这份PDF是关于**线性模型：回归（Linear Models: Regression）**的机器学习课程讲义。它涵盖了从基础的单变量线性回归到复杂的贝叶斯线性回归的内容。

以下是根据PDF内容整理的详细复习笔记，按逻辑板块划分，方便你记忆和理解：

一、 基础概念 (Slides 1-13)

1. **监督学习 (Supervised Learning)**：
   - 核心：给定每个训练样本的"正确答案"（标签）。
2. **回归 (Regression)**：
   - 目标：预测**连续型（Real-valued）**的输出。
   - 案例：智能手机价格预测（根据屏幕尺寸预测价格）。
3. **符号定义 (Notations)**：
   - m：训练样本的数量。
   - x：输入变量/特征（Input variable/features）。
   - y：输出变量/目标值（Output variable）。
   - $(x^{(i)}, y^{(i)})$：第 i 个训练样本。
4. **模型表示 (Hypothesis)**：
   - **单变量线性回归 (Univariate Linear Regression)**：
     $h_w(x) = w_0 + w_1 x$
   - w_0, w_1 是模型参数。

二、 代价函数与目标 (Slides 14-20)

1. **代价函数 (Cost Function)**：使用**均方误差 (Mean Squared Error, MSE)**。
   $J(w_0, w_1) = \frac{1}{2m} \sum_{i=1}^{m} (h_w(x^{(i)}) - y^{(i)})^2$
   - *注意*：分母中的 2 是为了在求导时抵消平方项，方便计算。
2. **优化目标 (Goal)**：找到使 $J(w_0, w_1)$ 最小的参数。
   $\arg\min_{w_0, w_1} J(w_0, w_1)$
3. **直观理解**：
   - 固定参数时，$h_w(x)$ 是关于 x 的函数（一条直线）。
   - $J(w_0, w_1)$ 是关于参数的函数（在 3D 空间是一个碗状的**凸函数**，在 2D 空间是等高线图）。

三、 梯度下降法 (Gradient Descent) (Slides 21-40)

1. **核心思想**：从某个初始参数出发，沿着代价函数减小最快的方向（负梯度方向）迭代更新参数。
2. **算法公式**：
   重复直至收敛 {
   $w_j := w_j - \alpha \frac{\partial}{\partial w_j} J(w_0, w_1) \quad (\text{for } j = 0 \text{ and } j = 1)$
   }
   - α：**学习率 (Learning rate)**，控制步长。
3. **关键点：同时更新 (Simultaneous Update)**：
   - 必须先计算出所有参数的偏导数，再一起更新。如果在计算 w_1 前就更新了 w_0，则不再是标准的梯度下降。
4. **学习率 α 的特性**：
   - 即使 α 固定，随着接近局部最小值，梯度会变小，梯度下降会自动采取更小的步长。因此，通常不需要随时间减小 α。
5. **线性回归的具体更新规则**：
   - $w_0 := w_0 - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_w(x^{(i)}) - y^{(i)})$
   - $w_1 := w_1 - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_w(x^{(i)}) - y^{(i)}) \cdot x^{(i)}$

四、 多变量线性回归 (Multivariate Linear Regression) (Slides 41-58)

1. **多特征表示**：
   - n：特征数量。
   - $x_j^{(i)}$：第 i 个样本的第 j 个特征。
2. **向量化表示 (Vectorized Notation)**：
   - 定义 $x_0 = 1$。
   - 特征向量 $\mathbf{x} = [x_0, x_1, \ldots, x_n]^T \in \mathbb{R}^{n+1}$。
   - 参数向量 $\mathbf{w} = [w_0, w_1, \ldots, w_n]^T \in \mathbb{R}^{n+1}$。
   - **假设函数**：$h_w(\mathbf{x}) = \mathbf{w}^T \mathbf{x}$。
3. **正规方程 (Normal Equation)**：一种直接求出最优参数的解析方法（不需要迭代）。
   $\mathbf{w} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$

- $\mathbf{X}$ 是所有输入特征组成的矩阵（包含 $x_0 = 1$ 列）。

五、 特征选择与广义线性回归 (Slides 59-65)

1. **特征变换**：可以通过组合现有特征生成新特征（例如：面积 = 长 × 宽）。
2. **多项式回归 (Polynomial Regression)**：
   - 将模型变为非线性曲线：$h_w(x) = w_0 + w_1 x + w_2 x^2 + w_3 x^3$。
   - 令 $x_1 = x, x_2 = x^2, x_3 = x^3$，它依然可以看作是一个线性回归问题。
3. **基函数 (Basis Functions)**：
   - 通用形式：$\Phi(\mathbf{x}) = [\phi_1(\mathbf{x}), \phi_2(\mathbf{x}), \ldots, \phi_n(\mathbf{x})]^T$。

六、 正则化 (Regularization) (Slides 66-72)

1. **过拟合 (Overfitting)**：
   - **欠拟合 (Underfit)**：高偏差 (High Bias)，模型太简单。
   - **过拟合 (Overfit)**：高方差 (High Variance)，模型太复杂，完美拟合训练数据但泛化能力差。
2. **正则化代价函数 (L2 Regularization)**：
   - 在代价函数后添加惩罚项，抑制参数 w 过大。
     $J(\mathbf{w}) = \frac{1}{2m}\left[\sum_{i=1}^{m}(h_w(\mathbf{x}^{(i)}) - y^{(i)})^2 + \lambda \sum_{j=1}^{n} w_j^2\right]$
   - *注意*：通常不惩罚 w_0。
3. **正则化下的参数更新**：
   - 梯度下降：$w_j := w_j(1 - \alpha\frac{\lambda}{m}) - \alpha \ldots$ (每次更新前先对 w_j 进行一定的衰减)。
   - 正规方程：$\mathbf{w} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{L})^{-1}\mathbf{X}^T\mathbf{y}$，其中 $\mathbf{L}$ 是除了第一个元素为0外，其余对角线为1的单位阵。

七、 概率视角与贝叶斯线性回归 (Slides 73-97)

1. **最大似然估计 (MLE) 与 最小二乘法 (Least Squares)**：
   - 假设噪声服从高斯分布 $\epsilon \sim \mathcal{N}(0, \beta^{-1})$。
   - 结论：**最小化平方和误差等价于最大化高斯噪声下的似然函数。**
2. **偏差-方差分解 (Bias-Variance Decomposition)**：

    - **期望损失 = 偏置^2 + 方差 + 噪声。**
    - 目标是找到三者之和最小的平衡点。
3. **贝叶斯处理 (Bayesian Treatment)**:
    - **先验 (Prior)**: $p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_0, \mathbf{S}_0)$。
    - **后验 (Posterior)**: $p(\mathbf{w}|\mathbf{t}) \propto p(\mathbf{t}|\mathbf{w})p(\mathbf{w})$。
    - **最大后验估计 (MAP)**: 等价于带正则化的最小二乘法，其中正则化系数 $\lambda = \alpha/\beta$。
4. **预测分布 (Predictive Distribution)**:
    - 贝叶斯不只给出一个 w 的点估计，而是给出一个分布。
    - 随着观测数据 N 增加:
        - 后验分布变窄（不确定性减小）。
        - 预测分布的方差趋向于数据本身的固有噪声 β^{-1}。

复习重点建议:

- 掌握梯度下降的**更新公式**（尤其是多变量和正则化版本）。
- 理解**正规方程**的适用场景。
- 深刻理解**过拟合**的原因及**正则化**的原理（惩罚大权重）。
- 贝叶斯部分：理解**MLE、MAP与最小二乘法**之间的数学联系。