Okay, I understand. I've incorporated Chinese explanations into the study notes to enhance clarity and understanding. Here's the revised version:

CBU5201 Principles of Machine Learning Methodology I & II

Part I: Methodology I

I. 导论 (Introduction)

1. **机器学习与传统编程 (Machine Learning vs. Traditional Programming):**
    - **机器学习 (Machine Learning):** 旨在构建在*目标总体样本*上运行良好的解决方案。它与传统编程的不同之处在于它缺乏对目标总体的精确*描述*。
        - "In machine learning, we lack a **description** of the target population." (在机器学习中，我们缺乏对目标总体的**描述**。)
        - "All we can do is extract samples from the population (known as **sampling the population**)." (我们所能做的就是从总体中提取样本（称为**对总体进行抽样**）。)
    - **数据集 (Dataset):** 提供目标总体*经验描述*的样本集合。
        - "A collection of samples extracted from a population forms a dataset. A dataset provides an **empirical** description of our target population." (从一个总体中提取的样本集合形成了一个数据集。数据集提供了对我们目标总体的**经验**描述。)
2. **抽样 (Sampling):**
    - **代表性 (Representative):** 数据集必须准确反映目标总体。
        - "Datasets are **representative**, i.e. provide a complete picture of the target population." (数据集具有**代表性**，即提供目标总体的完整 ⓒ。)
    - **独立同分布 (Independent and Identically Distributed (iid)):** 样本应独立提取且服从相同分布。
        - "Sampling mimics the mechanism that generates samples during deployment: Samples need to be extracted **independently**." (采样模拟了部署期间生成样本的机制：样本需要**独立**提取。)
        - "Samples need to be **independent and identically distributed** (**iid**)." (样本需要**独立同分布** (**iid**)。)

II. 机器学习中的核心任务 (Core Tasks in Machine Learning)

1. **部署性能 (Deployment Performance):**
    - **目标 (Goal):** 最终目标是构建具有高部署性能的模型，这意味着它们在目标总体上运行良好。
        - "The best model is the one with the highest **deployment performance**, i.e. the one that works best on the **target population**." (最佳模型是具有最高**部署性能**的模型，即在**目标总体**上效果最好的模型。)
    - **评估策略 (Evaluation Strategy):** 包括*质量指标*和如何使用*数据*评估性能的计划。
        - "A **quality metric** used to quantify the performance." (用于量化性能的**质量指标**。)
        - "How **data** will be used to assess the performance of a model." (如何使用**数据**来评估模型的性能。)
2. **测试 (Testing):**
    - **测试数据集 (Test Dataset):** 用于估计*真实部署性能*的数据子集。
        - "Instead we use a subset of data, the **test dataset**, to compute the **test deployment performance** as an **estimation** of the true performance." (相反，我们使用一个数据子集，即**测试数据集**，来计算**测试部署性能**，作为对真实性能的**估计**。)
    - **随机性 (Randomness):** 由于测试数据集提取的随机性，测试性能是一个随机变量。
        - "Test datasets are extracted randomly. Hence, the **test performance** is itself **random**, as different datasets generally produce different values." (测试数据集是随机提取的。因此，**测试性能**本身是**随机**的，因为不同的数据集通常会产生不同的值。)
    - **注意 (Caution):** 模型比较应考虑随机性；优势可能是偶然的。
        - "Caution should be used, as the test performance is a random quantity, hence some models might appear to be **superior by chance**!" (应谨慎使用，因为测试性能是一个随机量，因此某些模型可能看起来**侥幸更优**！)
3. **训练 (Training):**
    - **训练数据集 (Training Dataset):** 用于找到模型参数的最佳值。
        - "We use a subset of data, known as the **training dataset**, to (implicitly or explicitly) **reconstruct** the error surface needed during optimisation." (我们使用一个数据子集，称为**训练数据集**，来（隐式或显式地）**重构**优化期间所需的误差曲面。)
    - **经验误差曲面 (Empirical Error Surface):** 基于训练数据的真实误差曲面的近似值。
        - "The empirical and true error surfaces are in general different. Hence, their optimal models might differ, i.e. the best model for the training dataset might not be the best for the population." (经验误差曲面和真实误差曲面通常是不同的。因此，它们的最优模型可能不同，即训练数据集的最佳模型可能不是总体的最佳模型。)
4. **验证 (Validation):**
    - **目的 (Purpose):** 在最终训练之前比较不同的建模选项并选择最佳选项。
        - "**Validation**: Necessary to compare different modelling options and select the best one, the one that will be trained." (**验证**：比较不同建模选项并选择最佳选项（将要训练的选项）的必要步骤。)
    - **超参数 (Hyperparameter):** 用于控制学习过程的参数，例如多项式次数 D。

