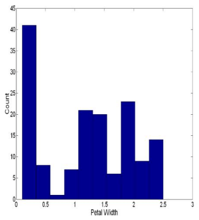
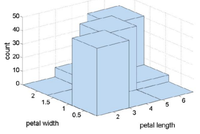
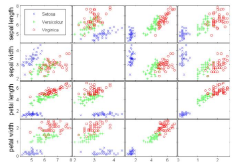
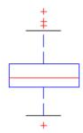
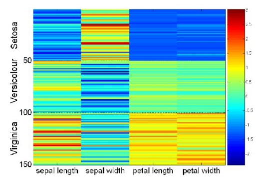
Topic2 Exploring Data

1. 数据是对象 (objects) 及其属性 (attributes) 的集合;变量类型：连续变量 (Continuous)；名义/类别变量 (Nominal/Categorical)；有序变量 (Ordinal)；区间变量 (Interval)；2.数据类型 (Types of Data)：数据矩阵 (Data matrix)、文本数据、图音视频数据、交易数据 (Transactions data) 以及图数据 (Graph data)；数据矩阵：n*p 矩阵；n 行代表对象，p 列代表变量；文本数据 (Text Data)当中通过 "文档-词项(term)" 矩阵展示词频统计；交易数据一种特殊的记录数据，每条记录（交易）包含一组项目 (items)；数据质量 (Data Quality)：包括噪声与异常值 (Noise and outliers)、缺失值 (Missing values) 以及采样偏差 (Sampling bias)；噪声是原始值的扰动；异常值是与其他对象显著不同的；缺失值成因：信息未收集和变量对案例不适用比如儿童没有年收入；采样偏差：随机采样与总体不匹配；原因：便利抽样 cov sam；类别不平衡；数据探索 data exploration：对数据进行初步探索以更好地理解其特征；EDA 探索性数据分析 John tukey 创立；侧重于可视化：描述性统计量：比如频率 fre 众数 mode 位置 location 离散度 spread 偏度 skewness；

2. 可视化的方法（图的类型）：直方图（histogram）/2D Histogram 箱线图 (Boxplot):包括中位数/上四分位下四分位/outliers/数据扩散范围两边；散点图 (Scatter Plot)不同变量两两之间的关系；矩阵图 (Matrix Plot)：当对象按类别排序时非常有用；平行坐标图 (Parallel Coordinates Plot):使用一组平行轴，每个对象的变量值在对应轴上描点并连线，每个对象由一条线表示；

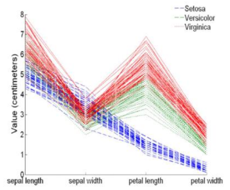


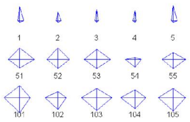
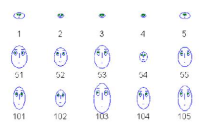


其他可视化技术 (Other Visualization Techniques):

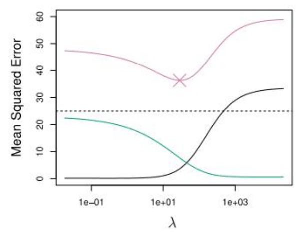
STAR PLOT AND CHERNOFF FACE

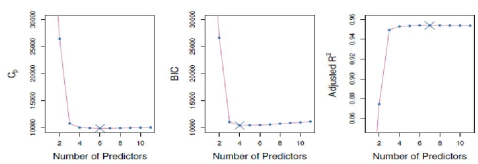


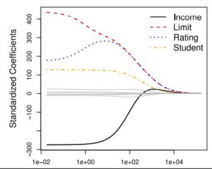



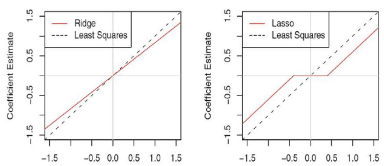
Topic5 Model Selection and Regularization

