f(x_0) - \hat{y}_0)^2] \\
&= E_T[(f(x_0) - E_T[\hat{y}_0] + E_T[\hat{y}_0] - \hat{y}_0)^2] \quad (\text{加减同一项 } E_T[\hat{y}_0]) \\
&= E_T[(E_T[\hat{y}_0] - f(x_0))^2] + E_T[(\hat{y}_0 - E_T[\hat{y}_0])^2] \\
&\quad + 2E_T[(E_T[\hat{y}_0] - f(x_0))(\hat{y}_0 - E_T[\hat{y}_0])] \\
&= (E_T[\hat{y}_0] - f(x_0))^2 + \text{Var}_T(\hat{y}_0) \\
&\quad + 2(E_T[\hat{y}_0] - f(x_0)) \underbrace{E_T[\hat{y}_0 - E_T[\hat{y}_0]]}_{=0} \quad (\text{中心化变量期望为 } 0) \\
&= \text{Var}_T(\hat{y}_0) + \text{Bias}^2(\hat{y}_0).
\end{aligned} \tag{7}$$

## 2. 线性回归

- 符号规定: 样本数  $N$ , 特征维度  $p$  (不包含截距)
  - 输入:  $X \in \mathbb{R}^{N \times (p+1)}$  (第  $i$  行为  $x_i^T$ , 向量形式), 参数:  $\beta \in \mathbb{R}^{p+1}$ , 输出:  $y \in \mathbb{R}^N$

模型:  $y = X\beta$

MSE 目标:

$$\hat{\beta} = \arg \min_{\beta} (y - X\beta)^T(y - X\beta) \tag{8}$$

**参数的形式解:**  $\boxed{\hat{\beta} = (X^T X)^{-1} X^T y}$

是为 OLS (Ordinary Least Squares, 普通最小二乘) 估计

推导:

- 目标函数对  $\beta$  求导并令其为 0:

$$\begin{aligned}
J(\beta) &= (y - X\beta)^T(y - X\beta) \\
&= y^T y - 2\beta^T X^T y + \beta^T X^T X \beta \quad (\text{交叉项均为标量可以合并}) \\
\nabla_{\beta} J(\beta) &= -2X^T y + 2X^T X \beta \quad (\text{用 } \nabla_{\beta}(\beta^T A \beta) = (A + A^T)\beta, \text{ 此处 } A = X^T X \text{ 对称}) \tag{9} \\
0 &= -2X^T y + 2X^T X \beta \\
X^T X \beta &= X^T y \\
\hat{\beta} &= (X^T X)^{-1} X^T y.
\end{aligned}$$

**参数协方差:**  $\boxed{\text{Cov}(\hat{\beta}) = \sigma^2 (X^T X)^{-1}}$

反映同一个回归问题, 反复抽不同训练集拟合, 参数估计怎么波动

噪声假设:  $E(\varepsilon) = 0$ ,  $\text{Cov}(\varepsilon) = \sigma^2 I_N$

推导:

$$\begin{aligned}
\hat{\beta} &= (X^T X)^{-1} X^T y \\
&= (X^T X)^{-1} X^T (X\beta + \varepsilon) \quad (\text{代入 } y = X\beta + \varepsilon) \\
&= (X^T X)^{-1} X^T X \beta + (X^T X)^{-1} X^T \varepsilon \\
&= \beta + (X^T X)^{-1} X^T \varepsilon.
\end{aligned} \tag{10}$$

因此

$$\begin{aligned}
\text{Cov}(\hat{\beta}) &= \text{Cov}(\beta + (X^T X)^{-1} X^T \varepsilon) \\
&= \text{Cov}((X^T X)^{-1} X^T \varepsilon) \quad (\beta \text{ 为常数项, 不影响协方差}) \\
&= A \text{Cov}(\varepsilon) A^T \quad (\text{令 } A = (X^T X)^{-1} X^T, \text{ 用 } \text{Cov}(A\varepsilon) = A\text{Cov}(\varepsilon)A^T) \\
&= (X^T X)^{-1} X^T (\sigma^2 I_N) X ((X^T X)^{-1} X^T)^T \\
&= \sigma^2 (X^T X)^{-1} X^T X (X^T X)^{-1} \quad (\text{因为 } I_N \text{ 可省, 且 } (AB)^T = B^T A^T) \\
&= \sigma^2 (X^T X)^{-1} (X^T X) (X^T X)^{-1} \\
&= \sigma^2 (X^T X)^{-1}.
\end{aligned} \tag{11}$$

**单点预测方差:**  $\boxed{\text{Var}(\hat{y}(x)) = \sigma^2 x^T (X^T X)^{-1} x}$

由于训练集随机性导致的，对均值响应  $E[Y | X = x]$  的估计不确定性。代表了 **模型稳定性**（反复抽训练集重拟合模型，同一个  $x$  处的方差），决定了该点处置信区间的宽度，而不包含观测噪声本身

对新点  $x \in \mathbb{R}^p$ （含“截距维”），预测

$$\hat{y}(x) = x^T \hat{\beta}. \quad (12)$$

$$\begin{aligned} \text{Var}(\hat{y}(x)) &= \text{Var}(x^T \hat{\beta}) \\ &= x^T \text{Cov}(\hat{\beta}) x \quad (\text{标量线性型: } \text{Var}(a^T u) = a^T \text{Cov}(u) a) \\ &= x^T (\sigma^2 (X^T X)^{-1}) x \\ &= \sigma^2 x^T (X^T X)^{-1} x. \end{aligned} \quad (13)$$

置信区间（Confidence Interval）：

$$\boxed{\hat{y}(x) \pm t_{1-\alpha/2, N-p} \sqrt{\hat{\sigma}^2 x^T (X^T X)^{-1} x}} \quad (14)$$

区分：若要预测新观测值  $Y_{\text{new}} = x^T \beta + \varepsilon_{\text{new}}$ ，则  
 $\text{Var}(Y_{\text{new}} - \hat{y}(x)) = \sigma^2 + \sigma^2 x^T (X^T X)^{-1} x$ （多出的  $\sigma^2$  是不可约噪声）

说明预测区间更大

## 噪声方差无偏估计 $\hat{\sigma}^2 = \text{RSS}/(N - p)$

上述两个公式里的  $\sigma^2$  现实不可知，需要用估计代替

先定义“帽子矩阵”与残差：

- $H = X(X^T X)^{-1} X^T$  ( $N \times N$ )
- $\hat{y} = Hy = X\hat{\beta}$
- 残差  $r = y - \hat{y} = (I - H)y$
- 残差平方和 (Residual Sum of Squares)  $\text{RSS} = r^T r = \|y - X\hat{\beta}\|^2$

### (1) 先把 RSS 写成噪声的二次型

$$\begin{aligned} r &= (I - H)y \\ &= (I - H)(X\beta + \varepsilon) \quad (\text{代入 } y = X\beta + \varepsilon) \\ &= (I - H)X\beta + (I - H)\varepsilon. \end{aligned} \quad (15)$$

又因为  $HX = X$ （代入约一下），所以  $(I - H)X = 0$ ，从而

$$\begin{aligned} r &= (I - H)\varepsilon, \\ \text{RSS} &= r^T r = \varepsilon^T (I - H)^T (I - H)\varepsilon = \varepsilon^T (I - H)\varepsilon \quad (16) \\ &\quad (\text{因 } H \text{ 对称且幂等: } H^T = H, H^2 = H, \text{ 故 } (I - H)^T (I - H) = I - H). \end{aligned}$$

### (2) 计算 $E(\text{RSS})$

当  $E(\varepsilon) = 0$ ,  $\text{Cov}(\varepsilon) = \sigma^2 I_N$ , 对任意对称矩阵  $A$  有：

$$E(\varepsilon^T A \varepsilon) = \sigma^2 \text{tr}(A). \quad (17)$$

取  $A = I - H$ :

$$\begin{aligned} E(\text{RSS}) &= E[\varepsilon^T (I - H)\varepsilon] \\ &= \sigma^2 \text{tr}(I - H) \\ &= \sigma^2 (\text{tr}(I) - \text{tr}(H)). \end{aligned} \quad (18)$$

### (3) 计算 $\text{tr}(H) = p$

因为  $H = X(X^T X)^{-1} X^T$ ，用迹的循环性质：

$$\begin{aligned} \text{tr}(H) &= \text{tr}(X(X^T X)^{-1} X^T) \\ &= \text{tr}((X^T X)^{-1} X^T X) \quad (\text{循环: } \text{tr}(ABC) = \text{tr}(BCA)) \\ &= \text{tr}(I_p) = p. \end{aligned} \quad (19)$$

所以

$$E(\text{RSS}) = \sigma^2 (N - p). \quad (20)$$

### (4) 得到无偏估计

因此令

$$\hat{\sigma}^2 = \frac{\text{RSS}}{N-p} \quad (21)$$

则

$$E(\hat{\sigma}^2) = \frac{E(\text{RSS})}{N-p} = \sigma^2, \quad (22)$$

即  $\hat{\sigma}^2$  是  $\sigma^2$  的无偏估计。

## 子集选择

保留一部分预测变量而丢弃剩余的变量（维度），得到一个可解释的、预测误差可能比全模型低的模型。

但由于是离散过程，经常表现为高方差而不会降低全模型预测误差，引入收缩方法（shrinkage methods）如下

## 岭回归（Ridge regression）

$$\hat{\beta}^{\text{ridge}} = \arg \min_{\beta} (y - X\beta)^T(y - X\beta) + \lambda \beta^T \beta, \quad \lambda \geq 0. \quad (23)$$

等价的约束形式

$$\hat{\beta}^{\text{ridge}} = \arg \min_{\beta} (y - X\beta)^T(y - X\beta) \text{ s.t. } \sum_{j=1}^p \beta_j^2 \leq s. \quad (24)$$

