Lecture1：机器学习介绍 machine learning

一、监督学习（supervised learning）

监督学习是一种机器学习的方法，其目标是根据已知的输入-输出对（训练样本）来建立一个预测模型，以便对新的输入进行预测或分类。在监督学习中，我们通常有一个标记的训练数据集，其中每个样本都有对应的输入特征和已知的输出标签。（Supervised learning is a method of machine learning whose goal is to build a prediction model based on known input-output pairs (training samples) in order to predict or classify new inputs. In supervised learning, we usually have a labeled training data set in which each sample has a corresponding input feature and a known output label.）

监督学习可以分为两类问题：回归问题和分类问题。

在回归问题中，我们的目标是预测连续的输出变量。例如，根据房屋的各种特征（如面积、卧室数量等），预测房屋的价格。回归模型通常使用线性回归、多项式回归、支持向量回归等算法。

在分类问题中，我们的目标是将输入样本分为不同的类别。例如，根据花瓣长度和宽度等特征，将花朵分为不同的品种。分类模型通常使用逻辑回归、决策树、支持向量机、神经网络等算法。

监督学习的基本步骤包括数据准备、特征选择和提取、模型选择和训练、模型评估和预测。在数据准备阶段，我们收集和清洗数据，并将其划分为训练集和测试集。然后，我们选择合适的特征，并进行必要的预处理，如标准化、归一化等。接下来，我们选择适当的模型，并使用训练集对模型进行训练。训练完成后，我们使用测试集评估模型的性能，并进行必要的调整和优化。最后，我们可以使用训练好的模型对新的未知样本进行预测或分类。（The basic steps of supervised learning include data preparation, feature selection and extraction, model selection and training, model evaluation and prediction. In the data preparation phase, we collect and clean the data and divide it into training sets and test sets. Then, we select suitable features and carry out necessary preprocessing, such as standardization, normalization, etc. Next, we select the appropriate model and train the model using the training set. After the training is complete, we use the test set to evaluate the performance of the model and make necessary adjustments and optimizations. Finally, we can use trained models to predict or classify new unknown samples.）

二、无监督学习（Unsupervised learning）

无监督学习是一种机器学习的方法，其目标是从未标记的数据中发现数据的内在结构、模式或关系，而不需要已知的输出标签。与监督学习不同，无监督学习的训练数据集没有标签或类别信息，模型需要自行探索数据的特征和结构。

无监督学习可以分为聚类和降维两类问题。（Unsupervised learning is an approach to machine learning where the goal is to discover the underlying structure, pattern, or relationship of data in unlabeled data without the need for a known output label. Different from supervised learning, the training data set of unsupervised learning has no label or category information, and the model needs to explore the characteristics and structure of the data by itself. Unsupervised learning can be divided into two categories, clustering and dimensionality reduction problem.）

在聚类问题中，无监督学习的目标是将相似的样本分组到同一簇中，而将不相似的样本分开。聚类算法通过测量样本之间的相似性或距离，将数据分为不同的群组。常见的聚类算法包括 K 均值聚类、层次聚类、DBSCAN 等。

在降维问题中，无监督学习的目标是减少数据的维度，保留数据的主要信息。降维算法通过将高维数据映射到低维空间，可以帮助我们理解数据的结构、可视化数据以及减少计算复杂度。常见的降维算法包括主成分分析（PCA）、独立成分分析（ICA）、t-SNE 等。

无监督学习的优点在于它可以在没有标签信息的情况下对数据进行探索和分析，帮助我们发现数据中的隐藏模式和结构。它也可以用于预处理数据，减少噪声和冗余，提高后续监督学习的性能。无监督学习的挑战在于缺乏明确的目标函数或评估指标，模型的性能评估相对困难。（The advantage of unsupervised learning is that it can explore and analyze data without label information, helping us to discover hidden patterns and structures in the data. It can also be used to preprocess data, reduce noise and redundancy, and improve the performance of subsequent supervised learning. The challenge of unsupervised learning lies in the lack of explicit objective function or evaluation index, and the performance evaluation of the model is relatively difficult.）

无监督学习在许多领域有广泛的应用，如聚类分析、异常检测、推荐系统、图像分割等。它可以帮助我们从大量未标记的数据中提取有用的信息，并为后续的分析和决策提供支持。

三、半监督学习

半监督学习是一种介于监督学习和无监督学习之间的机器学习方法。在半监督学习中，训练数据集中只有一部分样本带有标签（有监督信息），而剩下的样本没有标签（无监督信息）。半监督学习的目标是利用未标记样本的信息来增强监督学习的性能，从而提高模型的准确性和泛化能力。(Semi-supervised learning is a machine learning method between supervised learning and unsupervised learning. In semi-supervised learning, only a part of the samples in the training data set are labeled (supervised information), while the rest are unlabeled (unsupervised information). The goal of semi-supervised learning is to use information from unlabeled samples to enhance the performance of supervised learning, thereby improving the accuracy and generalization ability of the model.)