- "The polynomial degree D is a **hyperparameter**, as for each value of D a different family of models is obtained. How can we select the right value of a D?" (多项式次数 D 是一个**超参数**，因为对于每个 D 值，都会获得不同的模型族。我们如何选择正确的 D 值？)

- **验证方法 (Validation Methods):**
  - **验证集方法 (Validation Set Approach):** 将数据分成训练集和验证集。
    - "The **validation set approach** is the simplest method. It randomly splits the available dataset into a **training** and a **validation** (or hold-out) dataset." (**验证集方法**是最简单的方法。它将可用数据集随机分成**训练**和**验证**（或留出）数据集。)
  - **留一交叉验证 (Leave-One-Out Cross-Validation (LOOCV)):** 验证集仅包含一个样本，训练 N 次。
    - "**Leave-one-out cross-validation (LOOCV)**: This method also splits the available dataset into training and validation sets. However, the **validation set contains only one sample**" (**留一交叉验证 (LOOCV)**: 此方法也将可用数据集分成训练集和验证集。然而，**验证集只包含一个样本**。)
  - **k 折交叉验证 (k-Fold Cross-Validation):** 将数据分成 k 组，训练 k 次，每次使用一个组作为验证。
    - "***k*-fold cross-validation**: In this approach the **available dataset is divided into *k* groups** (also known as folds) of approximately equal size." (***k* 折交叉验证**：在这种方法中，**可用数据集被分成 *k* 个组**（也称为折叠）大小大致相等。)

III. 优化 (Optimization)

1. **优化理论 (Optimization Theory):**
   - **候选模型 (Candidate Models):** 一组潜在的模型。
   - **质量度量 (Quality Metric):** 模型性能的度量（例如，误差）。
   - **最优模型 (Optimal Model):** 在目标总体上具有最高质量的模型。
   - "Optimisation allows us to identify among all the candidate models the one that achieves the highest quality on the target population, i.e. the **optimal model**." (优化使我们能够从所有候选模型中找出在目标人群中达到最高质量的模型，即**最优模型**。)
2. **误差曲面 (Error Surface):**
   - **定义 (Definition):** 将每个*候选模型* w映射到其*误差*的函数 $E(w)$。
     - "The **error surface** (a.k.a. error, objective, loss or cost function) denoted by $E(w)$ maps each **candidate model** w to its **error**." (**误差曲面**（也称为误差、目标、损失或成本函数）用 $E(w)$ 表示，将每个**候选模型** w 映射到其**误差**。)
   - **梯度 (Gradient):** 误差曲面的斜率，$\nabla E(w)$。
     - "The **gradient** (slope) of the error surface, $\nabla E(w)$, is zero at the optimal model." (误差曲面的**梯度**（斜率）$\nabla E(w)$ 在最优点处为零。)
   - **梯度下降 (Gradient Descent):** 一种使用梯度迭代更新模型的优化方法。
     - "**Gradient descent** is a numerical optimisation method where we **update iteratively** our model using the gradient of the error surface." (**梯度下降**是一种数值优化方法，我们使用误差曲面的梯度**迭代更新**我们的模型。)
     - **学习率 (Learning Rate):** 控制更新步长的参数 (ε)。
       - "where ε is known as the **learning rate** or **step size**." (其中 ε 被称为**学习率**或**步长**。)
     - **参数调整 (Parameter Tuning):** 在优化期间调整模型参数的过程。
       - "With every iteration we adjust the parameters w of our model. This is why this process is also known as **parameter tuning**." (在每次迭代中，我们都会调整模型的参数 w。这就是为什么这个过程也被称为**参数调整**。)
3. **优化中的挑战 (Challenges in Optimization):**
   - **计算成本 (Computational Cost):** 评估每个候选模型可能很昂贵。
   - **局部最优 (Local Optima):** 梯度下降会陷入局部最优。
     - "**Local optima** (model with the lowest error within a region)." (**局部最优**（在某个区域内误差最小的模型）。)
     - "**Global optima** (model with the lowest error among all the models)." (**全局最优**（在所有模型中误差最小的模型）。)
   - **过拟合 (Overfitting):** 模型在训练数据上表现良好，但在未见数据上表现不佳。
     - "**Overfitting** and fooling ourselves: When **small datasets** and **complex models** are used, the differences between the two can be very large, resulting in trained models that work very well for the empirical error surface but very poorly for the true error surface." (**过拟合**和自欺欺人：当使用**小数据集**和**复杂模型**时，两者之间的差异可能非常大，导致训练好的模型在经验误差面上效果很好，但在真实误差面上效果很差。)
4. **正则化 (Regularization):**
   - **目的 (Purpose):** 约束模型参数并防止过拟合。
     - "**Regularisation** modifies the empirical error surface by adding a term that constrains the values that the model parameters can take on." (**正则化**通过添加一个项来修改经验误差曲面，该项约束模型参数可以取的值。)
   - **正则化误差曲面 (Regularized Error Surface):** $ER(w) = E(w) + \lambda wTw$
   - **成本与质量 (Cost vs. Quality):** 训练期间的质量概念（成本或目标函数）可能与目标质量指标不同。