Recall 拉格朗日对应：对任意  $s$ ，存在某个  $\lambda$  使两者解一致（在最优点处满足 KKT 条件）。直观上： $\lambda$  是“违反约束会付出的代价”，约束越紧相当于  $\lambda$  越大。

闭式解（推导过程同之前）： $\hat{\beta}_\lambda = (X^T X + \lambda I)^{-1} X^T y$

- Q：证明  $\|\hat{\beta}_\lambda\|$  随  $\lambda$  增大不增：
- 设  $X^T X = V \Lambda V^T$ ,  $\Lambda = \text{diag}(\mu_1, \dots, \mu_p)$ 。
- 记  $z = V^T X^T y$ , 则

$$\begin{aligned} \hat{\beta}_\lambda &= (X^T X + \lambda I)^{-1} X^T y = V(\Lambda + \lambda I)^{-1} z \quad (\text{正交变换不改变范数}) \\ \|\hat{\beta}_\lambda\|^2 &= \|(\Lambda + \lambda I)^{-1} z\|^2 = \sum_{i=1}^p \frac{z_i^2}{(\mu_i + \lambda)^2} \quad (\text{逐坐标展开}). \end{aligned} \quad (25)$$

因此若  $\lambda_2 > \lambda_1$ , 则每项分母变大, 故

$$\|\hat{\beta}_{\lambda_2}\|^2 \leq \|\hat{\beta}_{\lambda_1}\|^2 \Rightarrow \|\hat{\beta}_{\lambda_2}\| \leq \|\hat{\beta}_{\lambda_1}\|. \quad (26)$$

## SVD 视角

令  $X = UDV^T$ ,  $D = \text{diag}(d_1, \dots, d_p)$ 。

普通的线性回归：

$$\begin{aligned} \hat{y}_{\text{OLS}} &= X(X^T X)^{-1} X^T y \\ &= (UDV^T) \left( (UDV^T)^T (UDV^T) \right)^{-1} (UDV^T)^T y \quad (\text{代入 } X = UDV^T) \\ &= (UDV^T) \left( VD^T U^T UDV^T \right)^{-1} (VD^T U^T) y \\ &= (UDV^T) \left( VD^2 V^T \right)^{-1} (VDU^T) y \quad (U^T U = I_p, D^T = D) \\ &= (UDV^T) \left( V(D^2)^{-1} V^T \right) (VDU^T) y \\ &= U \underbrace{DV^T V}_{I_p} (D^2)^{-1} \underbrace{V V^T}_{I_p} D U^T y \quad (\text{用 } V \text{ 正交: } V^T V = VV^T = I_p) \\ &= U D (D^2)^{-1} D U^T y \\ &= U U^T y. \end{aligned} \quad (27)$$

Ridge:

$$\begin{aligned}
\hat{y}_{\text{ridge}} &= X\hat{\beta}_\lambda = X(X^T X + \lambda I)^{-1} X^T y \\
&= U D(D^2 + \lambda I)^{-1} D U^T y \quad (\text{代入 } X = UDV^T \text{ 并化简}) \\
&= \sum_{j=1}^p u_j \frac{d_j^2}{d_j^2 + \lambda} u_j^T y \quad (\text{每个方向乘滤波因子 } \frac{d_j^2}{d_j^2 + \lambda} \in (0, 1)).
\end{aligned} \tag{28}$$

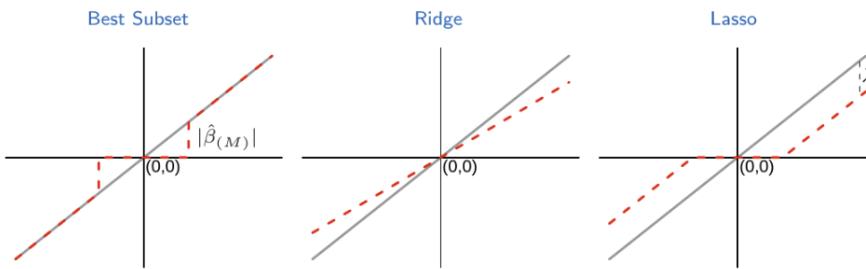
## Lasso 回归

Lasso: Least Absolute Shrinkage and Selection Operator, 使用绝对值 L1 惩罚

$$\hat{\beta}^{\text{lasso}} = \arg \min_{\beta} \|y - X\beta\|^2 + \lambda \sum_{j=1}^p |\beta_j| \iff \arg \min_{\beta} \|y - X\beta\|^2 \text{ s.t. } \sum_{j=1}^p |\beta_j| \leq s. \tag{29}$$

- Ridge(L2) 的系数通常“变小但不为 0”；而 LASSO(L1) 可产生稀疏解（很多  $\beta_j = 0$ ）
- 正交设计  $X^T X = I$  时，LASSO 解为软阈值：

$$\hat{\beta}_j = \text{sign}(x_j^T y) \max(|x_j^T y| - \lambda, 0). \tag{30}$$



图：OLS 普通解与不同正则化对应关系（正交情况）

## 3. 线性分类

贝叶斯分类器在现实中用不了，因为不知道真是的先验概率 / 给定类别数据分布；机器学习的任务是用有限样本去近似之。

- 一种方法是判别式，直接学  $P(G | X)$ : 逻辑斯蒂回归、SVM、神经网络
- 一种方法是生成式，先建立假设  $P(G | X)$  再用贝叶斯: LDA

### LDA (Linear Discriminant Analysis, 线性判别分析)

用  $f(x)$  表示  $P(X | G = k)$  (类别  $k$  中  $X$  的类别条件密度)，用  $\pi_k$  表示类别  $k$  的先验概率 (各类别和为 1)

每个类别密度用多元高斯分布建模

$$f_k(x) = \frac{1}{(2\pi)^{p/2} |\Sigma_k|^{1/2}} \exp\left(-\frac{1}{2}(x - \mu_k)^T \Sigma_k^{-1} (x - \mu_k)\right) \tag{31}$$

当满足同协方差矩阵假设时，有 线性 判别分析：

$$\Sigma_k = \Sigma, \forall k \tag{32}$$

#### 对数后验比 (两个类别的情况)

两类  $k, l$  的对数后验比：

$$\log \frac{\Pr(G = k | X = x)}{\Pr(G = l | X = x)} = \log \frac{f_k(x)}{f_l(x)} + \log \frac{\pi_k}{\pi_l} \tag{33}$$

代入高斯且  $\Sigma_k = \Sigma_l = \Sigma$  后，得到关于  $x$  的线性函数：

$$\log \frac{\Pr(G = k | X = x)}{\Pr(G = l | X = x)} = x^T \Sigma^{-1} (\mu_k - \mu_l) - \frac{1}{2} (\mu_k + \mu_l)^T \Sigma^{-1} (\mu_k - \mu_l) + \log \frac{\pi_k}{\pi_l} \tag{34}$$

因为右边对  $x$  只是一阶项，所以分界面（令该式为 0）是线性超平面。

#### 判别函数

定义 LDA 判别函数，多分类目标就是使判别函数最大：

$$\delta_k(x) = x^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + \log \pi_k \tag{35}$$

两类  $k, l$  的边界由  $\delta_k(x) = \delta_l(x)$  给出：

$$x^T \Sigma^{-1} (\mu_k - \mu_l) = \frac{1}{2} (\mu_k^T \Sigma^{-1} \mu_k - \mu_l^T \Sigma^{-1} \mu_l) - \log \frac{\pi_k}{\pi_l} \quad (36)$$

这是  $x$  的线性方程  $\Rightarrow$  分界面是直线/平面/超平面。

## 参数估计

用训练样本近似参数，不是代数约减

- 先验概率估计：

$$\hat{\pi}_k = \frac{N_k}{N} \quad (37)$$

- 类均值估计：

$$\hat{\mu}_k = \frac{1}{N_k} \sum_{i:y_i=k} x_i \quad (38)$$

- MLE 合并协方差

$$\hat{\Sigma} = \frac{1}{N} \sum_{k=1}^K \sum_{i:y_i=k} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T \quad (39)$$

**Q:** 样本数不同，分界面更靠近谁？

- 常用估计： $\pi_k \approx N_k/N$  ( $N_k$  为第  $k$  类样本数)
- 若  $N_1 > N_2$ , 则  $\pi_1 > \pi_2$ ,  $\log \pi_1$  更大, 使  $\delta_1(x)$  整体上升  
 $\Rightarrow$  为使  $\delta_1(x) = \delta_2(x)$  成立, 分界面会向样本数少的类移动  
 $\Rightarrow$  分界面更靠近样本数少的那类中心 (大类“占”更多区域)。
- 直观理解：样本多, 更常见, 判别空间更大, 推着分界线去较少的一类

## 拓展到 QDA (不做要求)

没有假设协方差相等时, 判别边界是与  $x$  有关的 二次等式

$$\delta_k(x) = -\frac{1}{2} \log |\Sigma_k| - \frac{1}{2} (x - \mu_k)^T \Sigma_k^{-1} (x - \mu_k) + \log \pi_k \quad (40)$$

## Reduced Rank LDA

背景：普通 LDA 使用的是原始  $p$  维特征空间, 需要估计  $\Sigma^{-1}$  (高维时很不稳定)。当  $p$  大但  $N$  不大时过拟合