半监督学习的核心思想是通过未标记样本的分布和相似性来辅助标记样本的学习过程。未标记样本可以提供更多的数据信息，有助于更好地捕捉数据的潜在结构和模式。半监督学习方法通常利用无监督学习算法对未标记样本进行聚类、降维或生成模型，然后将这些无监督学习的结果与带有标签的样本结合起来进行训练。(The core idea of semi-supervised learning is to assist the labelled sample learning process through the distribution and similarity of unlabelled samples. Unlabeled samples can provide more information about the data and help better capture the underlying structure and pattern of the data. Semi-supervised learning methods usually use unsupervised learning algorithms to cluster, reduce dimension or generate models for unlabeled samples, and then combine the results of unsupervised learning with labeled samples for training.)

半监督学习的优点在于可以充分利用未标记样本的信息，减少对标记样本的依赖。这对于数据集中标记样本较少的情况非常有用，因为标记样本的获取通常需要人工标注，耗时耗力。半监督学习可以通过使用大量的未标记样本来扩展训练数据集，从而提高模型的性能。(The advantage of semi-supervised learning is that it can make full use of the information of unlabeled samples and reduce the dependence on labeled samples. This is very useful when there are few marked samples in the data set, because the acquisition of marked samples usually requires manual marking, which is time-consuming and labor-intensive. Semi-supervised learning can improve the performance of the model by extending the training data set with a large number of unlabeled samples.)

然而，半监督学习也面临一些挑战。其中一个挑战是如何有效地利用未标记样本的信息，以提高模型的泛化能力。

四、强化学习 reinforcement learning

强化学习是一种机器学习方法，旨在让智能体（agent）通过与环境的交互学习如何做出最优的决策。在强化学习中，智能体通过观察环境的状态，执行特定的动作，接收环境的奖励或惩罚，并根据这些奖励或惩罚来调整其行为，以达到最大化长期累积奖励的目标。(Reinforcement learning is a machine learning method that aims to let agents learn how to make optimal decisions by interacting with the environment. In reinforcement learning, agents observe the state of the environment, perform specific actions, receive rewards or punishments from the environment, and adjust their behaviors according to these rewards or punishments in order to achieve the goal of maximizing long-term cumulative rewards.)

强化学习的关键要素包括：

1. 环境：智能体所处的外部环境，可以是真实的物理环境或虚拟的模拟环境。
2. 状态（State）：描述环境的特定瞬时情况，用于决策和行动选择。
3. 动作（Action）：智能体基于状态所采取的行动或决策。
4. 奖励（Reward）：环境根据智能体的行动给予的反馈信号，用于评估行动的好坏。
5. 策略（Policy）：智能体在给定状态下选择行动的策略，可以是确定性的或概率性的。
6. 值函数 (Value Function)：用于评估状态或状态-动作对的长期累积奖励的函数，可以指导智能体的决策过程。
7. 学习算法：智能体根据环境的反馈信号和奖励来更新策略和值函数，从而逐步改善其决策能力。

强化学习的目标是通过与环境的交互学习一个最优的策略，使智能体能够在不断变化的环境中做出最佳的决策，以最大化累积奖励。强化学习在许多领域具有广泛的应用，包括机器人控制、游戏策略、自动驾驶等。

五、机器学习类型

批量学习（Batch Learning）和在线学习（Online Learning）是机器学习中两种不同的学习方式。

1. 批量学习（Batch Learning）：批量学习是一种离线学习的方式，它在训练阶段一次性使用所有可用的训练数据

来训练模型。在批量学习中，整个训练数据集被划分为若干批次（batch），每个批次中的样本同时输入模型进行训练，并通过计算损失函数来更新模型的参数。批量学习通常需要较大的内存和计算资源，适用于数据规模较小且可以一次性加载到内存中的情况。（Batch Learning: Batch learning is an off-line learning method that uses all available training data at once to train the model during the training phase. In batch learning, the whole training data set is divided into several batches. The samples in each batch are input into the model for training at the same time, and the parameters of the model are updated by calculating the loss function. Batch learning usually requires large memory and computing resources, and is suitable for small amounts of data that can be loaded into the memory at one time.）