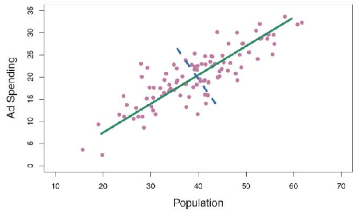
稀疏回归 (Sparse Regression)：**给定训练集 (x_i, y_i)，假设线性模型 $y_i = \beta_0 + \sum_{j=1}^{p_0} \beta_j x_{ij} + \epsilon_i$** 稀疏性假设：假设 $p_0 \ll p$ 即真实相关的预测变量远小于总变量数；变量选择的目标：正确检测出包含信息量的预测变量集合；**区分出冗余变量**；**主要关注线性回归模型**；为什么关注变量选择？多重共线性(Multicollinearity)显著性掩盖；**方差膨胀**；总变量太大过拟合导致维度灾难；可解释性 (Interpretability) 无关变量模型更复杂；常用的技术：最优子集选择 **b subset selc**；信息准则交叉验证；逐步变量选择 forward/backward；收缩方法 (Shrinkage methods)：如 Lasso 及其变体；PCA 降维；**最优子集选择**：遍历拟合所有可能的预测变量组合；筛选出每个变量数量下的最佳模型(最小 rss/最大 r2)；最后利用 cv/adj r2/等确定唯一全局最优模型；模型选择准则 (Model Selection Criteria)：Mallow`s C_p 小/AIC/BIC 小/Adjust R^2 大；前向/后向选择：空模型添加变量/全模型逐个剔除变量来构建候选模型最后通过统计准则 CV/BIC 选出最优模型；缺点无法保证找到全











局最优模型；岭回归 (Ridge Regression)：使用 L2 范数惩罚；参数 lambda 控制回归拟合和系数收缩之间的平衡：0 为最小二乘估计；趋于无穷时估计值趋于 0；展示了随着 lambda 增加，各变量的标准化系数趋向于 0 的过程；增加 lambda 会增加偏差(Bias) 但减少方差(Variance)；LASSO：使用 L1 范数惩罚；通常没有显式解，需使用二次规划 (QP) 算法求解；Ridge vs. Lasso：都会收缩系数估计并引入偏差；Lasso 优势：产生更简单、更具可解释性的模型；Bridge 估计量 (Bridge Estimators)使用 Lr 范数 R=1 lasso /2 ridge /∞最大值惩罚；非负 Garrote：通过缩放因子最小化误差；s.t c_j ＞=0；$\min_c \frac{1}{2}\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p} c_j\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p} c_j$ 其他扩展 (Other Extensions)：Elastic Net (弹性网络)混合 L1 与 L2 惩罚；Group Lasso：当变量分组时，整组地包含或剔除变量；惩罚系数控制回归拟合精度与系数收缩程度之间的权衡；随着模型复杂度的增加，偏差逐渐减小，而方差逐渐增大；主成分分析 (PCA)：寻找数据主成分方向以及方差最大方向（但这可能不是信息量最大的）；主成分回归 (Principal Component Regression, PCR) 展示了随着主成分数量增加，PCR 的平方偏差、测试 MSE 和方差的变化趋势

2. 岭回归 (Ridge Regression)

定义与公式

- 岭回归使用 L_2**-范数惩罚**，即系数的平方和：$\lambda\|\beta\|_2^2 = \lambda\sum\beta_j^2$。
- **优化目标**：$\hat{\beta}_\lambda^{ridge} = \mathrm{argmin}(y - X\beta)^T(y - X\beta) + \lambda\|\beta\|_2^2$。
- **解**：岭回归有解析解（Closed-form solution）：$\hat{\beta}_\lambda^{ridge} = (X^TX + \lambda I_p)^{-1}X^Ty$。

性质

- **收缩作用**：惩罚项会将系数估计值向零收缩（Shrink towards zero）。
- **λ 的影响**：
  - 当 $\lambda = 0$ 时，就是普通的最小二乘法（LSE）。
  - 当 $\lambda \to \infty$ 时，所有系数趋近于零。
- **非稀疏性**：岭回归虽然会让系数变小，但通常**不会**让系数完全等于零（除非 λ 无穷大）。它保留了所有变量。