核心：对于  $k$  个类别, 最多只有  $k - 1$  个判别方向有意义

- 类均值  $\mu_k$  的变化都落在一个至多  $K - 1$  维的仿射子空间里
- 超过这个维度, 对区分类别没有新增信息

目标：寻找一组最有判别力的正交方向：

寻找 LDA 的最优子空间序列, 过程：

- 计算  $K \times p$  的类别形心矩阵  $M$  以及共同协方差矩阵  $W$   
(组内 (within-class) 协方差矩阵)
- 使用  $W$  的特征值分解计算

$$M^* = MW^{-1/2} \quad (41)$$

$W^{-1/2}$  是白化, 先把类内协方差变成单位阵, 在这个标准化空间中类间差异才是纯粹可分性

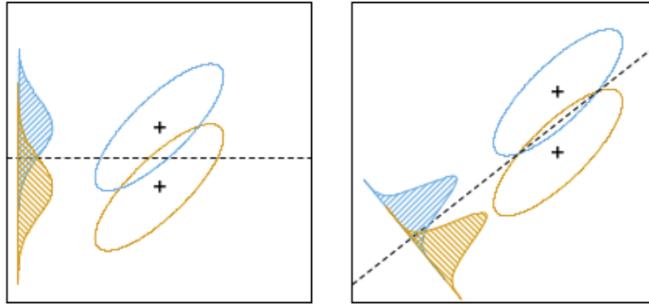
- 计算  $M^*$  的协方差矩阵  $B^*$  ( $B$  是组间 (between-class) 协方差矩阵), 并进行特征值分解：

$$B^* = V^* D_B V^{*T} \quad (42)$$

其中  $V^*$  的列向量  $v_\ell^*$  从第一个到最后一个, 依次定义了最优子空间的坐标方向。

结合上述操作, 第  $\ell$  个判别变量 (discriminant variable) 定义为：

$$Z_\ell = v_\ell^T X, \quad \text{其中 } v_\ell = W^{-1/2} v_\ell^*. \quad (43)$$



**FIGURE 4.9.** Although the line joining the centroids defines the direction of greatest centroid spread, the projected data overlap because of the covariance (left panel). The discriminant direction minimizes this overlap for Gaussian data (right panel).

图4.9. 尽管连接形心的直线定义了最大形心分散的方向，但是由于协方差（左图）投影数据会发生重叠。判别边界的线使得高斯数据的重叠最小（右图）。

## 4. 基展开 (Basis Expansion)

从“学一个函数”变成“学一堆参数的线性回归”，将函数表示为基函数的线性组合：

$$f(x) = \sum_{k=1}^N \beta_k h_k(x) = \beta^T h(x) \quad (44)$$

常见基函数：多项式基 ( $1, x, x^2 \dots$ )，分段多项式，B-Spline...

### 方差边际效应

拟合函数的方差（类比第2节的单点预测方差）：

$$\text{Var}(\hat{f}(x)) = h(x)^T (H^T H)^{-1} h(x) \sigma^2 \quad (45)$$

- 边界处基函数支持少  $\Rightarrow$  方差大  $\Rightarrow$  不稳定
- 基于事实（B-spline/spline 基关键性质）：**local support**，每个  $h_k(x)$  只在一个很小的区间内非零，对任意固定的  $x$ ，只有很少几个基函数满足  $h_k(x) \neq 0$ 。  
处于样本密集区的  $x$  有多个 spline basis 覆盖，非零变量多、贡献被多个参数方向“分摊”

### 光滑样条

- 惩罚型目标函数：

$$\text{RSS}(f, \lambda) = \sum_{i=1}^N (y_i - f(x_i))^2 + \lambda \int (f''(t))^2 dt \quad (46)$$

- $\lambda$  控制拟合-光滑折中：
  - $\lambda = 0$ : 当函数空间足够大，可以做到训练误差为 0（对数据插值），非常抖
  - $\lambda \rightarrow \infty$ : 线性回归
  - 中间：bias-variance tradeoff

### 自然三次样条（不做考试要求）

最优解为自然三次样条：

$$f(x) = \sum_{j=1}^N N_j(x) \theta_j \quad (47)$$

- $N_j(x)$ : 自然三次样条的一组基函数（也可用 B-spline 作为数值实现）
- $\theta \in \mathbb{R}^n$ : 系数（注意这里基的维数可取到  $n$ ）

求解：

- $N \in \mathbb{R}^{n \times n}$ : 样条设计矩阵

$$N_{ij} = N_j(x_i). \quad (48)$$

- $\Omega_N \in \mathbb{R}^{n \times n}$ : 惩罚矩阵（你问的 omega 是什么）

$$(\Omega_N)_{ij} = \int N_i''(t) N_j''(t) dt. \quad (49)$$

解释：因为

$$f''(t) = \sum_{j=1}^n \theta_j N_j''(t), \quad (50)$$

所以

$$\begin{aligned} \int (f''(t))^2 dt &= \int \left( \sum_i \theta_i N_i''(t) \right) \left( \sum_j \theta_j N_j''(t) \right) dt \\ &= \sum_{i,j} \theta_i \theta_j \int N_i''(t) N_j''(t) dt \\ &= \theta^T \Omega_N \theta. \end{aligned} \quad (51)$$

令  $f(x_i) = (N\theta)_i$ , 则

$$\text{RSS}(\theta, \lambda) = (y - N\theta)^T (y - N\theta) + \lambda \theta^T \Omega_N \theta. \quad (52)$$

求导可得：

$$\hat{\theta} = (N^T N + \lambda \Omega_N)^{-1} N^T y. \quad (53)$$

是为 广义岭回归

---

### 拟合值与平滑矩阵 (Smoothen Matrix) (不做考试要求)

$$\hat{f} = N\hat{\theta} = N(N^T N + \lambda \Omega_N)^{-1} N^T y \equiv S_\lambda y. \quad (54)$$

- $S_\lambda \in \mathbb{R}^{n \times n}$ : 平滑矩阵 (smoother matrix / hat matrix)
  - 重要性质： $\hat{f}$  对  $y$  线性 (线性平滑器)，所以很多统计量可写成矩阵形式
- 

### 有效自由度 (Effective Degrees of Freedom) (不做考试要求)

量化模型复杂度 (此处不是由参数个数直接决定)

$$\text{df}_\lambda = \text{trace}(S_\lambda). \quad (55)$$

解释 (直觉) :

- 在线性回归中，hat matrix 的 trace 等于参数个数 (自由度)
  - 在平滑样条中， $S_\lambda$  不是投影矩阵但仍刻画“模型对数据的响应强度”
  - $\lambda$  小  $\Rightarrow S_\lambda$  更接近插值 (更“灵活”)  $\Rightarrow$  trace 更大  $\Rightarrow$  df 更大
  - $\lambda$  大  $\Rightarrow$  更平滑 (更“刚性”)  $\Rightarrow$  trace 更小  $\Rightarrow$  df 更小
- 

### 交叉验证 (Cross-Validation, CV) 自动选 $\lambda$ (不做考试要求)

LOOCV: 令  $\hat{f}_\lambda^{(-i)}$  表示“去掉第  $i$  个样本再拟合得到的函数”，则

$$\text{CV}(\lambda) = \sum_{i=1}^n (y_i - \hat{f}_\lambda^{(-i)}(x_i))^2. \quad (56)$$

由于  $\hat{f} = S_\lambda y$ , 存在闭式：

$$\text{CV}(\lambda) = \sum_{i=1}^n \left( \frac{y_i - \hat{f}_\lambda(x_i)}{1 - S_\lambda(i, i)} \right)^2. \quad (57)$$

- $S_\lambda(i, i)$ : 平滑矩阵的第  $i$  个对角元素 (leverage)
- 作用：避免对每个  $i$  都重拟合一次 (从  $O(n)$  次拟合降到一次拟合)

实际：

- 在一组候选  $\lambda$  上计算  $\text{CV}(\lambda)$
- 取最小点：

$$\hat{\lambda} = \arg \min_{\lambda} \text{CV}(\lambda). \quad (58)$$

## 5. EM 算法, 以 2-GMM 为例

我们通常用极大似然估计 (Maximum Likelihood Estimation, MLE) 来估计参数 (对数形式) :

$$\ell(\theta) = \sum_{i=1}^n \log p(x_i | \theta) \quad (59)$$

当模型含“隐变量/缺失变量” $z$  时:

$$p(x_i | \theta) = \sum_{z_i} p(x_i, z_i | \theta) \Rightarrow \ell(\theta) = \sum_{i=1}^n \log \left( \sum_{z_i} p(x_i, z_i | \theta) \right) \quad (60)$$

由于  $\log$  与  $\sum$  纠缠, 直接最大化困难。

由此引入 EM 算法, 考试考察重点以 双高斯混合模型 (2-component GMM) 为例:

### (1) 引入隐变量 (latent variable)

为每个样本引入隐变量  $z_i \in \{0, 1\}$ :

- $z_i = 1$  表示  $x_i$  来自第 1 个高斯
- $z_i = 0$  表示  $x_i$  来自第 2 个高斯

先验:

$$P(z_i = 1) = \alpha, \quad P(z_i = 0) = 1 - \alpha. \quad (61)$$

条件分布:

$$x_i | z_i = 1 \sim \mathcal{N}(\mu_1, \Sigma_1), \quad x_i | z_i = 0 \sim \mathcal{N}(\mu_2, \Sigma_2). \quad (62)$$

写成联合分布 (常用的“指数形式”):

$$p(x_i, z_i | \theta) = \left[ \alpha \mathcal{N}(x_i | \mu_1, \Sigma_1) \right]^{z_i} \left[ (1 - \alpha) \mathcal{N}(x_i | \mu_2, \Sigma_2) \right]^{1-z_i}. \quad (63)$$

完整数据对数似然 (complete-data log-likelihood) :

$$\begin{aligned} \log p(X, Z | \theta) &= \sum_{i=1}^n \left( z_i \log \alpha + z_i \log \mathcal{N}(x_i | \mu_1, \Sigma_1) \right. \\ &\quad \left. + (1 - z_i) \log(1 - \alpha) + (1 - z_i) \log \mathcal{N}(x_i | \mu_2, \Sigma_2) \right). \end{aligned} \quad (64)$$