2. 在线学习（Online Learning）：在线学习是一种增量学习的方式，它可以逐步地从流式数据中进行学习和更新模型。在线学习适用于数据以流的形式产生，并且可能随着时间的推移发生变化的情况。在在线学习中，模型会逐个接收样本并进行预测，然后根据实际的预测结果来更新模型的参数。在线学习具有实时性和适应性的优势，可以灵活地处理大规模和不断变化的数据。（Online Learning: Online learning is a way of incremental learning, which can gradually learn and update models from streaming data. Online learning is suitable for situations where data is produced in streams and may change over time. In online learning, the model will receive samples one by one and make predictions, and then update the parameters of the model according to the actual prediction results. Online learning has the advantage of timeliness and adaptability, and can be flexible in dealing with large-scale and constantly changing data.）

批量学习和在线学习各有其优缺点，选择哪种学习方式取决于具体的应用场景和需求。批量学习适用于静态数据集和离线分析的任务，能够全面利用所有训练数据进行训练，但对资源要求较高。在线学习适用于动态数据和实时决策的任务，可以随着数据的到达进行增量更新，但可能对历史数据的遗忘和训练的稳定性有一定影响。在实际应用中，可以根据数据的特点和需求的实时性来选择合适的学习方式。

基于实例的学习（Instance-based Learning）和基于模型的学习（Model-based Learning）是机器学习中两种不同的学习方法。

1. 基于实例的学习：基于实例的学习也称为记忆学习（Memory-based Learning）或懒惰学习（Lazy Learning），它通过将训练数据保存起来，在预测时直接使用与测试实例最相似的训练实例进行预测。基于实例的学习不需要显式地构建模型，而是将训练数据作为记忆存储，并通过计算相似度度量（如欧氏距离、余弦相似度等）来找到最相似的训练实例。基于实例的学习在训练阶段不进行显式的模型构建和参数优化，而是将实例存储起来，因此在预测阶段速度较快。典型的基于实例的学习算法包括K最近邻算法（K-Nearest Neighbors, KNN）。（Instance-based Learning: Instance-based Learning, also known as Memory-based learning or Lazy Learning, saves training data and directly uses the training instance that is most similar to the test instance for prediction. Instance-based learning does not require explicitly building a model, but takes training data as memory storage and finds the most similar training instance by calculating similarity measures (such as Euclidean distance, cosine similarity, etc.). Case-based learning does not carry out explicit model construction and parameter optimization in the training stage, but stores the instance, so it is faster in the prediction stage. Typical instance-based learning algorithms include K-Nearest Neighbors (KNN).）
2. 基于模型的学习：基于模型的学习是通过构建一个显式的模型来学习和预测。在基于模型的学习中，训练数据被用来训练模型的参数或结构，从而得到一个能够对新数据进行预测的模型。基于模型的学习可以根据具体任务选择不同的模型，如线性回归、决策树、支持向量机等。在训练阶段，基于模型的学习会使用训练数据来拟合模型，并进行参数优化，以使模型能够最好地适应训练数据。在预测阶段，基于模型的学习会使用学习到的模型对新的输入数据进行预测。（Model-based learning: Model-based learning is learning and prediction by constructing an explicit model. In model-based learning, training data is used to train the parameters or structure of the model, resulting in a model capable of making predictions about new data. Model-based learning can choose different models according to specific tasks, such as linear regression, decision tree, support

vector machine, etc. In the training phase, model-based learning will use training data to fit the model and optimize parameters so that the model can best adapt to the training data. In the prediction phase, model-based learning uses the learned model to predict new input data.)

