这份PDF是关于**线性模型：分类（Linear Models: Classification）**的课程讲义。为了方便你复习和记忆，我将内容提炼为四个核心模块，并补充了相关的背景知识、数学推导要点和面试常见考点。

模块一：分类模型基础 (Slide 4-15)

1. 核心概念

- **目标**：将输入向量 $\mathbf{x}$ 分配到 K 个离散类别 $\mathcal{C}_k$ 中的一个。
- **判别函数 (Discriminant Functions)**：直接将 $\mathbf{x}$ 映射到类别的函数。
- **线性判别边界**：
    - 二分类方程：$y(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}\mathbf{x} + w_0$。
    - 决策规则：若 $y(\mathbf{x}) \geq 0$，判为 $\mathcal{C}_1$；否则判为 $\mathcal{C}_2$。

2. 几何意义（重点记忆）

- **方向**：权重向量 $\mathbf{w}$ 是决策平面的**法向量**，决定了平面的方向。
- **位置**：偏置 w_0 决定了平面的位置。原点到平面的垂直距离为 $\frac{-w_0}{\|\mathbf{w}\|}$。
- **距离**：任意点 $\mathbf{x}$ 到决策平面的符号距离为 $r = \frac{y(\mathbf{x})}{\|\mathbf{w}\|}$。

3. 多分类策略及其缺陷

- **One-versus-the-rest (一对其余)**：需 $K-1$ 个分类器。缺陷：会出现“无法确定的区域”（如图 Slide 11 绿色部分）。
- **One-versus-one (一对一)**：需 $K(K-1)/2$ 个分类器。缺陷：同样存在决策冲突区域。
- **线性多分类判别器（推荐）**：使用 K 个线性函数 $y_k(\mathbf{x}) = \mathbf{w}_k^{\mathrm{T}}\mathbf{x} + w_{k0}$。选 y_k 最大者为结果。
    - **重要特性**：其决策区域是**单连通且凸的 (Convex)**。

模块二：最小二乘法分类 (Least Squares for Classification, Slide 16-23)

1. 原理

- 将分类任务看作回归任务，目标值 t 采用 **1-of-K 编码**（如类别2为 $[0,1,0]$）。
- 通过最小化平方误差和 (Sum-of-squares error) 来求解参数矩阵 $\tilde{\mathbf{W}}$。

- **解**：$\tilde{\mathbf{W}} = (\tilde{\mathbf{X}}^{\mathrm{T}}\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^{\mathrm{T}}\mathbf{T} = \tilde{\mathbf{X}}^{\dagger}\mathbf{T}$（伪逆矩阵）。

2. 优缺点分析 (常考点)

- **优点**：有解析解，计算直接。
- **缺点（致命伤）**：
    - **缺乏鲁棒性**：对离群点（Outliers）极其敏感。因为平方误差会给远离边界的正确分类点很大的惩罚，导致边界偏移（Slide 22）。
    - **“屏蔽”现象 (Masking)**：当某些类别的点被夹在中间时，最小二乘法可能会完全漏掉某些类别（Slide 23）。

模块三：感知机 (Perceptron, Slide 24-39)

1. 模型定义

- **形式**：$y(\mathbf{x}) = f(\mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}))$。
- **激活函数**：阶跃函数（Step function），输出 $\{+1, -1\}$。
- **目标编码**：$t = 1$ 对应 $\mathcal{C}_1$，$t = -1$ 对应 $\mathcal{C}_2$。

2. 感知机准则 (Perceptron Criterion)

- **误差函数**：$E_P(\mathbf{w}) = -\sum_{n\in\mathcal{M}} \mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}_n)t_n$。
    - 其中 $\mathcal{M}$ 是**被错分**的样本集合。
    - 正确分类时 $\mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}_n)t_n > 0$，错误时 < 0，故加上负号使误差为正。

3. 优化与收敛

- **算法**：随机梯度下降 (SGD)。
- **更新规则**：$\mathbf{w}^{(\tau+1)} = \mathbf{w}^{(\tau)} + \eta\phi(\mathbf{x}_n)t_n$。
    - *直观理解*：如果分错了，就把权重向正确方向“拨”一下。
- **感知机收敛定理 (重点)**:
    - 如果训练数据是**线性可分**的，算法保证在**有限步**内收敛。
    - 如果数据**不可分**，算法将**永远不会收敛**（产生震荡）。

模块四：逻辑回归 (Logistic Regression, Slide 40-70)

1. 核心模型

- **目标**：预测概率 $0 \leq h_w(\mathbf{x}) \leq 1$。
- **假设函数**：$h_w(\mathbf{x}) = \sigma(\mathbf{w}^{\mathrm{T}}\mathbf{x}) = \frac{1}{1+\exp(-\mathbf{w}^{\mathrm{T}}\mathbf{x})}$。
- **Sigmoid函数特性**：将实数映射到 $(0, 1)$。

2. 决策边界

- **线性**：当 $\mathbf{w}^{\mathrm{T}}\mathbf{x} = 0$ 时，预测概率为 0.5。
- **非线性**：通过引入高级特征（如 x_1^2, x_2^2），逻辑回归可以产生圆形成或其他形状的决策边界（Slide 56-58）。

3. 代价函数 (Cost Function) - 重点理解

- **为什么不用最小二乘?**：逻辑回归结合 MSE 会导致非凸 (Non-convex) 函数，容易陷入局部最优，且梯度消失。
- **对数损失 (Log Loss / Cross-Entropy)**：
  - $cost(h_w(\mathbf{x}), t) = -t\log(h_w(\mathbf{x})) - (1-t)\log(1-h_w(\mathbf{x}))$
  - *逻辑*：如果 $t = 1$，希望 h_w 接近 1；如果 $t = 0$，希望 h_w 接近 0。不符则惩罚无限大（Slide 62）。

4. 优化：梯度下降

- **更新公式**：$w_j := w_j - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_w(\mathbf{x}^{(i)}) - t^{(i)}) x_j^{(i)}$。
- **注意**：这个公式的形式竟然与线性回归完全一致，但不同之处在于 $h_w(\mathbf{x})$ 的定义。

复习建议与考点预测

1. **比较题**：感知机 vs 逻辑回归。
   - *相同*：都是线性分类器，都用梯度下降。
   - *不同*：感知机输出离散值（阶跃函数），不可导，只能处理线性可分；逻辑回归输出概率（Sigmoid），可导，可处理概率问题。
2. **几何题**：给定 $\mathbf{w}$ 和 w_0，判断点的类别或计算到边界的距离。
3. **计算题**：执行一步感知机更新（$\mathbf{w} + \eta\mathbf{x}t$）或逻辑回归的梯度计算。
4. **理解题**：为什么最小二乘法不适合分类？（Outliers 和 Masking 问题）。

笔记完善建议：
在复习时，可以尝试手画 Slide 22 (Outlier) 和 Slide 34-36 (Perceptron update) 的过程，这能帮你极大地加深对线性模型局限性和工作原理的理解。