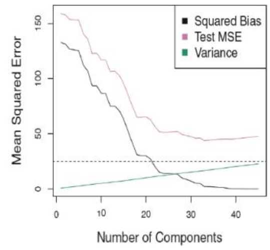
3. LASSO (Least Absolute Shrinkage and Selection Operator)

定义与公式

- LASSO 使用 L_1**-范数惩罚**，即系数的绝对值之和：$\lambda\|\beta\|_1 = \lambda\sum|\beta_j|$。
- **优化目标**：$\hat{\beta}_\lambda^{lasso} = \mathrm{argmin}(y - X\beta)^T(y - X\beta) + \lambda\|\beta\|_1$。
- **解**：一般情况下没有显式解析解，需要通过二次规划（QP）算法求解。

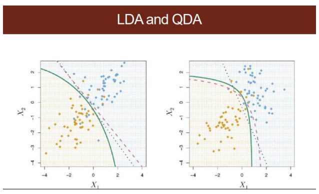
性质

- **稀疏解 (Sparse Solution)**：LASSO 的一个关键特性是它能产生稀疏解。这意味着某些系数会**恰好变为零**。因此，LASSO 可以同时进行**参数估计和连续的变量选择**。
- **模型解释性**：由于它剔除了部分变量，LASSO 生成的模型通常比岭回归更简单、更易解释。



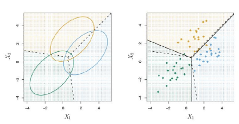
Topic 6 Classification

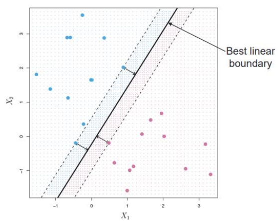
交叉验证;交叉验证是一种模型选择技术。它的主要目的是通过调整训练误差来估计测试误差（Test Error），从而解释过拟合现象;数学定义：输出变量是定性的；输入变量 x；分类器 G（x）$\hat{G}(x) = argmax_k \hat{h}_k(x)$ 目标是最小化误分类率 err（G）；分类函数通过 h_k(x)来构建分类器；二分类就是 k=2；分类器为 I 或者 sign（h（x））；分类边界就是类别 k 与 l：$\{x: h_k(x) = h_l(x)\}$；线性回归用于分类的问题：编码问题：对于多分类问题（如三种糖尿病），如果简单地将其编码为 1, 2, 3 并使用线性回归，由于类别间距离和顺序的假设（如类型 2 介于类型 1 和妊娠期糖尿病之间）可能不成立，会导致模型偏差；二分类可行性：对于二分类，将其转换为虚拟变量 (dummy variable) 并拟合线性回归



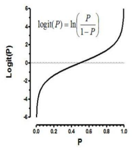
决策边界对比：

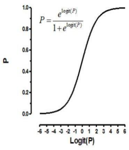
- **黑色点线 (Black: LDA)**：这是 LDA 强行拟合出的**直线**边界。由于它忽略了方差的差异，导致在左上角和右下角有明显的分类错误。
- **绿色实线 (Green: QDA)**：这是 QDA 拟合出的**曲线**边界。它呈现出弯曲的形状，能够更好地包裹住数据分布的特征，适应了不同类别方差不一致的情况。
- **紫色虚线 (Purple: Bayes rule)**：真实的最优边界也是弯曲的，QDA 比 LDA 更接近这条真实边界。





是可行的；线性回归估计的概率 h_k(X)可能小于 0 或大于 1，且存在“遮蔽问题 (Masking problem)”贝叶斯理论:最小化误差定义最优分类器把 hx 换成 px 后验概率；估计 px 的方法：判别分析；逻辑回归；分类树；深度神经网络；估计 G（x）：SVM；Boosting;Bagging;线性判别分析 (LDA)and 二次判别分析 (QDA);朴素贝叶斯 (Naïve Bayes):不假设特定的分布族，而是假设在每个类别内，预测变量之间是相互独立的；