基于实例的学习和基于模型的学习各有其特点和适用场景。基于实例的学习适用于数据较小且特征维度较高的情况，能够灵活地适应训练数据的分布，但可能对噪声和冗余数据敏感。基于模型的学习适用于大规模数据和复杂的模式识别问题，能够通过模型的泛化能力进行预测，但可能需要更多的计算资源和时间来构建和优化模型。在实际应用中，可以根据数据集的规模、问题的复杂性和需求的实时性来选择合适的学习方法。

Lecture2：数据处理 processing

一、数据相关性

数据相关性指的是两个或多个变量之间的关联程度。在数据分析和机器学习中，了解变量之间的相关性对于理解数据的特征和进行预测非常重要。

相关性可以通过相关系数来衡量，最常用的是皮尔逊相关系数。皮尔逊相关系数的取值范围在-1 到 1 之间，表示变量之间的线性关系强度和方向。具体而言：

- 当相关系数为 1 时，表示两个变量完全正相关，即一个变量增加时，另一个变量也相应增加。
- 当相关系数为-1 时，表示两个变量完全负相关，即一个变量增加时，另一个变量减少。
- 当相关系数接近于 0 时，表示两个变量之间没有线性关系。

除了皮尔逊相关系数，还有其他的相关系数可用于衡量非线性相关性，例如斯皮尔曼相关系数和刻尔吉斯塔德相关系数。

通过计算变量之间的相关系数，可以获得以下信息：

1. 正相关和负相关：相关系数的符号表示变量之间的方向关系，正相关表示变量随着另一个变量的增加而增加，负相关表示变量随着另一个变量的增加而减少。
2. 强弱关系：相关系数的绝对值大小表示变量之间的关联程度，绝对值越接近 1，关联程度越强。
3. 线性关系：相关系数衡量的是线性关系，即变量之间的直线关系。如果变量之间存在非线性关系，则相关系数可能无法准确反映其关联程度。

了解数据的相关性可以帮助我们理解变量之间的关系，选择合适的特征进行建模和预测，并进行特征选择、降维或变量变换等数据处理操作。

二、数据预处理 data pre-processing

数据预处理是指在应用机器学习算法之前对原始数据进行转换、清洗和规范化的过程。数据预处理的目的是为了提高数据质量、减少噪声和异常值的影响，以及使数据适应机器学习算法的需求。（Data preprocessing refers to the process of converting, cleaning, and normalizing raw data before applying machine learning algorithms. The purpose of data preprocessing is to improve data quality, reduce the impact of noise and outliers, and adapt data to the needs of machine learning algorithms.）

数据预处理通常包括以下步骤：

1. 数据清洗：这一步骤主要处理数据中的缺失值、异常值和重复值。缺失值可以通过填充、删除或插值等方法进行处理。异常值可以通过统计方法或基于领域知识进行识别和处理。重复值可以通过去重操作进行处理。（Data cleansing: This step deals with missing values, outliers, and duplicate values in the data. Missing values can be processed by filling, deleting, or interpolating. Outliers can be identified and processed by statistical methods or based on domain knowledge. Duplicate values can be handled by a deduplication operation.）
2. 特征选择：特征选择是从原始数据中选择最具有信息量和预测能力的特征。可以通过统计分析、相关性分析、特征重要性评估等方法进行特征选择。（Feature selection: Feature selection is to select the most informative and predictive features from the original data. Feature selection can be carried out by statistical analysis, correlation analysis, feature importance assessment and other methods.）
3. 特征变换：特征变换是对原始特征进行变换或组合，以提取更有用的特征表示。常见的特征变换方法包括标准化（使特征具有相同的尺度）、归一化（将特征缩放到一定的范围）、离散化（将连续特征转换为离散值）等。（Feature transformation: Feature transformation is the transformation or combination of original features to extract a more useful feature representation. Common feature transformation methods include normalization (making the features have the same scale), normalization (scaling the features to a certain range), discretization (converting continuous features to discrete values), etc.）
4. 数据集划分：将原始数据集划分为训练集和测试集。训练集用于模型的训练和参数调优，测试集用于评估模型的性能和泛化能力。（Data set partitioning: Divides the original data set into training set and test set. The training set is used for the training and parameter tuning of the model, and the test set is used to evaluate the performance and generalization ability of the model.）

5. 数据集平衡：对于分类问题中存在类别不平衡的情况，可以采取欠采样、过采样或合成新样本等方法来平衡数据集，以避免模型对多数类样本过度拟合。（Data set balance: In case of category imbalance in classification problems, undersampling, oversampling or new sample synthesis can be adopted to balance the data set, so as to avoid over-fitting of the model to most class samples.）
6. 数据编码：将非数值型数据（如类别型数据）转换为数值型数据，以便机器学习算法的处理。常用的编码方法包括独热编码、标签编码等。（Data coding: Converting non-numerical data (such as categorical data) into numerical data for processing by machine learning algorithms. Common coding methods include thermal coding, tag coding and so on.）
7. 数据降维：对于高维数据，可以通过降维方法如主成分分析（PCA）、线性判别分析（LDA）等将其映射到低维空间，以减少特征维度和计算复杂度。（Data dimension reduction: For high-dimensional data, dimensionality reduction methods such as principal component analysis (PCA) and linear discriminant analysis (LDA) can be used to map it to low-dimensional space to reduce the feature dimension and computational complexity.）