这一步的意义: 如果  $z_i$  已知, 就相当于“分好两堆数据”, 每堆做高斯 MLE 很容易。

### (2) EM 的核心: 用“软分配”替代不可见的 $z$

EM 每轮迭代包含两步:

- E-step: 在当前参数  $\theta^{(t)}$  下, 计算后验  $P(z_i = 1 | x_i)$  (软标签/责任度)
- M-step: 把这些后验当作权重, 最大化“期望完整对数似然”得到新参数

### (3) E-step: 计算责任度 (responsibility)

定义责任度 (软分配权重) :

$$\gamma_i := P(z_i = 1 | x_i, \theta^{(t)}). \quad (65)$$

由 Bayes 公式:

$$\gamma_i = \frac{\alpha^{(t)} \mathcal{N}(x_i | \mu_1^{(t)}, \Sigma_1^{(t)})}{\alpha^{(t)} \mathcal{N}(x_i | \mu_1^{(t)}, \Sigma_1^{(t)}) + (1 - \alpha^{(t)}) \mathcal{N}(x_i | \mu_2^{(t)}, \Sigma_2^{(t)})}. \quad (66)$$

同时:

$$P(z_i = 0 | x_i, \theta^{(t)}) = 1 - \gamma_i. \quad (67)$$

直觉:  $\gamma_i$  越接近 1,  $x_i$  越像来自第 1 个高斯; 越接近 0 越像来自第 2 个高斯。

### (4) M-step: 用加权 MLE 更新参数 (“软分好两堆”)

先定义有效样本数 (effective sample size) :

$$N_1 = \sum_{i=1}^n \gamma_i, \quad N_2 = \sum_{i=1}^n (1 - \gamma_i). \quad (68)$$

更新混合系数：

$$\alpha^{(t+1)} = \frac{N_1}{n}. \quad (69)$$

含义：第 1 个高斯“被分到的样本权重总量”占比。

更新均值：

$$\mu_1^{(t+1)} = \frac{1}{N_1} \sum_{i=1}^n \gamma_i x_i, \quad \mu_2^{(t+1)} = \frac{1}{N_2} \sum_{i=1}^n (1 - \gamma_i) x_i. \quad (70)$$

含义：把  $\gamma_i$  当成“属于第 1 类的程度”，做加权平均。

更新协方差（多维）：

$$\Sigma_1^{(t+1)} = \frac{1}{N_1} \sum_{i=1}^n \gamma_i (x_i - \mu_1^{(t+1)}) (x_i - \mu_1^{(t+1)})^T, \quad (71)$$

$$\Sigma_2^{(t+1)} = \frac{1}{N_2} \sum_{i=1}^n (1 - \gamma_i) (x_i - \mu_2^{(t+1)}) (x_i - \mu_2^{(t+1)})^T. \quad (72)$$

一维时把外积换成平方：

$$(\sigma_1^2)^{(t+1)} = \frac{1}{N_1} \sum_{i=1}^n \gamma_i (x_i - \mu_1^{(t+1)})^2, \quad (\sigma_2^2)^{(t+1)} = \frac{1}{N_2} \sum_{i=1}^n (1 - \gamma_i) (x_i - \mu_2^{(t+1)})^2. \quad (73)$$


---

## 初始化问题

EM 只保证“似然不下降”，但可能收敛到局部最优，所以初始化决定效果。

### 随机初始化

1. 随机给每个点一个初始标签  $z_i^{(0)} \in \{0, 1\}$  (或随机责任度  $\gamma_i^{(0)} \in (0, 1)$ )

2. 用“硬分配”或“软分配”的加权公式算初始参数：

$$\circ \quad \alpha^{(0)} = \frac{1}{n} \sum z_i^{(0)} \quad (\text{或 } \frac{1}{n} \sum \gamma_i^{(0)})$$

$\circ \quad \mu_k^{(0)}$ 、 $\Sigma_k^{(0)}$  用对应集合的样本均值/协方差

优点：实现快；缺点：不稳定，容易坏局部解。

### K-means 初始化（推荐，常用）

1. 对数据做 K-means ( $K=2$ )，得到两簇

2. 令

$\circ \quad \mu_k^{(0)}$  为第  $k$  簇样本均值

$\circ \quad \Sigma_k^{(0)}$  为第  $k$  簇样本协方差

$\circ \quad \alpha^{(0)} = N_k/n$

优点：稳定、收敛快；缺点：K-means 对形状/尺度敏感。

### 选取两个点做均值

- 从数据里选两个相距较远的点作为  $\mu_1^{(0)}, \mu_2^{(0)}$

- 方差用整体方差的一部分，如  $\sigma_k^{2(0)} = \text{Var}(X)$

- $\alpha^{(0)} = 0.5$

优点：快；缺点：粗糙，可能慢或陷入差解。

### 多次随机重启

- 用不同初始化跑 EM 多次

- 取最终观测对数似然  $\ell(\theta)$  最大的一次作为结果

## 6. 模型评估与选择

---

## Bias-Variance 分解 (升级版)

相比于第一章线性函数的推导，引入了真实世界的噪声。也就是真实函数变为：

$$Y = f(X) + \varepsilon, \quad E(\varepsilon) = 0, \quad \text{Var}(\varepsilon) = \sigma^2 \quad (74)$$

在输入点  $X = x_0$  处拟合值的期望预测误差：

$$\begin{aligned} \text{Err}(x_0) &= E_{T,\varepsilon}[(f(x_0) + \varepsilon - \hat{y}_0)^2] \quad (\text{代入 } Y = f(x_0) + \varepsilon, \text{ 此处 } x_0 \text{ 固定}) \\ &= E_{T,\varepsilon}[(f(x_0) - \hat{y}_0)^2] + E_{T,\varepsilon}[\varepsilon^2] + 2E_{T,\varepsilon}[\varepsilon(f(x_0) - \hat{y}_0)] \quad (\text{展开平方: } (a+b)^2 = a^2 + b^2 + 2ab) \\ &= E_T[(f(x_0) - \hat{y}_0)^2] + \underbrace{E[\varepsilon^2]}_{= \text{Var}(\varepsilon) = \sigma^2} + 2E_T[(f(x_0) - \hat{y}_0)E(\varepsilon)] \quad (\text{对 } \varepsilon \text{ 先取期望; } \hat{y}_0 \text{ 仅依赖 } T) \\ &= E_T[(f(x_0) - \hat{y}_0)^2] + \sigma^2 \quad (\text{因 } E(\varepsilon) = 0, \text{ 交叉项为 } 0) \\ &= \sigma^2 + E_T[(f(x_0) - \hat{y}_0)^2] \\ &= \sigma^2 + \text{Var}_T(\hat{y}_0) + \text{Bias}^2(\hat{y}_0) \quad (\text{第一章公式}) \end{aligned} \quad (75)$$

观测噪声这一项是 不可约误差

### KNN 的 $\text{Err}(x_0)$ 推导 (需要写出来, 可能推导? )

设定：训练集  $T = \{(x_i, y_i)\}_{i=1}^N$ ,  $y_i = f(x_i) + \varepsilon_i$ ; 对固定测试点  $x_0$ , 令它的  $k$  个最近邻输入为

$$x_{(1)}, x_{(2)}, \dots, x_{(k)}, \quad (76)$$

对应输出为  $y_{(1)}, \dots, y_{(k)}$ 。KNN 回归器：

$$\hat{f}_k(x_0) = \frac{1}{k} \sum_{\ell=1}^k y_{(\ell)}. \quad (77)$$

预测误差 (条件在  $X = x_0$ , 且训练输入视为固定) :

$$\text{Err}(x_0) = E[(Y - \hat{f}_k(x_0))^2 | X = x_0]. \quad (78)$$

注意：这里的  $Y$  指“在同一个  $x_0$  处新观测一次”的输出： $Y = f(x_0) + \varepsilon_0$ , 并且  $\varepsilon_0$  与训练噪声  $\{\varepsilon_i\}$  独立同分布。

#### Step 1: 把 $Y$ 与 $\hat{f}_k(x_0)$ 写成“真函数 + 噪声”的形式

$$\begin{aligned} Y - \hat{f}_k(x_0) &= (f(x_0) + \varepsilon_0) - \frac{1}{k} \sum_{\ell=1}^k (f(x_{(\ell)}) + \varepsilon_{(\ell)}) \\ &= \underbrace{\left( f(x_0) - \frac{1}{k} \sum_{\ell=1}^k f(x_{(\ell)}) \right)}_{\text{(A) 纯偏差项: 局部平均与 } f(x_0) \text{ 的差}} + \underbrace{\left( \varepsilon_0 - \frac{1}{k} \sum_{\ell=1}^k \varepsilon_{(\ell)} \right)}_{\text{(B) 纯噪声项}}. \end{aligned} \quad (79)$$

令

$$a := f(x_0) - \frac{1}{k} \sum_{\ell=1}^k f(x_{(\ell)}), \quad b := \varepsilon_0 - \frac{1}{k} \sum_{\ell=1}^k \varepsilon_{(\ell)}. \quad (80)$$

则  $Y - \hat{f}_k(x_0) = a + b$ , 所以

$$\text{Err}(x_0) = E[(a + b)^2] = E[a^2] + E[b^2] + 2E[ab]. \quad (81)$$

#### Step 2: 交叉项 $E[ab] = 0$

$$\begin{aligned} E[ab] &= E \left[ a \left( \varepsilon_0 - \frac{1}{k} \sum_{\ell=1}^k \varepsilon_{(\ell)} \right) \right] \\ &= a \left( E[\varepsilon_0] - \frac{1}{k} \sum_{\ell=1}^k E[\varepsilon_{(\ell)}] \right) \quad (\text{这里 } a \text{ 在“固定输入”假设下是常数}) \\ &= a(0 - 0) = 0. \end{aligned} \quad (82)$$