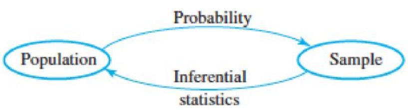




虽然独立性假设通常不成立（引入偏差），但它减少了方差。这种偏差-方差的权衡使得朴素贝叶斯在样本量 n $p(X)=\frac{e^{\beta_0+x^T\beta}}{1+e^{\beta_0+x^T\beta}}$。 相对 p 较小时表现良好;二元逻辑回归 (Binary Logistic Regression);直观展示了线性回归（直线，可能超出 0-1）与逻辑回归（S 形曲线，限制在 0-1）在预测违约概率时的区别;逻辑回归的拟合:极大似然估计 (MLE):通过最大化对数似然函数 l(beta)来估计系数;逻辑回归 vs LDA:两者都产生线性决策边界（Logit 是线性的）。区别在于拟合过程：逻辑回归使用 MLE（更少假设），LDA 使用均值和协方差的估计（基于正态假设）;分离超平面 (Separating Hyperplane):分类规则：根据 f(x^*) > 0 或 < 0 将测试点归类为 1 或 -1。点离超平面越远，分类置信度越高;最优分离超平面 (Optimal Separating Hyperplane):能完美分离训练数据的超平面可能有无限多个，需要选择最好的一个;最优超平面应使得到最近数据点的距离（间隔 Margin）最大化;通过拉格朗日乘子法转化为 Wolfe 对偶形式，利用二次规划 (Quadratic Programming, QP) 求解.分类模型评估:误分类误差；ROC 曲线（TP 对 FP 的曲线）；Precision=TP/P;F1 score；AUC =ROC 曲线的面积；混淆矩阵的格式.

True Neg. (TN)	False Pos. (FP)
False Neg. (FN)	True Pos. (TP)

Topic1 Review



总体 (Population): 我们试图得出结论的个体的全体集合；样本 (Sample): 被观察到的总体的一部分；关系: 概率论 (Probability) 从总体推导样本性质，推断统计 (Inferential statistics) 从样本推断总

两个事件：$P(A\cup B)=P(A)+P(B)-P(A\cap B)$（减去重复计算的交集部分）。

三个事件：$P(A\cup B\cup C)$ 需要用到容斥原理，加回被减去过多的三者交集部分 $+P(A\cap B\cap C)$。

体性质；试验 (Experiment): 任何产生观测结果的动作；样本空间

Page 13: 条件概率与贝叶斯 (Conditional Probability)

- 条件概率: $P(A|B)=\frac{P(A\cap B)}{P(B)}$。
- 独立性: 若 $P(A\cap B)=P(A)P(B)$ 或 $P(A|B)=P(A)$，则 A 与 B 独立。
- 贝叶斯定理: $P(A_i|B)=\frac{P(A_i)P(B|A_i)}{\sum P(A_k)P(B|A_k)}$。

(Sample Space, S): 试验所有可能结果的集合；随机变量 (Random Variable)；离散型随机变量 (Discrete Random Variable)：概率质量函数 (PMF) 和 累积分布函数 (CDF)；常见离散分布；连续型随机变量 (Continuous Random Variable)：结果构成实数轴上的区间：概率密度函数 (PDF) 和 累积分布函数 (CDF, 积分形式);中心极限定理(CLT): 样本均值的标准化形式趋近于标准正态分布 N(0,1)统计推断 (Statistical Inference)；点估计 (Point Estimation)；无偏估计量 (Unbiased Estimator)：定义：若 E(hat{theta})=theta，则 hat{theta} 是无偏的；最小方差无偏估计 (MVUE)：在所有无偏估计量中方差最小的一个；比较：比较不同估计量的方差大小；矩估计法 (Method of Moments, MM)；令样本矩等于总体矩，解方程求出参数；

- 伯努利分布: $Bern(p)$。
- 二项分布: $Bin(n,p)$，公式 $C_n^x p^x(1-p)^{n-x}$。
- 泊松分布: $Poi(\lambda)$，公式 $\frac{\lambda^x}{x!}e^{-\lambda}$。