数据预处理的目标是为了准备好适合机器学习算法使用的数据，使其更易于理解和分析，提高机器学习模型的性能和泛化能力。数据预处理过程需要根据具体问题和数据特点进行调整和选择，以确保得到高质量的输入数据。（The goal of data preprocessing is to prepare data suitable for use by machine learning algorithms, make it easier to understand and analyze, and improve the performance and generalization ability of machine learning models. The data preprocessing process needs to be adjusted and selected according to specific problems and data characteristics to ensure high quality input data.）

三、标准化 normalization

1. Z-Score 标准化（零均值标准化）：
    - 对于特征 x，标准化后的数值为：(x - mean) / std
    - 其中，mean 为特征的均值，std 为特征的标准差。
2. 最小-最大标准化：
    - 对于特征 x，标准化后的数值为：(x - min) / (max - min)
    - 其中，min 为特征的最小值，max 为特征的最大值。
3. 小数定标标准化：
    - 对于特征 x，标准化后的数值为：x / (10^j)
    - 其中，j 为使得最大绝对值小于 1 的数量级。
4. 归一化：
    - 对于样本 x，标准化后的数值为：x / ||x||
    - 其中，||x||为样本 x 的范数。

四、特征选择和特征提取

特征选择和特征提取都是在机器学习和数据挖掘任务中对原始特征进行预处理的方法，旨在减少特征维度、提高模型效果或降低计算成本。

特征选择（Feature Selection）是指从原始特征中选择一部分最具有代表性和重要性的特征，而舍弃其他特征。常见的特征选择方法包括过滤式方法、包裹式方法和嵌入式方法。过滤式方法通过计算特征与目标变量之间的相关性或统计指标来评估特征的重要性，并选择具有高分值的特征。包裹式方法使用特定的机器学习模型来评估特征的重要性，通过逐步选择或消除特征来得到最佳子集。嵌入式方法将特征选择与模型训练过程结合起来，通过正则化或惩罚项来约束模型的复杂度，从而自动选择重要的特征。（Feature Selection refers to selecting some of the most representative and important features from the original features, while abandoning other features. Common feature selection methods include filter method, wrap method and embedded method. Filtering method evaluates the importance of features by calculating the correlation or statistical index between features and target variables, and selects features with high scores. The wrapped approach uses a specific machine learning model to assess the importance of features and to get the best subset by progressively selecting or eliminating features. The embedded method combines feature selection with model training

process, constrains the complexity of the model through regularization or penalty terms, and then automatically selects important features.)

特征提取（Feature Extraction）是指从原始特征中抽取新的、更有意义的特征表示。常见的特征提取方法包括主成分分析（PCA）、线性判别分析（LDA）、非负矩阵分解（NMF）等。这些方法通过将原始特征进行变换或组合，提取出能够更好地表示数据分布和区分不同类别的新特征。(Feature Extraction refers to extracting new and more meaningful feature representations from original features. Common feature extraction methods include principal component analysis (PCA), linear discriminant analysis (LDA), and non-negative matrix decomposition (NMF). By transforming or combining original features, these methods extract new features that can better represent data distribution and distinguish different categories.)

Lecture3：分类

一、二元分类器 binary classifier

二元分类器是一种用于将样本分为两个类别的模型。它通常用于解决二元分类问题，即将输入样本划分为正类和负类两个类别。（A binary classifier is a model used to divide a sample into two categories. It is usually used to solve the binary classification problem, that is, the input sample is divided into positive and negative categories.）

常见的二元分类器包括逻辑回归、支持向量机（SVM）、朴素贝叶斯分类器等。这些分类器使用不同的算法和方法来建立分类模型，以预测样本属于哪个类别。（Common binary classifiers include logistic regression, support vector machine (SVM), naive Bayes classifier and so on. These classifiers use different algorithms and methods to build classification models to predict which category a sample belongs to.）