#### Step 3: 计算噪声项 $E[b^2]$

噪声独立、同方差：

$$\text{Var}(\varepsilon_0) = \sigma_\varepsilon^2, \quad \text{Var}\left(\frac{1}{k} \sum_{\ell=1}^k \varepsilon_{(\ell)}\right) = \frac{1}{k^2} \sum_{\ell=1}^k \text{Var}(\varepsilon_{(\ell)}) = \frac{\sigma_\varepsilon^2}{k}. \quad (83)$$

又由于  $\varepsilon_0$  与训练噪声独立:

$$\text{Var}\left(\varepsilon_0 - \frac{1}{k} \sum_{\ell=1}^k \varepsilon_{(\ell)}\right) = \text{Var}(\varepsilon_0) + \text{Var}\left(\frac{1}{k} \sum_{\ell=1}^k \varepsilon_{(\ell)}\right) = \sigma_\varepsilon^2 + \frac{\sigma_\varepsilon^2}{k}. \quad (84)$$

且  $E[b] = 0$ , 所以  $E[b^2] = \text{Var}(b)$ , 得到

$$E[b^2] = \sigma_\varepsilon^2 + \frac{\sigma_\varepsilon^2}{k}. \quad (85)$$

#### Step 4: 合并

$$\text{Err}(x_0) = \sigma_\varepsilon^2 + \left[ f(x_0) - \frac{1}{k} \sum_{\ell=1}^k f(x_{(\ell)}) \right]^2 + \frac{\sigma_\varepsilon^2}{k} \quad (86)$$

解释:

- 第一项  $\sigma_\varepsilon^2$ : 新观测点  $Y$  自带噪声 (不可约误差)
- 中间平方: kNN 的“局部平均偏差”
- 最后一项  $\sigma_\varepsilon^2/k$ : 用  $k$  个带噪声邻居做平均, 噪声方差被平均掉了

#### 线性模型的 $\text{Err}(x_0)$ 推导 (需要写出来, 可能推导? )

模型:  $y = X\beta + \varepsilon$ , 其中

- $X \in \mathbb{R}^{N \times p}$  为设计矩阵 (每行是  $x_i^T$ ), 视为固定;
- $\varepsilon \in \mathbb{R}^N$ , 满足  $E(\varepsilon) = 0$ ,  $\text{Cov}(\varepsilon) = \sigma_\varepsilon^2 I_N$ ;
- 真函数在训练点上可写为  $f = (f(x_1), \dots, f(x_N))^T$ , 且  $y = f + \varepsilon$ .

OLS 拟合与预测:

$$\hat{\beta} = (X^T X)^{-1} X^T y, \quad \hat{f}_p(x_0) = x_0^T \hat{\beta}. \quad (87)$$

#### Step 1: 把 $\hat{f}_p(x_0)$ 写成训练输出 $y$ 的线性组合 (引出 $h(x_0)$ )

$$\begin{aligned} \hat{f}_p(x_0) &= x_0^T (X^T X)^{-1} X^T y \\ &= \underbrace{(X(X^T X)^{-1} x_0)^T}_{=: h(x_0)^T} y \quad (\text{把标量写成 } h(x_0)^T y; \text{ 定义线性权重向量 } h(x_0)) \\ &= h(x_0)^T y. \end{aligned} \quad (88)$$

因此

$$h(x_0) = X(X^T X)^{-1} x_0 \in \mathbb{R}^N. \quad (89)$$

#### Step 2: 写出预测误差并拆成“偏差部分 + 噪声部分”

令新的观测  $Y = f(x_0) + \varepsilon_0$  ( $\varepsilon_0$  与训练噪声独立同分布) :

$$\begin{aligned} \text{Err}(x_0) &= E[(Y - \hat{f}_p(x_0))^2 | X = x_0] \\ &= E[(f(x_0) + \varepsilon_0 - \hat{f}_p(x_0))^2] \\ &= E[(f(x_0) - \hat{f}_p(x_0))^2] + E[\varepsilon_0^2] + 2E[\varepsilon_0(f(x_0) - \hat{f}_p(x_0))] \\ &= \sigma_\varepsilon^2 + E[(f(x_0) - \hat{f}_p(x_0))^2] \quad (\text{交叉项为 } 0; E\varepsilon_0 = 0 \text{ 且独立}). \end{aligned} \quad (90)$$

接下来对  $E[(f(x_0) - \hat{f}_p(x_0))^2]$  做 bias-var:

$$E[(f(x_0) - \hat{f}_p(x_0))^2] = (f(x_0) - E[\hat{f}_p(x_0)])^2 + \text{Var}(\hat{f}_p(x_0)). \quad (91)$$

#### Step 3: 计算 $\text{Var}(\hat{f}_p(x_0))$ (出现 $\|h(x_0)\|^2$ )

由 Step 1:  $\hat{f}_p(x_0) = h(x_0)^T y = h(x_0)^T(f + \varepsilon) = h(x_0)^T f + h(x_0)^T \varepsilon$ 。其中  $h(x_0)^T f$  为常数, 所以

$$\text{Var}(\hat{f}_p(x_0)) = \text{Var}(h(x_0)^T \varepsilon). \quad (92)$$

用协方差定义:

$$\begin{aligned} \text{Var}(h^T \varepsilon) &= E[(h^T \varepsilon)^2] \quad (\text{因 } E\varepsilon = 0) \\ &= E[h^T \varepsilon \varepsilon^T h] \\ &= h^T E[\varepsilon \varepsilon^T] h \\ &= h^T (\sigma_\varepsilon^2 I_N) h \\ &= \sigma_\varepsilon^2 h^T h = \sigma_\varepsilon^2 \|h\|^2. \end{aligned} \quad (93)$$

代回  $h = h(x_0)$ :

$$\text{Var}(\hat{f}_p(x_0)) = \sigma_\varepsilon^2 \|h(x_0)\|^2. \quad (94)$$

#### Step 4: 合并

$$Err(x_0) = \sigma_\varepsilon^2 + (f(x_0) - E[\hat{f}_p(x_0)])^2 + \|h(x_0)\|^2 \sigma_\varepsilon^2 \quad (95)$$

#### 样本内 (in-sample) 平均预测误差

把  $Err(x_i) = \sigma_\varepsilon^2 + (f(x_i) - E[\hat{f}(x_i)])^2 + \text{Var}(\hat{f}(x_i))$  在  $i = 1..N$  求平均:

$$\frac{1}{N} \sum_{i=1}^N Err(x_i) = \sigma_\varepsilon^2 + \frac{1}{N} \sum_{i=1}^N (f(x_i) - E[\hat{f}(x_i)])^2 + \frac{p}{N} \sigma_\varepsilon^2 \quad (96)$$

其中最后一项来自上面证明的  $\frac{1}{N} \sum_i \|h(x_i)\|^2 = \frac{p}{N}$ 。

#### Akaike Information Criterion (AIC)

面对一系列候选模型 (不同特征数、不同多项式阶数、不同 spline 自由度 / 正则化强度), 每个模型都能拟合训练集

- 训练误差偏乐观: 模型越复杂, 训练误差越小但泛化性不一定好。

AIC / BIC 目标: 用“训练集拟合好坏”+“对复杂度乘法”来近似测试误差 / 泛化误差, 帮助模型选择。它们只是评估得分 (选分数最小的模型), 不是用来迭代优化的

经典形式:

$$\text{AIC} = -2 \ell(\hat{\theta}) + 2d \quad (97)$$

若写成“每样本平均”:

$$\text{AIC} = -2 \ell(\hat{\theta}) + \frac{2d}{N} \quad (98)$$

- $\ell(\hat{\theta})$ : 极大似然估计代回对数似然
- $d$ : 模型自由参数个数
- 惩罚强度:  $2d$  (与  $N$  无关)

高斯噪声模型下近似:

$$\begin{aligned} \text{AIC} &= -2 \ell(\hat{\theta}) + 2d \\ &\doteq \frac{\text{RSS}(\hat{\theta})}{\sigma_\varepsilon^2} + 2d = \frac{N \overline{err}}{\sigma_\varepsilon^2} + 2d \end{aligned} \quad (99)$$

#### 推导

模型设定 (回归 + 高斯噪声) : 给定训练输入  $x_1, \dots, x_N$ , 假设  $y_i = f(x_i; \theta) + \varepsilon_i$ , 其中  $\varepsilon_i \stackrel{iid}{\sim} N(0, \sigma_\varepsilon^2)$ 。记模型在  $x_i$  的预测均值为  $\mu_i(\theta) = f(x_i; \theta)$ 。

单点条件密度 (Gaussian likelihood) :

$$p(y_i | x_i, \theta, \sigma_\varepsilon^2) = \frac{1}{\sqrt{2\pi\sigma_\varepsilon^2}} \exp\left(-\frac{(y_i - \mu_i(\theta))^2}{2\sigma_\varepsilon^2}\right) \quad (100)$$

独立样本联合密度 (**likelihood**) :

$$L(\theta) = p(y | X, \theta, \sigma_\varepsilon^2) = \prod_{i=1}^N p(y_i | x_i, \theta, \sigma_\varepsilon^2) \quad (101)$$

对数似然 (**log-likelihood**) 定义:

$$\ell(\theta) = \log L(\theta) = \sum_{i=1}^N \log p(y_i | x_i, \theta, \sigma_\varepsilon^2) \quad (102)$$