        - "We usually call our notion of quality during training **cost** or **objective function**, to distinguish it from the **target quality metric**." (我们通常将训练期间的质量概念称为**成本**或**目标函数**，以将其与**目标质量指标**区分开来。)

IV. 正确使用数据 (Using Data Correctly)

1. **数据角色 (Data Roles):** 测试、训练和验证任务都涉及数据。
2. **数据集创建 (Dataset Creation):** 首先考虑任务，然后创建数据集。
3. **关键原则 (Key Principles):**
    - **代表性 (Representativeness):** 数据集应代表目标总体。
    - **独立性 (Independence):** 应独立提取样本。
    - **训练前设计 (Design Before Training):** 测试策略应在训练前设计。
    - **随机性 (Randomness):** 性能估计是随机量。
    - **模型质量 (Model Quality):** 取决于模型类型、优化策略和数据代表性。

Part II: Methodology II

I. 超越基本模型：流水线 (Beyond Basic Models: Pipelines)

1. **流水线 (Pipeline):** 将输入数据转换为输出预测的操作序列。
    - "The term **pipeline** describes a **sequence of operations**." (术语**流水线**描述了**操作序列**。)
    - "Note that the term **pipeline** is often used to describe **workflows**. We use the term workflow to describe a **sequence of steps** that we take, e.g. we first formulate a problem, then collect data, select a model, train it..." (请注意，术语**流水线**通常用于描述**工作流**。我们使用术语“工作流”来描述我们采取的**一系列步骤**，例如，我们首先制定一个问题，然后收集数据，选择一个模型，训练它.....)
2. **流水线的组成部分 (Components of a Pipeline):**
    - **转换阶段 (Transformation Stages):** 处理输入数据。
    - **机器学习模型 (Machine Learning Models):** 可以并行运行。
    - **聚合阶段 (Aggregation Stage):** 组合各个输出。
3. **训练和部署 (Training and Deployment):**
    - **手工制作与训练 (Hand-crafted vs. Trained):** 阶段可以手动设计或从数据中学习。
    - **固定参数 (Fixed Parameters):** 训练后，流水线参数保持固定。
    - **部署 (Deployment):** 部署整个流水线，而不仅仅是单个模型。

II. 数据归一化 (Data Normalization)

1. **目的 (Purpose):** 缩放属性，使其值落在相似的范围内。
    - "**Data normalisation** is a **tunable transformation** stage that allows us to scale attributes so that their values belong to similar ranges." (**数据归一化**是一个**可调的转换**阶段，它允许我们缩放属性，使其值属于相似的范围。)
2. **方法 (Methods):**
    - **最小-最大归一化 (Min-Max Normalization):** 将值缩放到 [0, 1] 范围内。
        - "**Min-max normalisation** produces values within the same continuous range [0, 1], i.e. greater (or equal) than 0 and less (or equal) than 1." (**最小-最大归一化**产生相同连续范围 [0, 1] 内的值，即大于（或等于）0 且小于（或等于）1。)
        - $z = (x - min(x)) / (max(x) - min(x))$
    - **标准化 (Standardization):** 将数据转换为均值为 0，标准差为 1。
        - "**Standardisation** is a common procedure defined by the transformation $z = (x - \mu) / \sigma$" (**标准化**是由变换 $z = (x - \mu) / \sigma$ 定义的常见过程。)
3. **注意事项 (Considerations):**
    - **超出范围的值 (Out-of-Range Values):** 预期部署期间会出现超出训练范围的值。
    - **异常值 (Outliers):** 会对归一化产生负面影响。
    - **非线性缩放 (Non-linear Scaling):** 存在 softmax 或对数缩放等选项。
    - **影响和失真 (Effects and Distortions):** 考虑缩放的意外后果。

III. 数据转换 (Data Transformations)

1. **定义 (Definition):** 改变样本表示形式的数据操作，在空间之间移动样本。
    - "**Transformations** are data manipulations that change the way that we represent our samples. They can be seen as moving samples from one space to another." (**转换**是改变我们表示样本方式的数据操作。它们可以看作是将样本从一个空间移动到另一个空间。)
2. **类型 (Types):**
    - **线性变换 (Linear Transformations):** 可以看作是旋转和缩放（例如，PCA）。
        - "**A linear transformations** can be seen as a rotation and scaling." (可以将**线性变换**视为旋转和缩放。)