在二元分类任务中，通常会将其中一个类别定义为正类（如标记为 1），另一个类别定义为负类（如标记为 0）。分类器的目标是根据输入样本的特征，将其正确地分为正类或负类，并进行预测。（In a binary classification task, it is common to define one of the categories as a positive class (such as marked 1) and the other as a negative class (such as marked 0). The goal of a classifier is to correctly classify input samples into positive or negative categories based on their characteristics and make predictions.）

二元分类器的训练过程通常涉及选择适当的特征表示、选择合适的模型和算法、定义损失函数并进行参数优化等步骤。通过训练，分类器会学习到样本的特征与类别之间的关系，从而能够对新的未见样本进行分类预测。

二元分类器在许多领域中都有广泛的应用，例如垃圾邮件过滤、疾病诊断、欺诈检测等。根据具体的问题和数据特点，选择适合的二元分类器和相应的算法是非常重要的。

二、模型的评估 performance measures

1. 混淆矩阵（ Confusion matrix）

混淆矩阵是用于评估分类模型性能的一种表格形式的工具。它对分类模型的预测结果和真实类别进行了统计，以便分析模型的准确性、召回率、精确度和特异度等指标。

混淆矩阵的表格通常是一个 2x2 的矩阵，其中行表示真实类别，列表示模型预测的类别。矩阵的四个单元格包括了四种不同的分类情况：

- 真正例（True Positive，TP）：模型将正类样本预测为正类。
- 假正例（False Positive，FP）：模型将负类样本预测为正类。
- 假反例（False Negative，FN）：模型将正类样本预测为负类。
- 真反例（True Negative，TN）：模型将负类样本预测为负类。

通过统计这些分类情况，可以计算出一系列分类指标，例如准确率（Accuracy）、召回率（Recall）、精确度（Precision）、特异度（Specificity）等。这些指标可以帮助我们评估模型的性能和判断分类器在不同类别上的表现。

混淆矩阵在机器学习中广泛应用于二元分类问题，也可以扩展到多类别分类问题。它提供了对分类模型进行全面评估的信息，帮助我们了解模型在不同类别上的表现，并作出相应的调整和改进。

2. 评估二元分类模型的标准

- 召回率（Recall）：也称为灵敏度（Sensitivity）或真正例率（True Positive Rate），表示模型正确预测为正类的样本数占所有正类样本数的比例。计算公式为 Recall = TP / (TP + FN)。
- 精确度（Precision）：表示模型预测为正类的样本中真正为正类的比例。计算公式为 Precision = TP / (TP + FP)。
- F1 分数（F1 Score）：综合考虑了召回率和精确度的指标，是召回率和精确度的调和平均值。F1 分数可以平衡召回率和精确度之间的权衡，对不均衡数据集或关注错误分类的情况有用。计算公式为 F1 Score = 2 * (Precision * Recall) / (Precision + Recall)。
- Fβ 分数（Fβ Score）：是 F1 分数的一般化形式，通过引入一个参数 β 来调整召回率和精确度之间的权衡。当 β=1 时，即为 F1 分数；当 β>1 时，更加重视召回率；当 β<1 时，更加重视精确度。计算公式为 Fβ Score = (1 + β^2) * (Precision * Recall) / (β^2 * Precision + Recall)。
- 特异度（Specificity）：表示模型正确预测为负类的样本数占所有负类样本数的比例。计算公式为 Specificity = TN / (TN + FP)。

- 虚警率（Fall-out）：也称为误报率（False Positive Rate），表示模型错误地将负类样本预测为正类的比例。计算公式为 Fall-out = 1 – Specificity= FP / (TN + FP)。
- 漏报率（Miss）：也称为错误率（False Negative Rate），表示模型错误地将正类样本预测为负类的比例。计算公式为 Miss = 1 – Recall= FN / (TP + FN)。

3. 阈值的选择 ROC 与 AUC

阈值的选择在二元分类中非常重要，它决定了将模型的输出预测为正例还是负例的界限。选择适当的阈值可以平衡模型的召回率和精确率，以及根据具体应用场景的需求进行调整。

一般来说，阈值的选择取决于具体的优化目标。以下是几种常见的阈值选择策略：

1. 默认阈值：对于大多数二元分类模型，默认阈值通常为 0.5。如果模型的输出概率大于 0.5，则将其预测为正例，否则预测为负例。

2. 受限阈值：根据特定的业务需求和问题场景，可以将阈值设置为其他值，例如 0.3 或 0.7，以便更注重召回率或精确率。

3. ROC 曲线和 AUC：通过绘制 ROC 曲线并计算 AUC，可以选择在曲线上最接近理想点（0,1）的阈值作为最佳阈值。

4. 成本敏感阈值：如果模型对于误报和漏报有不同的成本或代价，可以根据成本敏感性选择阈值。例如，对于医疗诊断中的异常检测，漏报（将异常样本误判为正常）的成本可能更高，因此可以选择较低的阈值来提高召回率。