定义: k阶总体矩 $E(X^k)$，k阶样本矩 $\frac{1}{n}\Sigma X_i^k$

MLE：使似然函数 L(θ)=f(x1,⋯,xn; θ)最大的 θ 值;正态分布：μ^ =x¯，σ^2=1/nΣ(xi−x¯)2；似然函数：将概率密度函数（pdf）中的参数视为变量、数据视为固定值，得到的关于参数的函数；；**置信区间**：构建一个 $\mathrm{CI}=估计值\pm(z_{临界值}\times 标准误)$ 区间，使其以 1-alpha 的概率覆盖真实参数；含义：重复抽样中约 100(1−α)%的区间包含参数 θ（参数固定，区间随机）；假设检验：用数据在 H0（原假设，最初假定为真，通常含等号）和 Ha（备择假设，对立，通常含不等号）之间做决策；决策：拒绝 H0 or 不拒绝 H0（不拒绝≠接受）。错误类型：Type I 错误（拒绝真 H0，概率 ＝α，即显著性水平）；Type II 错误（不拒绝假 H0，概率 ＝β）；功效 ＝1−β（正确拒绝假 H0 的概率）。通常 α↑⇒β↓。检验构造原则：在控制 α 的前提下，最小化 β 最大化功效）。P 值：在 H0 为真的假设下，得到至少与观测值一样极端的检验统计量的概率；也是能拒绝 H0 的最小显著

2. 线性判别分析 (LDA)

此部分主要依据 **Topic 6 Classification** (第 12-15 页)。

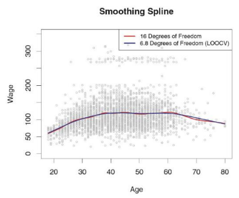
核心理论

- **贝叶斯定理**：LDA 试图通过估计 $f_k(X)$（类条件密度）和 π_k（先验概率）来计算后验概率 $p_k(X)$。
- **关键假设**：
  1. **正态分布**：假设每类数据服从多元正态分布 $X|y=k\sim N_p(\mu_k,\Sigma)$。
  2. **方差相等 (Equal Covariance)**：假设**所有类别**共用同一个协方差矩阵 Σ。这是 LDA 最核心的假设。
- **结果**：由于假设协方差矩阵相同，判别函数中的二次项 $X^T\Sigma^{-1}X$ 相互抵消，剩下的判别函数 $\delta_k(X)$ 是关于 X 的**线性函数**。
- **决策边界**：分类边界由线性方程决定，因此在几何上表现为直线（或超平面）。

λ 的作用：

- $\lambda=0$：函数 g 会穿过所有训练点（插值），可能过拟合。
- $\lambda\to\infty$：平滑样条退化为简单的线性回归（直线）。

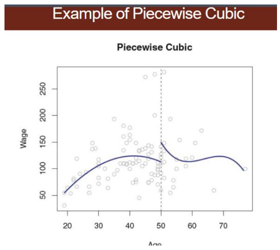
选择：λ 控制偏差-方差权衡，可以通过交叉验证确定。



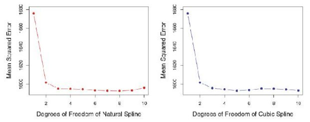
性水平。决策规则：p-value＜αα 则拒绝 H0;常见 α=0.05 但需谨慎使用。P 值越小，反对 H0 的证据越强。

Topic 7 Moving beyond Linearity: 多项式回归 (Polynomial

模型： 多项式逻辑回归假设 logit 概率与预测变量呈多项式关系，即 $logit(Pr(y_i > 250|x_i)) = \beta_0 + \beta_1 x_i + \cdots + \beta_d x_i^d$。

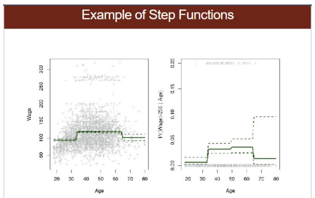


Regression) 模型形式为 $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_d x_i^d + \epsilon_i$ 多项逻辑回归



(Polynomial Logistic Regression);阶梯函数 (Step Functions):阶梯函数用于局部近似非线性结构。它将连续变量转换为有序的二元变量（哑变量）;在 X 的范围内设定 k 个断点。构建 k+1 个变量利用指示函数 I 来表示 x 落在哪个区间；所有指示函数之和为 1；X 必须为 K+1 个区间当中一个；

