第一部分：引言与维数灾难 (Pages 1-13)

1. 降维的背景：

- **初衷：** 在实际应用中，特征数往往达到数百甚至上千。直观上，增加特征能提供更多信息，从而提高分类性能。
- **现实：** 当特征数超过一定点后，性能反而下降。

2. 维数灾难 (Curse of Dimensionality)：

- **现象：** 随着维度增加，分类精度先升后降（在训练样本有限的情况下）。
- **核心挑战：**
    - **分类精度：** 样本密度随维度呈指数级稀释。
    - **计算复杂度：** 模型参数剧增，计算开销变大。
- **原因：** 模型选择假设错误、高维观测下的估计误差（过拟合）。
- **经验法则：** 训练样本数 n 与特征数 d 的比例应满足 $n/d > 10$。

3. 特征降维 (Feature Reduction)：

- **方式：** 特征选择（子集）或特征组合（变换）。
- **优势：** 降低估计复杂度、方便可视化、减少冗余。
- **分类：** 线性变换（PCA, LDA） vs. 非线性变换（Isomap, LLE）。

第二部分：主成分分析 PCA (Pages 14-40)

1. 核心目标：
PCA 寻找一个最优的线性投影，将 d 维数据映射到 d' 维（$d' < d$），目标是：

- **最大化投影方差：** 保持数据的多样性。
- **最小化重构误差：** 在最小二乘意义下，让降维后的数据最能代表原始数据。

2. 数学表述：

- **样本方差矩阵：** $S = \frac{1}{N}\sum_{n=1}^{N}(x_n - \bar{x})(x_n - \bar{x})^T$。
- **特征方程：** $Su_i = \lambda_i u_i$。
- **选择：** 选取前 M 个最大的特征值对应的特征向量（主成分轴）。

3. PCA 的变体：

- **概率 PCA (Probabilistic PCA):**
  - 将 PCA 视为一个生成式潜变量模型：$x = Wz + \mu + \epsilon$。
  - 假设潜变量 z 和噪声 ϵ 均服从高斯分布。
  - 优点：可以处理缺失数据，能用于生成新样本。
- **核 PCA (Kernel PCA):**
  - 通过非线性映射 $\phi(x)$ 将数据映射到高维空间，再在该空间执行线性 PCA。
  - 利用**核技巧 (Kernel Trick)**，无需显式计算高维映射，直接在低维计算内积。

第三部分：自编码器 Autoencoder (Pages 41-43)

- **定义：** 一种旨在将输入映射到自身（输入=输出）的神经网络。
- **结构：** 输入层 $\to$ 隐藏层（瓶颈层） $\to$ 输出层。
- **结论：**
  - 如果隐藏层是线性的，自编码器等价于 **PCA**。
  - 加入非线性激活函数和更多隐藏层，可以实现**非线性降维**。

第四部分：线性判别分析 LDA (Pages 44-61)

1. PCA vs. LDA：

- **PCA：** 无监督，追求“表示/重构”，寻找方差最大的方向。
- **LDA：** **有监督**，追求“分类/判别”，寻找最能区分不同类别的方向。

2. Fisher 线性判别准则 (Fisher Criterion):

- **核心思想：** **类内方差最小，类间方差最大**（让同类样本抱团，异类样本远离）。
- **代价函数：** $J(w) = \frac{w^T S_B w}{w^T S_W w}$。
  - S_B (Between-class scatter): 类间散度矩阵。
  - S_W (Within-class scatter): 类内散度矩阵。
- **局限性：** 投影后的维度最多为 $c-1$（c 为类别数），因为 S_B 的秩最大为 $c-1$。

第五部分：流形学习 Manifold Learning (Pages 62-90)

流形学习旨在发现嵌入在高维空间中的低维非线性结构（如"瑞士卷"数据）。

1. Isomap (等距映射)：

- **核心：** 保持样本间的**测地线距离 (Geodesic distance)**。
- **步骤：**
    1. 构建近邻图。
    2. 计算图中所有点对的最短路径（近似测地距离）。
    3. 应用多维尺度变换 (MDS) 进行嵌入。
- **优点：** 能够捕捉全局非线性结构。

2. LLE (局部线性嵌入)：

- **核心：** 保持样本间的**局部线性重构关系**。
- **原理：** 假设局部邻域是平坦的，每个点可以用邻居的线性组合表示，降维后保持这些权重不变。
- **优点：** 计算效率高（处理稀疏矩阵特征值问题），擅长捕捉局部几何特征。

3. SNE 与 t-SNE：

- **SNE：** 将高维空间中点的相似度转化为概率，通过最小化 KL 散度在低维空间重建。
- **t-SNE：**
    - 改进了 SNE，使用 **Student t-分布** 代替高斯分布。
    - 解决了"拥挤问题" (Crowding problem)，是目前**数据可视化**（如 MNIST）效果最好的算法之一。

第六部分：度量学习 Metric Learning (Pages 91-100)

1. 动机：
降维的本质是寻找合适的特征空间，而空间定义了距离度量。既然如此，直接"学习"一个合适的距离度量矩阵 M 可能更有效。

2. 马氏距离 (Mahalanobis Distance)：

- 公式：$dist_{mah}(x_i, x_j) = \sqrt{(x_i - x_j)^T M (x_i - x_j)}$。
- 要求：矩阵 M 必须是**半正定 (PSD)** 的。

3. 常见度量学习方法：

- **NCA (近邻成分分析):** 目标是直接优化最近邻分类器的留一法 (LOO) 分类准确率。
- **基于约束的学习 (Xing et al.):**
    - **Must-link (必连):** 相似样本间的距离要小。
    - **Cannot-link (勿连):** 不相似样本间的距离要大（通常设置一个下限阈值）。

第七部分：降维方法总结对照表 (Pages 101-103)

方法	属性/类别	核心特点
PCA	线性, 无监督	快速, 基于特征向量, 最小重构误差
LDA	线性, 有监督	判别性强, 基于类间/类内散度比
Kernel PCA	非线性, 无监督	利用核技巧处理非线性分布
Autoencoder	线性/非线性	神经网络结构, 灵活, 线形时等价于PCA
Isomap	非线性, 流形	保持全局测地距离
LLE	非线性, 流形	保持局部重构权重
t-SNE	非线性, 概率	极佳的可视化效果, 解决拥挤问题

复习建议：

1. **推导能力：** 重点掌握 PCA 的特征值分解推导和 LDA 的 Fisher 准则函数推导。
2. **对比理解：** 能够清晰区分 PCA（投影后方差大）和 LDA（投影后类别分得开）。
3. **非线性选择：** 记住 Isomap 关注“路径/距离”，LLE 关注“重构/邻居”。
4. **维数灾难：** 记住样本数与维度的 $10:1$ 比例。

祝你复习顺利！如果对某个具体公式有疑问，可以随时问我。