5. 动态阈值：根据不同样本的特征和上下文，使用动态阈值进行预测。例如，根据样本的重要性、先验知识或其他相关因素来调整阈值。

需要注意的是，阈值的选择是一项实践性工作，需要结合具体问题和需求进行调整和优化。同时，评估模型性能时应综合考虑多个指标（如召回率、精确率、F1-score 等），而不仅仅依赖于单一的阈值选择。

ROC 曲线（Receiver Operating Characteristic curve）和 AUC（Area Under the Curve）是用于评估二元分类模型性能的常用工具。(Receiver Operating Characteristic curve (ROC) and Area Under the Curve (AUC) are commonly used to evaluate the performance of binary classification models.)

ROC 曲线是以模型的真正例率（TPR，True Positive Rate）为纵轴，模型的虚警率（FPR，False Positive Rate）为横轴绘制的曲线。ROC 曲线展示了模型在不同阈值下的表现，阈值用于将模型的输出转换为二元分类的预测结果。曲线上的每个点表示模型在不同阈值下的 TPR 和 FPR。ROC 曲线越靠近左上角，说明模型的性能越好，因为这意味着模型在保持高召回率的同时降低了误报率。(ROC curve was drawn with the True Positive Rate (TPR) of the model as the vertical axis and the False Positive Rate (FPR) of the model as the horizontal axis. ROC curves show the model's performance under different thresholds, which are used to convert the model's output into binary classification predictions. Each point on the curve represents the TPR and FPR of the model under different thresholds. The closer the ROC curve is to the top left, the better the performance of the model, because it means that the model has a lower false positive rate while maintaining a high recall rate.)

AUC 是 ROC 曲线下的面积，取值范围在 0 到 1 之间。AUC 提供了一个单一的数值来度量模型的整体性能。AUC 等于 0.5 表示模型的预测能力与随机预测相当，AUC 大于 0.5 表示模型优于随机预测，AUC 接近 1 表示模型具有很高的预测能力。(AUC is the area under the ROC curve. The value ranges from 0 to 1. The AUC provides a single value to measure the overall performance of the model. AUC equal to 0.5 indicates that the prediction ability of the model is equal to that of random prediction, AUC greater than 0.5 indicates that the model is better than random prediction, and AUC close to 1 indicates that the model has high prediction ability.)

ROC 曲线和 AUC 提供了综合评估模型在不同阈值下的性能，并且不受类别不平衡问题的影响。在实际应用中，可以根据业务需求和阈值的重要性选择最合适的模型。

4. 交叉验证（cross validation）

交叉验证是一种评估机器学习模型性能的技术。它将数据集划分为训练集和验证集，并多次重复地使用不同的划分方式来训练和评估模型。具体而言，交叉验证的步骤如下：(Cross validation is a technique for evaluating the performance of

machine learning models. It divides the data set into a training set and a validation set, and uses different partitions many times to train and evaluate the model. Specifically, the steps of cross validation are as follows:)

1. 数据集划分：将原始数据集划分为 K 个近似大小的子集（通常称为折）。每个子集都充当一次验证集，而其他的 K-1 个子集合并作为训练集。(Data set partitioning: Partitioning the original data set into K approximately sized subsets (often called folds). Each subset acts as a verification set, while the other K-1 subsets are combined as a training set.)
2. 模型训练与评估：对于每一次交叉验证，使用 K-1 个子集进行模型训练，并在保留的子集上进行模型评估。评估指标可以是准确率、精确率、召回率、F1-score 等。(Model training and evaluation: For each cross-validation, K-1 subsets are used for model training, and model evaluation is performed on the retained subsets. The evaluation indexes can be accuracy rate, accuracy rate, recall rate, F1-score, etc.)
3. 重复步骤 2：重复步骤 2，直到所有的 K 个子集都被用作了一次验证集。(Repeat Step 2: Repeat Step 2 until all K subsets have been used as a verification set.)
4. 性能指标计算：将每次交叉验证的评估指标进行平均，得到模型的最终性能指标。(Calculation of performance indicators: The evaluation indicators of each cross-validation are averaged to obtain the final performance indicators of the model.)