模型： 线性回归函数被替换为 $y_i = \beta_0 + \beta_1 C_1(X_i) + \cdots + \beta_K C_K(X_i) + \epsilon_i$。 ；



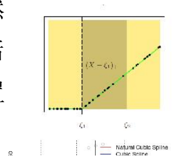
交互项回归模型 (Interaction Regression Model)：有多个预测变量时，可以考虑交互作用：例如 $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_{12} x_{i1} x_{i2} + \epsilon_i$ ；分段多项式

For example, a piecewise cubic regression model is

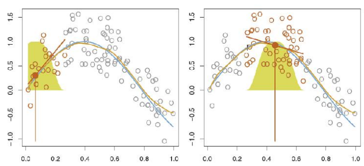
$$y_i = \begin{cases} \beta_{01} + \beta_{11}x_i + \beta_{21}x_i^2 + \beta_{31}x_i^3 + \epsilon_i & if\, x_i < c \\ \beta_{02} + \beta_{12}x_i + \beta_{22}x_i^2 + \beta_{32}x_i^3 + \epsilon_i & if\, x_i \geq c \end{cases}$$

c is a **knot**.

(Piecewise Polynomials)；连续分段多项式 (Continuous Piecewise Polynomials)：可以添加约束条件，强制拟合曲线必须是连续的，一阶导二阶导也连续，保证曲线平滑；基函数 (Basis Functions)：**通用框架：** 许多非线性模型可以统一写成 $y_i = \beta_0 + \sum \beta_k b_k(x_i)$ 的形式，其中 $b_k(\cdot)$ 是已知且固定的基函数。例如阶梯函数 bk 是指示函数 I；三次样条 (Cubic Spline)：具有连续一阶和二阶导数的连续分段多项式；截断幂基函数（truncated power basis function）：$h(x, \xi) = (x - \xi)_+^3$



自然样条 (Natural Cubic Spline)：自然样条要求模型在边界区域（即第一个节点之前和最后一个节点之后）必须是线性的（红色实线）；确定节点 (Determining Knots)：均匀放置或者交叉验证之后选择节点数量；展示了均方误差（MSE）随自由度（Degrees of Freedom，与节点数量相关）变化的曲线。呈现 U 型或 L 型，用于选择最优的模型复杂度；平滑样条的性质：最优解 \hat{g}(x)是一个自然三次样条，其节点位于每一个唯一的 x_i 上；调节参数 (Tuning Parameter)；红色线（16 Degrees of Freedom）：看起来比较“抖动”，拟合过度；蓝色线（6.8 Degrees of Freedom, LOOCV 选定）：通过留一法交叉验证选出的最优平滑度，曲线更加平滑且合理；局部线性回归 (Local Linear Regression)目标：最小化加权残差平方和；黄色区域/钟形曲线表示权重函数；左图：针对某个数据点(橙色实心点)，在局部范围内拟合一条直线(橙色线)，该直线只受附近点的影响；广义加性模型 (GAMs)：优点：能自动对每个变量进行非线性建模；具有可加性，解释性强（可以单独考察每个变量的影响）；加性形式排除了变量间的交互作用。如果需要交互，必须手动将交互项 f_{jk}(X_j, X_k)加入模型；



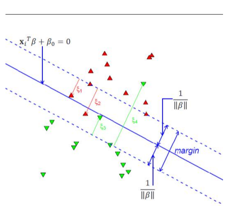
Topic 8 Classification and Regression Trees (CART)

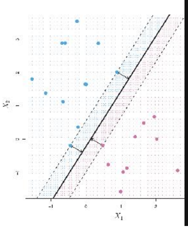
Topic 9 SVM

最大间隔分类器 (Maximal Margin Classifier) —— 理想的线性情况；基本概念：当数据是线性可分（Linearly Separable）的，即可以用一条直线（或超平面）完美地将两类数据分开时，SVM 试图找到一个“最优”的超平面 。什么是“最优”？：能将两类数据分开的超平面可能有无数个。SVM 选择的是那个拥有最大间隔（Maximal Margin）的超平面 。间隔 (Margin)：是指数据点到超平面的最小垂直距离。最大化间隔意味着在这个