推导  $-2\ell(\theta)$  与 RSS 的关系:

$$\begin{aligned} \ell(\theta) &= \sum_{i=1}^N \log \left[ \frac{1}{\sqrt{2\pi\sigma_\varepsilon^2}} \exp \left( -\frac{(y_i - \mu_i(\theta))^2}{2\sigma_\varepsilon^2} \right) \right] \\ &= \sum_{i=1}^N \left[ \log \left( \frac{1}{\sqrt{2\pi\sigma_\varepsilon^2}} \right) + \log \left( \exp \left( -\frac{(y_i - \mu_i(\theta))^2}{2\sigma_\varepsilon^2} \right) \right) \right] \quad (\text{对数把乘积拆成加和}) \\ &= \sum_{i=1}^N \left[ -\frac{1}{2} \log(2\pi\sigma_\varepsilon^2) - \frac{(y_i - \mu_i(\theta))^2}{2\sigma_\varepsilon^2} \right] \quad (\text{用 } \log \exp(a) = a) \\ &= -\frac{N}{2} \log(2\pi\sigma_\varepsilon^2) - \frac{1}{2\sigma_\varepsilon^2} \sum_{i=1}^N (y_i - \mu_i(\theta))^2 \end{aligned} \quad (103)$$

$$\begin{aligned} -2\ell(\theta) &= N \log(2\pi\sigma_\varepsilon^2) + \frac{1}{\sigma_\varepsilon^2} \sum_{i=1}^N (y_i - \mu_i(\theta))^2 \\ &= \underbrace{N \log(2\pi\sigma_\varepsilon^2)}_{\text{与 } \theta \text{ 无关的常数项}} + \frac{\text{RSS}(\theta)}{\sigma_\varepsilon^2} \quad (\text{定义 } \text{RSS}(\theta) = \sum_{i=1}^N (y_i - \mu_i(\theta))^2) \end{aligned}$$

因此在 Gaussian 回归里 (给定  $\sigma_\varepsilon^2$ ) , 比较模型时常数项可忽略:

$$-2\ell(\hat{\theta}) \doteq \frac{\text{RSS}(\hat{\theta})}{\sigma_\varepsilon^2} \quad (\doteq \text{表示相差一个与模型无关的常数}) \quad (104)$$

定义平均训练误差 (**mean training error / MSE on training**) :

$$\overline{\text{err}} = \frac{1}{N} \text{RSS}(\hat{\theta}) \quad (105)$$

则:

$$-2\ell(\hat{\theta}) \doteq \frac{N\overline{\text{err}}}{\sigma_\varepsilon^2} \quad (106)$$

代入

## Bayesian Information Criterion (BIC)

定义 (经典形式) :

惩罚随样本量  $N$  增强

$$\text{BIC} = -2\ell(\hat{\theta}) + (\log N) d \quad (107)$$

同样代入 Gaussian 的  $-2\ell(\hat{\theta})$  (忽略常数项) :

$$\begin{aligned} \text{BIC} &= -2\ell(\hat{\theta}) + (\log N) d \\ &\doteq \frac{\text{RSS}(\hat{\theta})}{\sigma_\varepsilon^2} + (\log N) d = \frac{N\overline{\text{err}}}{\sigma_\varepsilon^2} + (\log N) d \end{aligned} \quad (108)$$

## 区别

- AIC 目标更关注“预测好不好”，更容易选择复杂模型，不那么保守
- BIC 目标更关注“找真模型”，在大样本下更能有效压制过拟合。真模型在候选集合时更能选到、但不在时可能牺牲预测性能

## 7. 支持向量机

### 基本公式

两条支持超平面：

$$w^T x + b = \pm 1 \quad (109)$$

间隔宽度（优化目标是最大化之）：

$$m = \frac{2}{\|w\|} \quad (110)$$

## 硬间隔

$$\begin{aligned} \min_{w,b} \quad & \frac{1}{2} \|w\|^2 \\ \text{s.t.} \quad & y_i(w^T x_i + b) \geq 1, \forall i \end{aligned} \quad (111)$$

转为拉格朗日函数：

$$\mathcal{L}(w, b, \alpha) = \frac{1}{2} \|w\|^2 - \sum_i \alpha_i [y_i(w^T x_i + b) - 1], \quad \alpha_i \geq 0 \quad (112)$$

$$\min_{w,b} \max_{\alpha \geq 0} \mathcal{L}(w, b, \alpha) \quad (113)$$

## 对偶问题

分别对  $w, b$  求导：

$$\frac{\partial \mathcal{L}}{\partial w} = w - \sum_i \alpha_i y_i x_i = 0 \Rightarrow w = \sum_i \alpha_i y_i x_i \quad (114)$$

$$\frac{\partial \mathcal{L}}{\partial b} = - \sum_i \alpha_i y_i = 0 \Rightarrow \sum_i \alpha_i y_i = 0 \quad (115)$$

固定  $w, b$ ；求  $\alpha$ ，将上述最优  $w$  与  $\sum_i \alpha_i y_i = 0$  代回  $\mathcal{L}$ ：

$$\begin{aligned} \mathcal{L} &= \frac{1}{2} \|w\|^2 - \sum_i \alpha_i y_i w^T x_i - \sum_i \alpha_i y_i b + \sum_i \alpha_i \\ &= \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j x_i^T x_j - \sum_{i,j} \alpha_i \alpha_j y_i y_j x_i^T x_j + \sum_i \alpha_i \quad (\sum_i \alpha_i y_i b = b \sum_i \alpha_i y_i = 0) \quad (116) \\ &= \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j x_i^T x_j \end{aligned}$$

视为对偶问题，注意始终有约束  $\alpha_i \geq 0$

## 支持向量与偏置

由 KKT 互补条件：

$$\alpha_i [y_i(w^T x_i + b) - 1] = 0 \quad (117)$$

点刚好落在间隔边界上的，定义为支持向量

$$\alpha_i > 0 \Rightarrow y_i(w^T x_i + b) = 1 \quad (118)$$

对任一支持向量，可以计算偏置

$$b = y_s - w^T x_s \quad (119)$$

## 软间隔

$$\begin{aligned} \min_{w,b,\xi} \quad & \frac{1}{2} \|w\|^2 + C \sum_i \xi_i \\ \text{s.t.} \quad & y_i(w^T x_i + b) \geq 1 - \xi_i \end{aligned} \quad (120)$$

允许少量错误点存在。 $C$  是权衡参数，越大越“硬”

核技巧：将  $x$  映射到高维空间

对偶问题只出现内积  $x_i^T x_j \Rightarrow K(x_i, x_j) = \phi(x_i)^T \phi(x_j)$ 。不显式算  $\phi$ ，只算核函数

## 8. 无监督学习

### Principal Components (PCA)

数据分布方差大、投影后重构误差极小 两种方法

板书：投影方向 数据矩阵特征值最大方向 -> 重构误差极小

问题设定：给定中心化数据

$$x_1, \dots, x_n \in \mathbb{R}^p, \quad \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad (121)$$

目标：用一个  $q$  维子空间 ( $q < p$ ) 近似数据，做数据降维。目标的两种视角：

1. 找“最有信息”的方向：把数据投影到一条线 / 一个低维子空间上，希望投影后的数据尽可能“分散”（最大化投影方差）
- 方差大：数据变化多 => 信息多
2. 用低维空间“最好地近似原数据”（最小化重构误差）

### 优化函数（重构误差）

投影后点：

$$\hat{x}_i = \mu + \lambda_i V_q \quad (122)$$

- $\mu$ : 均值
- $V_q \in \mathbb{R}^{p \times q}$ , 列正交
- $\lambda_i \in \mathbb{R}^q$

重构误差（与最小化目标）：

$$\|x_i - \hat{x}_i\|^2 \quad (123)$$

$$\{\mu, \lambda, V_q\} = \underset{\mu, \lambda, V_q}{\text{argmin}} \sum_{i=1}^n \|x_i - (\mu + \lambda_i V_q)\|^2 \quad \text{s.t. } V_q^T V_q = I_q. \quad (124)$$

消元： $V_q$ , 求最优  $\mu$  与  $\lambda_i$

对每个  $i$ , 最小化

$$\min_{\lambda_i} \|x_i - \mu - V_q \lambda_i\|^2. \quad (125)$$

$\Leftrightarrow r_i(\lambda_i) = x_i - \mu - V_q \lambda_i$ , 则

$$\|r_i\|^2 = r_i^T r_i. \quad (126)$$

对  $\lambda_i$  求导并令 0:

$$\begin{aligned} \frac{\partial}{\partial \lambda_i} \|x_i - \mu - V_q \lambda_i\|^2 &= \frac{\partial}{\partial \lambda_i} [(x_i - \mu - V_q \lambda_i)^T (x_i - \mu - V_q \lambda_i)] \\ &= -2V_q^T (x_i - \mu - V_q \lambda_i) = 0 \\ &\quad (\text{二次型求导: } \partial \|a - B\lambda\|^2 / \partial \lambda = -2B^T(a - B\lambda)) \\ \Rightarrow V_q^T V_q \lambda_i &= V_q^T (x_i - \mu). \end{aligned} \quad (127)$$

用约束  $V_q^T V_q = I_q$ :

$$\boxed{\lambda_i = V_q^T (x_i - \mu)}. \quad (128)$$

把  $\lambda_i$  代回目标函数后，目标变为

$$J(\mu, V_q) = \sum_{i=1}^N \|(I - V_q V_q^T)(x_i - \mu)\|^2. \quad (129)$$

注意  $P_\perp := I - V_q V_q^T$  是投影矩阵（对称且幂等:  $P_\perp^T = P_\perp$ ,  $P_\perp^2 = P_\perp$ ）。

对  $\mu$  求导（略去中间标准步骤）可得最优为样本均值：

$$\boxed{\mu = \bar{x}}. \quad (130)$$

（直觉：整体平移到中心使平方误差最小；形式上是凸二次函数，唯一最小点在均值。）

### 求解 $V_q$ , 新的代价函数

代回最优  $\mu$  和  $\lambda_i$ , 新的代价函数：

$$J(\bar{x}, V_q) = \sum_{i=1}^N \| (I - V_q V_q^T)(x_i - \bar{x}) \|^2, \quad V_q^T V_q = I_q. \quad (131)$$