- **非线性变换 (Non-linear Transformations):** 没有唯一的描述。
- **降维 (Dimensionality Reduction):** 目标空间具有较少的维度。
  - "If the destination space has fewer dimensions than the original one, we talk about **dimensionality reduction**." (如果目标空间的维数少于原始空间，我们就说**降维**。)
  - **特征选择 (Feature Selection):** 目标空间由原始属性的子集定义。
    - "In particular, after **feature selection** the destination space is defined by a subset of the original attributes." (特别是，在**特征选择**之后，目标空间由原始属性的子集定义。)
  - **特征提取 (Feature Extraction):** 新属性是从原始属性的操作中派生出来的。
    - "In **feature extraction** the new attributes are defined as operations on the original attributes. Common when using complex types." (在**特征提取**中，新属性被定义为对原始属性的操作。使用复杂类型时很常见。)

3. **示例 (Examples):**
   - **主成分分析 (Principal Components Analysis (PCA)):** 识别样本排列的方向并为成分分配分数。
     - "**Principal components analysis (PCA)** identifies the **directions** along which samples are aligned." (主成分分析 (PCA)** 确定样本排列的**方向**。)
   - **非线性映射 (Non-linear Mapping):** 转换到基于到中心的距离的空间。

IV. 特征选择 (Feature Selection)

1. **目标 (Goal):** 通过选择属性的相关子集来降低维度。
2. **方法 (Approaches):**
   - **过滤 (Filtering):** 使用模型和验证性能单独评估每个属性。
     - "The simplest approach to feature selection is to consider each attribute individually. A score can be assigned by fitting a model and obtaining its validation performance." (特征选择的最简单方法是单独考虑每个属性。可以通过拟合模型并获得其验证性能来分配分数。)
   - **封装 (Wrapping):** 一起评估属性子集，考虑交互。
     - "**Wrapping** approaches consider possible interaction between predictors by:" (**封装**方法通过以下方式考虑预测变量之间可能的相互作用：)

V. 特征提取 (Feature Extraction)

1. **目的 (Purpose):** 通过创建一组较小的信息特征来降低维度。
2. **技术 (Techniques):** 数字信号和图像处理提供了广泛的特征提取方法。

VI. 集成方法 (Ensembles)

1. **概念 (Concept):** 组合多个*基模型*以提高性能。
   - "**Ensemble** methods allow us to create a new model that **combines** the strengths of **base models**." (**集成**方法允许我们创建一个新模型，该模型**结合**了**基模型**的优点。)
   - "Base models need to be as **diverse** as possible and can be created by training:" (基模型需要尽可能**多样化**，并且可以通过训练来创建：)
2. **方法 (Methods):**
   - **装袋法 (Bagging):** 使用自举采样在数据的不同子集上训练多个模型。
     - "Given a training dataset, **bagging** generates K sub-datasets by bootstrapping and trains K simple base models with each sub-dataset." (给定一个训练数据集，**装袋法**通过自举生成 K 个子数据集，并使用每个子数据集训练 K 个简单的基模型。)
   - **随机森林 (Random Forests):** 决策树的集合，同时随机化样本和预测变量。
     - "**Random forests** are an ensemble of decision trees." (**随机森林**是决策树的集合。)
   - **提升法 (Boosting):** 训练一系列模型，每个模型都关注先前模型错误分类的样本。
     - "**Boosting** follows a different approach: it generates a **sequence of simple base models**, where each successive model focuses on the samples that the **previous models could not handle** properly." (**提升法**遵循不同的方法：它生成**一系列简单的基模型**，其中每个后续模型都关注**先前模型无法正确处理**的样本。)
3. **决策树 (Decision Trees):**
   - **分区 (Partitioning):** 使用树状结构将预测变量空间划分为决策区域。
     - "**Decision tree** classifiers partition the predictor space into multiple decision regions by implementing sequences of splitting rules using one **predictor** only." (**决策树**分类器通过仅使用一个**预测变量**实现分裂规则序列，将预测变量空间划分为多个决策区域。)
   - **目标 (Goal):** 创建主要包含来自同一类别的样本的*纯叶子*。
   - **分类 (Classification):** 将样本分配给其叶子中的*多数类*。
   - **递归构建 (Recursive Construction):** 通过分裂区域递归地构建树。

VII. 关键要点 (Key Takeaways)

1. **流水线 (Pipelines):** 部署操作序列，而不仅仅是模型。
2. **转换 (Transformations):** 使用适当的数据表示形式。
3. **集成 (Ensembles):** 组合模型以提高性能。
4. **数据至关重要 (Data is Crucial):** 首先设计任务，然后创建数据集。
5. **仔细评估 (Careful Evaluation):** 考虑随机性和潜在的陷阱。

These detailed study notes, now with comprehensive Chinese explanations, should provide a solid foundation for your understanding of the material. Remember to refer to the original PDFs for further details and examples. Good luck with your studies!