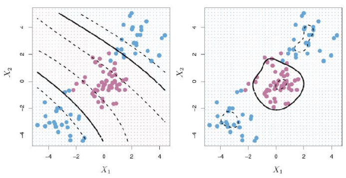
3. 支持向量机 (SVM) —— 非线性情况

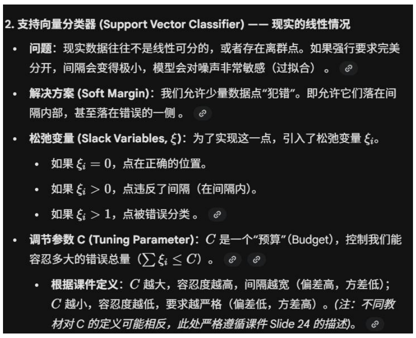
- **核心思想：** 当数据在当前维度（如 2D）线性不可分时，SVM 通过将数据映**射到更高维的空间**（如 3D 或更高），使得数据在那个高维空间中变得线性可分。
- **核技巧 (Kernel Trick)：** 直接在高维空间计算是非常耗时的。SVM 使用“核函数”$K(x, x')$，它能够在低维空间中直接计算出高维空间的内积，从而避免了显式的维度转换，极大地提高了计算效率。
- **常用核函数：**
  - **多项式核 (Polynomial)：** $K(x, x') = (1 + x^T x')^d$。
  - **径向基核 (Radial/Gaussian)：** $K(x, x') = \exp(-\gamma||x - x'||^2)$。这是一种局部核，只有附近的点才会影响预测。





超平面两侧留出了最宽的“无人区”（Slab），这使得模型对新数据的泛化能力更强 。支持向量 (Support Vectors)：那些恰好落在间隔边缘上的**数据点**被称为**支持向量**。它们支撑着这个超平面，如果移动非支持向量的数据点，决策边界不会改变；但如果移动支持向量，决策边界就会改变；**图解**：展示了“最优”的那条分界线（实线）及其间隔（虚线）；支持向量：请注意那些正好落在虚线上的有箭头的点（两个绿色，一个红色）。这些点就是支持向量；展示了支持向量分类器（软间隔）的思想。为了获得更宽、更稳健的间隔（由虚线表示），我们允许少量数据点违反规则（即落在间隔内或被错分）。这是一种“退一步海阔天空”的策略；超参数 C 越大：意味着对错误的容忍度极低（惩罚重）。模型会试图正确分类每一个点，导致间隔变窄；展示了如果我们引入多项式特征；在原始空间看，分界线就变成了非线性的曲线，从而能解决复杂分类问题；左 1 展示了不同核函数（三阶多项式核/径向基核 RBF）的特性。RBF 核具有很强的局部性 (Local behavior)，非常适合处理这种“中心包围”或局部聚集的复杂分布；





2. 支持向量分类器 (Support Vector Classifier) —— 现实的线性情况

- **问题**：现实数据往往不是线性可分的，或者存在离群点。如果强行要求完美分开，间隔会变得极小，模型会对噪声非常敏感（过拟合）。
- **解决方案 (Soft Margin)**：我们允许少量数据点“犯错”。即允许它们落在间隔内部，甚至落在错误的一侧。
- **松弛变量 (Slack Variables, ξ)**：为了实现这一点，引入了松弛变量 ξ_i。
  - 如果 $\xi_i = 0$，点在正确的位置。
  - 如果 $\xi_i > 0$，点违反了间隔（在间隔内)。
  - 如果 $\xi_i > 1$，点被错误分类。
- **调节参数 C (Tuning Parameter)**：C 是一个“预算”(Budget)，控制我们能容忍多大的错误总量（$\sum \xi_i \leq C$）。
  - **根据课件定义**：C 越大，容忍度越高，间隔越宽（偏差高，方差低）；C 越小，容忍度越低，要求越严格（偏差低，方差高）。*(注：不同教材对 C 的定义可能相反，此处严格遵循课件 Slide 24 的描述)。*