记  $P_\perp := I - V_q V_q^T$  (投影矩阵有  $P_\perp^T = P_\perp$ ,  $P_\perp^2 = P_\perp$ )

(1) 先把范数平方写成二次型 (以下为板书)

$$\begin{aligned} J(\bar{x}, V_q) &= \sum_{i=1}^N \| P_\perp(x_i - \bar{x}) \|^2 \\ &= \sum_{i=1}^N (P_\perp(x_i - \bar{x}))^T (P_\perp(x_i - \bar{x})) \quad (\text{用 } \|a\|^2 = a^T a) \\ &= \sum_{i=1}^N (x_i - \bar{x})^T P_\perp^T P_\perp (x_i - \bar{x}) \quad (\text{把 } P_\perp \text{ 移到中间}) \\ &= \sum_{i=1}^N (x_i - \bar{x})^T P_\perp P_\perp (x_i - \bar{x}) \quad (\text{因 } P_\perp^T = P_\perp) \\ &= \sum_{i=1}^N (x_i - \bar{x})^T P_\perp (x_i - \bar{x}) \quad (\text{因 } P_\perp^2 = P_\perp). \end{aligned} \quad (132)$$

(2) 变成 trace

对每一项 (它是标量) :

$$\begin{aligned} (x_i - \bar{x})^T P_\perp (x_i - \bar{x}) &= \text{tr}((x_i - \bar{x})^T P_\perp (x_i - \bar{x})) \quad (\text{标量 } a = \text{tr}(a)) \\ &= \text{tr}(P_\perp (x_i - \bar{x})(x_i - \bar{x})^T) \quad (\text{trace 循环: } \text{tr}(ABC) = \text{tr}(BCA)). \end{aligned} \quad (133)$$

因此

$$\begin{aligned} J(\bar{x}, V_q) &= \sum_{i=1}^N \text{tr}(P_\perp (x_i - \bar{x})(x_i - \bar{x})^T) \\ &= \text{tr}\left(P_\perp \sum_{i=1}^N (x_i - \bar{x})(x_i - \bar{x})^T\right) \quad (\text{trace 线性: } \sum \text{tr}(\cdot) = \text{tr}(\sum \cdot)). \end{aligned} \quad (134)$$

(3) 引入样本协方差

定义样本协方差矩阵 (中心化形式) :

$$\hat{\Sigma} := \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})(x_i - \bar{x})^T. \quad (135)$$

于是

$$\sum_{i=1}^N (x_i - \bar{x})(x_i - \bar{x})^T = N \hat{\Sigma}. \quad (136)$$

代回上式得到板书那行:

$$J(\bar{x}, V_q) = \text{tr}(P_\perp N \hat{\Sigma}) = N \text{tr}((I - V_q V_q^T) \hat{\Sigma}). \quad (137)$$

因为  $N$  是常数, 等价优化为:

$$\min_{V_q^T V_q = I_q} \text{tr}((I - V_q V_q^T) \hat{\Sigma}). \quad (138)$$

(4) 最小化重构误差"等价于"最大化投影方差"

展开 trace:

$$\begin{aligned} \text{tr}((I - V_q V_q^T) \hat{\Sigma}) &= \text{tr}(\hat{\Sigma}) - \text{tr}(V_q V_q^T \hat{\Sigma}) \quad (\text{线性: } \text{tr}(A - B) = \text{tr}(A) - \text{tr}(B)) \\ &= \text{tr}(\hat{\Sigma}) - \text{tr}(V_q^T \hat{\Sigma} V_q) \quad (\text{循环: } \text{tr}(V_q V_q^T \hat{\Sigma}) = \text{tr}(V_q^T \hat{\Sigma} V_q)). \end{aligned} \quad (139)$$

$\text{tr}(\hat{\Sigma})$  与  $V_q$  无关 (常数), 所以:

$$\min_{V_q^T V_q = I_q} \text{tr}((I - V_q V_q^T) \hat{\Sigma}) \iff \max_{V_q^T V_q = I_q} \text{tr}(V_q^T \hat{\Sigma} V_q). \quad (140)$$

其中  $\text{tr}(V_q^T \hat{\Sigma} V_q)$  就是"把数据投影到  $q$  维子空间后, 总方差 (方差之和) 最大"。

### (5) 最终解 (特征向量)

若  $\hat{\Sigma}$  的特征分解为  $\hat{\Sigma} = Q\Lambda Q^T$  (特征值降序) ,  
则最优  $V_q$  取最大的  $q$  个特征值对应的特征向量:

$$V_q = [q_1, \dots, q_q]. \quad (141)$$

### 奇异值分解视角

#### 中心化数据矩阵

$$X = \begin{bmatrix} (x_1 - \bar{x})^T \\ \vdots \\ (x_N - \bar{x})^T \end{bmatrix} \in \mathbb{R}^{N \times p}, \quad \bar{x} = \frac{1}{N} \sum_{i=1}^N x_i \quad (142)$$

#### SVD (奇异值分解)

$$X = UDV^T \quad (143)$$

- $U \in \mathbb{R}^{N \times p}$ ,  $U^T U = I_p$  (列正交)
- $D = \text{diag}(d_1, \dots, d_p)$ ,  $d_1 \geq \dots \geq d_p \geq 0$
- $V \in \mathbb{R}^{p \times p}$ ,  $V^T V = VV^T = I_p$  (正交)

为什么  $V_q$  是  $V$  的前  $q$  列 (主方向)

$$\hat{\Sigma} = \frac{1}{N} X^T X = \frac{1}{N} (UDV^T)^T (UDV^T) = \frac{1}{N} V D^2 V^T \quad (\text{用 } U^T U = I_p) \quad (144)$$

所以  $\hat{\Sigma}$  的特征向量就是  $V$  的列向量  $v_j$ , 特征值  $\lambda_j = d_j^2/N$ ; 取最大的  $q$  个方向:

$$V_q = [v_1, \dots, v_q]. \quad (145)$$

#### 投影矩阵

$$H_q = V_q V_q^T, \quad \hat{x} = H_q x \quad (\text{投到 } \text{span}(V_q) \text{ 上}) \quad (146)$$

$X = SA^T$  是换符号

常见取  $S := UD$ ,  $A := V$ , 则

$$X = UDV^T = (UD)V^T = SA^T. \quad (147)$$

$A (= V)$  给主方向;  $S (= UD)$  给样本在主方向上的坐标 (scores) 。

### Independent Component Analysis (ICA)

- 混合模型:  $x(k) = A s(k) + \varepsilon(k)$ 
  - $s(k)$ : 独立源信号;  $x(k)$ : 观测;  $A$ : 混合矩阵;  $\varepsilon(k)$ : 噪声
- 目标: 找  $W$  使  $y = Wx$  的分量尽可能独立 (不仅是不相关)
- 一个代表性目标 (非高斯性/独立性) :  $\max_W \sum_i \text{Negentropy}(w_i^T x)$
- 场景: EEG/MEG 盲源分离、语音鸡尾酒会问题

### 梯度下降和流形的区别

- 梯度下降 (欧氏空间无约束) :
$$\theta^{(t+1)} = \theta^{(t)} - \eta \nabla_{\theta} L(\theta)$$
- 流形优化 (参数受约束在  $\mathcal{M}$  上) :
$$\theta^{(t+1)} = \text{Proj}_{\mathcal{M}}(\theta^{(t)} - \eta \nabla L)$$
- 直觉: 流形=“走一步再拉回合法集合” (如正交约束  $V^T V = I$ )

### Stochastic Neighbor Embedding 和 t-SNE

t-SNE 主要用于可视化, 而不是“作为通用降维方法去做下游建模”

它优化的是“邻居关系”而不是保全局几何, 强烈强调: 高维里“近”的点在低维里也要“近”。结果是: 局部团簇常很清晰, 但团簇之间的距离、相对位置、大小往往不可信 (全局结构容易被扭曲)

## 9. Latent Variable Model

- 观测数据:  $X$

- 隐变量:  $Z$
- 模型参数:  $\theta$

目标: 最大化观测数据对数似然

$$\ell(\theta; X) = \log p(X | \theta) \quad (148)$$

但由于存在隐变量:

$$p(X | \theta) = \int p(X, Z | \theta) dZ \quad (149)$$

积分通常不可解析。

## ELBO 推导

$$\begin{aligned} \log p(X | \theta) &= \log \int p(X, Z | \theta) dZ \\ &= \log \int \frac{p(X, Z | \theta)}{q(Z | \phi)} q(Z | \phi) dZ \quad (\text{乘除同一个 } q(Z | \phi)) \\ &= \log E_{q(Z|\phi)} \left[ \frac{p(X, Z | \theta)}{q(Z | \phi)} \right] \quad (\text{期望定义}) \\ &\geq E_{q(Z|\phi)} \left[ \log \frac{p(X, Z | \theta)}{q(Z | \phi)} \right] \quad (\text{Jensen 不等式; log 凹}) \\ &= E_{q(Z|\phi)} [\log p(X, Z | \theta)] - E_{q(Z|\phi)} [\log q(Z | \phi)]. \end{aligned} \quad (150)$$

定义 ELBO:

$$\boxed{\text{ELBO}(\theta, \phi) = E_{q(Z|\phi)} [\log p(X, Z | \theta)] - E_{q(Z|\phi)} [\log q(Z | \phi)]} \quad (151)$$

