机器学习复习笔记：模型评估与选择

一、 核心概念：误差与拟合 (Error & Fitting)

1. **误差分类**：
    - **错误率 (Error Rate)**：$E = a/m$（a个样本分类错误，m个总样本）。
    - **精度 (Accuracy)**：$1 - E$（分类正确的比例）。
    - **训练误差 (Training/Empirical Error)**：模型在训练集上的误差。
    - **泛化误差 (Generalization Error)**：模型在新样本（未来样本）上的误差。**学习的目标是最小化泛化误差。**
2. **拟合程度**：
    - **欠拟合 (Underfitting)**：模型对训练样本的一般性质没学好。
    - **过拟合 (Overfitting)**：模型把训练样本自身的特性当成了所有样本的共性。
    - **重要结论**：过拟合无法完全避免，只能缓解。机器学习问题通常是NP难的，而有效的算法需在多项式时间内完成，若能彻底避免过拟合则意味着证明了$P = NP$。

二、 评估方法：如何获得"泛化误差"的近似 (Evaluation Methods)

为了评估模型，需将数据集 $\mathcal{D}$ 划分为 **训练集** $\mathcal{S}$ 和 **测试集** $\mathcal{T}$（互斥）。

1. **留出法 (Hold-out)**：
    - 直接划分为两个互斥集合。
    - 注意点：需保持数据分布一致性（使用**分层采样 Stratified Sampling**）。
    - 比例：通常训练集占 2/3 ~ 4/5。
2. **交叉验证法 (Cross Validation)**：
    - **k折交叉验证 (k-fold)**：将 $\mathcal{D}$ 分成 k 个子集，轮流用 $k-1$ 个训练，1个测试，取 k 次结果的均值（常用 $k = 10$）。
    - **留一法 (LOO)**：$k = m$（样本数）的极端情况。优点是评估最准，缺点是数据集大时计算成本极高。
3. **自助法 (Bootstrapping)**：
    - **有放回采样**：产生一个和原数据集一样大的训练集。
    - **数学结论**：约有 **36.8%** 的样本从未被抽到，称为"**包外估计 (out-of-bag estimate)**"，可直接作为测试集。
    - 优点：在数据集较小时非常有用。

4. **调参与验证集**：
    - **参数调节 (Parameter Tuning)**：通过范围和步长搜索最优超参数（如网格搜索）。
    - **验证集 (Validation Set)**：在模型选择和调参过程中用于评估的数据集，需与最终评估性能的测试集区分开。

三、 性能度量 (Performance Measures)

性能度量是衡量模型泛化能力的数学标准。

1. **回归任务**：常用 **均方误差 (MSE)**。
2. **分类任务基础：混淆矩阵 (Confusion Matrix)**
    - **TP** (真正), **FP** (假正), **TN** (真负), **FN** (假负)。
3. **查准率 (Precision) vs. 查全率 (Recall)**：
    - $P = TP/(TP + FP)$：预测为正的样本中有多少是真的正例（侧重"准"）。
    - $R = TP/(TP + FN)$：真实正例中有多少被预测出来了（侧重"全"）。
    - 两者通常相互矛盾。
4. **综合指标**：
    - **F1 Measure**：P和R的调和平均，$\frac{1}{F1} = \frac{1}{2}(\frac{1}{P} + \frac{1}{R})$。
    - F_β **Measure**：引入 β 指定重要程度。$\beta > 1$ 侧重查全率（如医疗诊断），$\beta < 1$ 侧重查准率。
    - **宏平均 (Macro)**：先计算每个类的P、R，再求平均。
    - **微平均 (Micro)**：先将所有类的TP、FP等元素平均，再计算P、R。

四、 评价曲线与面积 (Visualizing Performance)

1. **P-R 曲线**：
    - 横轴 Recall，纵轴 Precision。
    - 比较：若曲线 A 完全包住 B，则 A 优于 B；否则看**平衡点 (BEP)**（$P = R$的点）或曲线下方的面积。
2. **ROC 曲线与 AUC**：
    - **横轴 FPR (假正例率)**：$FP/TN + FP$（负例被错判的比例）。
    - **纵轴 TPR (真正例率)**：$TP/TP + FN$（同查全率）。
    - **AUC (Area Under ROC Curve)**：ROC 曲线下的面积。

    - **重要意义**：AUC 反映了模型对样本的**排序质量**。$AUC = 1 - \ell_{rank}$，其中 ℓ_{rank} 是排序损失。
3. **代价敏感 (Cost Sensitive)**：
    - 不同类型的错误后果不同（如漏诊癌症 vs 误诊癌症），在 ROC 曲线中通过调整工作点（阈值）来平衡不同错误造成的代价。

五、 实践：如何手绘 ROC 曲线

- **第一步**：将模型预测的概率从高到低排序。
- **第二步**：依次将每个样本的预测概率作为“分类阈值”。
- **第三步**：
    - 若当前样本为**真正例**，坐标向上移动（$1/m^+$）。
    - 若当前样本为**假正例**，坐标向右移动（$1/m^-$）。
- **最终结果**：形成一条从 (0,0) 到 (1,1) 的折线。

复习建议：

- **高频考点**：P、R、$F1$ 的计算；ROC 与 P-R 曲线的区别；自助法 36.8% 的来源；过拟合与 $P \neq NP$ 的哲学联系。
- **理解难点**：宏平均与微平均的区别（Macro 是平均的平均，Micro 是加权后的平均）；AUC 与排序损失的关系。