## ELBO 与 KL 的严格关系

由 Bayes:

$$p(Z | X, \theta) = \frac{p(X, Z | \theta)}{p(X | \theta)}. \quad (152)$$

把  $p(X, Z | \theta) = p(Z | X, \theta) p(X | \theta)$  代回 ELBO:

$$\begin{aligned} \text{ELBO}(\theta, \phi) &= E_{q(Z|\phi)} \left[ \log \frac{p(X, Z | \theta)}{q(Z | \phi)} \right] \\ &= E_{q(Z|\phi)} \left[ \log \frac{p(Z | X, \theta) p(X | \theta)}{q(Z | \phi)} \right] \\ &= E_{q(Z|\phi)} [\log p(X | \theta)] + E_{q(Z|\phi)} \left[ \log \frac{p(Z | X, \theta)}{q(Z | \phi)} \right] \quad (153) \\ &= \log p(X | \theta) - \underbrace{E_{q(Z|\phi)} \left[ \log \frac{q(Z | \phi)}{p(Z | X, \theta)} \right]}_{= \text{KL}(q(Z|\phi) \| p(Z|X,\theta))}. \end{aligned}$$

因此:

$$\boxed{\log p(X | \theta) = \text{ELBO}(\theta, \phi) + \text{KL}(q(Z | \phi) \| p(Z | X, \theta))} \quad (154)$$

## General EM 算法

### E-step (或 Variational E-step)

固定  $\theta^{(k)}$ , 选择  $\phi^{(k)}$  使 KL 最小:

$$\phi^{(k)} = \arg \min_{\phi} \text{KL}\left(q(Z | \phi) \| p(Z | X, \theta^{(k)})\right). \quad (155)$$

- 若  $q$  不受限制 (可表达任意分布), 最优解是:

$$\boxed{q(Z | \phi^{(k)}) = p(Z | X, \theta^{(k)})} \quad (156)$$

- 若  $q$  受限 (比如均值场), 就只能近似最优后验 (这就是 VI/变分推断)。

### M-step

固定  $\phi^{(k)}$  (即固定  $q(Z | \phi^{(k)})$ ), 更新  $\theta$ :

$$\theta^{(k+1)} = \arg \max_{\theta} E_{q(Z|\phi^{(k)})} [\log p(X, Z | \theta)] \quad (157)$$

## VAE

- VAE 是一种 带概率建模的自编码器，用隐变量模型学习数据的生成分布
- 编码器输出的不是确定向量，而是 隐变量分布参数（均值  $\mu$ 、方差  $\sigma^2$ ）
- 通过 重参数化技巧 从分布中采样，训练时最大化 ELBO

项目	AE	VAE
隐空间	确定向量	随机变量（分布）
编码器输出	$z$	$(\mu, \sigma)$
目标函数	重构误差	重构误差 + KL 正则
生成能力	弱	强（可采样）
理论基础	表征学习	概率生成模型

## 10. 深度学习

基础知识，见《智能计算系统》课程笔记

### 注意力

给定一组“要查的东西” Key/Value (K,V) 和一个“查询” Query (Q)，注意力做三步：

1. 相似度打分：用 Q 和每个 K 算“像不像”；
2. 归一化：把分数变成权重 (softmax)；
3. 加权汇总：用权重对 V 做加权平均，得到输出 (context)。

#### 点积注意力 (Dot-Product Attention)

矩阵形式（多查询同时算）： $Q \in \mathbb{R}^{n_q \times d_k}, K \in \mathbb{R}^{n_k \times d_k}, V \in \mathbb{R}^{n_k \times d_v}$ ：

$$\text{Attention}(Q, K, V) = \text{softmax}(QK^\top) V \quad (158)$$

#### 缩放点积注意力 (Scaled Dot-Product Attention, Transformer 常用)

为避免  $d_k$  大时点积数值过大，先除以  $\sqrt{d_k}$ （对应分子的方差）：

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right) V \quad (159)$$

### Diffusion

目标：学一个生成模型，最后能“从纯噪声采样出像数据一样的样本”

先加噪，逐步往数据里掺入高斯噪声，使分布最终变得简单（接近  $N(0, I)$ ）。前向过程  $q$  人为规定，需要学习的是反向链（从噪声还原到数据）：

$$p(z_{t-1} | z_t) \quad (t = T, \dots, 1) \quad (160)$$

但真实的反向条件分布未知，所以用神经网络近似，输入  $t$  步样本预测去除的噪声：

$$p_\theta(z_{t-1} | z_t) \quad (161)$$

#### 两种定义

定义 A (递推视角)：

$$z_t = \sqrt{1 - \beta_t} z_{t-1} + \sqrt{\beta_t} \epsilon_t, \quad \epsilon_t \sim N(0, I) \quad (162)$$

定义 B (概率视角)：

$$q(z_1 | x) = N(\sqrt{1 - \beta_1} x, \beta_1 I), \quad q(z_t | z_{t-1}) = N(\sqrt{1 - \beta_t} z_{t-1}, \beta_t I) \quad (163)$$

#### 推导直接采样公式

$$z_t = \sqrt{\alpha_t} x + \sqrt{1 - \alpha_t} \epsilon$$

先手动展开前两步（和 PPT 一样），把  $z_2$  写成  $x$  和独立高斯噪声的线性组合：

- 第一步:

$$z_1 = \sqrt{1 - \beta_1} x + \sqrt{\beta_1} \epsilon_1 \quad (164)$$

- 第二步:

$$\begin{aligned} z_2 &= \sqrt{1 - \beta_2} z_1 + \sqrt{\beta_2} \epsilon_2 \\ &= \sqrt{1 - \beta_2} (\sqrt{1 - \beta_1} x + \sqrt{\beta_1} \epsilon_1) + \sqrt{\beta_2} \epsilon_2 \\ &= \sqrt{(1 - \beta_2)(1 - \beta_1)} x + \underbrace{\left( \sqrt{(1 - \beta_2)\beta_1} \epsilon_1 + \sqrt{\beta_2} \epsilon_2 \right)}_{\triangleq \tilde{\epsilon}_2} \end{aligned} \quad (165)$$

关键:  $\tilde{\epsilon}_2$  仍是高斯 (独立高斯线性组合还是高斯), 且均值为 0。它的协方差:

$$\begin{aligned} \text{Cov}(\tilde{\epsilon}_2) &= (1 - \beta_2)\beta_1 I + \beta_2 I \quad (\epsilon_1, \epsilon_2 \text{ 独立, 交叉项为 } 0) \\ &= (\beta_1 + \beta_2 - \beta_1\beta_2)I = (1 - (1 - \beta_1)(1 - \beta_2))I \end{aligned} \quad (166)$$

所以可写成标准形式  $\tilde{\epsilon}_2 = \sqrt{1 - (1 - \beta_1)(1 - \beta_2)} \epsilon$ , 其中  $\epsilon \sim N(0, I)$ :

$$z_2 = \sqrt{(1 - \beta_1)(1 - \beta_2)} x + \sqrt{1 - (1 - \beta_1)(1 - \beta_2)} \epsilon \quad (167)$$

现在推广到一般  $t$ : 定义

$$\alpha_t = \prod_{k=1}^t (1 - \beta_k) \quad (168)$$

则反复展开 (同样的结构不断出现) 可得到闭式:

$$z_t = \sqrt{\alpha_t} x + \sqrt{1 - \alpha_t} \epsilon, \quad \epsilon \sim N(0, I) \quad (169)$$

推广到  $t$  的高斯视角归纳:

假设第  $t - 1$  步已经是一个条件高斯, 我们要证明对任意  $t$  都有

$$q(z_t | x) = N(\sqrt{\alpha_t} x, (1 - \alpha_t)I), \quad \alpha_t = \prod_{k=1}^t (1 - \beta_k) \quad (170)$$

归纳假设 (IH) : 对某个  $t - 1$  成立:

$$z_{t-1} | x \sim N(\sqrt{\alpha_{t-1}} x, (1 - \alpha_{t-1})I) \quad (171)$$

从采样式:

$$z_t = \sqrt{1 - \beta_t} z_{t-1} + \sqrt{\beta_t} \epsilon_t \quad (172)$$

在给定  $x$  时:

- $z_{t-1} | x$  是高斯 (由 IH)
- $\epsilon_t$  是高斯, 且与  $z_{t-1}$  独立

先看均值:

$$\begin{aligned} E[z_t | x] &= E[\sqrt{1 - \beta_t} z_{t-1} + \sqrt{\beta_t} \epsilon_t | x] \\ &= \sqrt{1 - \beta_t} E[z_{t-1} | x] + \sqrt{\beta_t} E[\epsilon_t] \\ &= \sqrt{1 - \beta_t} \sqrt{\alpha_{t-1}} x \end{aligned} \quad (173)$$

令  $\alpha_t = (1 - \beta_t)\alpha_{t-1}$ , 则

$$E[z_t | x] = \sqrt{\alpha_t} x \quad (174)$$

再看协方差:

因为  $\epsilon_t$  与  $z_{t-1}$  独立, 且常数缩放下协方差按平方缩放:

$$\text{Cov}(aU) = a^2 \text{Cov}(U) \quad (175)$$

所以

$$\begin{aligned}
\text{Cov}(z_t \mid x) &= \text{Cov}\left(\sqrt{1-\beta_t} z_{t-1} + \sqrt{\beta_t} \epsilon_t \mid x\right) \\
&= (1-\beta_t) \text{Cov}(z_{t-1} \mid x) + \beta_t \text{Cov}(\epsilon_t) \\
&= (1-\beta_t)(1-\alpha_{t-1})I + \beta_t I \\
&= ((1-\beta_t) - (1-\beta_t)\alpha_{t-1} + \beta_t)I \\
&= (1-(1-\beta_t)\alpha_{t-1})I \\
&= (1-\alpha_t)I
\end{aligned} \tag{176}$$

因此归纳完成：

$$z_t \mid x \sim N(\sqrt{\alpha_t} x, (1-\alpha_t)I) \tag{177}$$

并且  $\alpha_t = \prod_{k=1}^t (1-\beta_k)$  (由递推展开)。

既然  $z_t \mid x$  是上述高斯，那么一定存在  $\epsilon \sim N(0, I)$  使得

$$z_t = \sqrt{\alpha_t} x + \sqrt{1-\alpha_t} \epsilon \tag{178}$$