交叉验证的主要优势是充分利用了数据集，提供了对模型性能的更准确估计。它可以帮助评估模型的稳定性和泛化能力，减少因数据集划分方式引起的偶然性。同时，交叉验证也能够帮助选择最佳的超参数设置，以提高模型的性能。(The main advantage of cross-validation is that it makes full use of the data set and provides a more accurate estimate of model performance. It can help to evaluate the stability and generalization ability of the model and reduce the contingency caused by the way the data set is divided. At the same time, cross-validation can also help select the best hyperparameter Settings to improve the performance of the model.)

常见的交叉验证方法包括 k 折交叉验证（k-fold cross-validation）、留一交叉验证（leave-one-out cross-validation）、留 p 交叉验证（leave-p-out cross-validation）等。其中，k 折交叉验证是最常用的方法，通常选择 k=5 或 k=10。

Lecture4：线性回归（Linear Regression）

一、介绍

线性回归是一种用于建立连续目标变量与自变量之间线性关系的统计模型。它的目标是通过拟合一个线性函数来预测目标变量的值。在线性回归中，假设自变量与目标变量之间存在一个线性关系，并且误差项服从正态分布。(Linear regression is a statistical model used to establish the linear relationship between continuous target variables and independent variables. Its goal is to predict the value of the target variable by fitting a linear function. In linear regression, it is assumed that there is a linear relationship between the independent variable and the target variable, and the error term follows a normal distribution.)

线性回归模型的一般形式可以表示为：

y = β0 + β1x1 + β2x2 + ... + βnxn + ε

其中，y 是目标变量，x1, x2, ..., xn 是自变量，β0, β1, β2, ..., βn 是回归系数，ε 是误差项。

线性回归的目标是找到最优的回归系数，使得预测值与实际观测值的差距最小化。通常使用最小二乘法来估计回归系数，即通过最小化误差平方和来确定最优参数值。(The goal of linear regression is to find the optimal regression coefficient that minimizes the difference between the predicted value and the actual observed value. The least square method is usually used to estimate the regression coefficient, that is, the optimal parameter values are determined by minimizing the sum of squares of errors.)

线性回归模型的优点包括简单易理解、计算效率高、可解释性强。它适用于连续目标变量与自变量之间存在线性关系的问题，并且可以用于预测、关联分析和因果推断等任务。

然而，线性回归也有一些限制，例如对非线性关系的建模能力较弱，可能会受到异常值和离群点的影响，以及对数据的假设要求（如线性关系、误差项的独立性、方差齐性等）。在实际应用中，可以根据问题的特点选择适当的回归模型和考虑其他的非线性方法。

对于线性回归模型，假设我们有 m 个样本，其中第 i 个样本的特征向量为 x^(i)，对应的真实值为 y^(i)，模型预测值为 y_hat^(i)。代价函数可以定义为：

J(theta) = (1/2m) * Σ[(y_hat^(i) - y^(i))^2]

其中，J(theta)表示代价函数，theta 表示模型的参数（包括截距和斜率），m 表示样本数量，Σ 表示求和符号。代价函数通过计算所有样本的预测值与真实值之间的差的平方，并取平均值，得到模型在整个数据集上的平均误差。

线性回归的目标是最小化代价函数，即找到使代价函数达到最小值的模型参数。一种常用的方法是使用梯度下降算法，通过迭代优化模型参数来逐步减小代价函数的值，从而得到最优的线性回归模型。

二、梯度下降（gradient descent）

梯度下降是一种常用的优化算法，用于求解函数的最小值。在机器学习中，梯度下降常用于调整模型的参数，使模型能够拟合训练数据并最小化代价函数。(Gradient descent is a common optimization algorithm used to solve the minimum value of a function. In machine learning, gradient descent is often used to adjust model parameters so that the model can fit the training data and minimize the cost function.)

梯度下降的基本思想是通过迭代更新参数，每次迭代都朝着代价函数梯度的反方向移动一定的步长，从而逐渐接近最优解。具体来说，梯度下降根据当前参数的梯度方向计算更新步长，并将参数沿着负梯度方向进行更新。(The basic idea of gradient descent is to update parameters iteratively, and each iteration moves a certain step in the opposite direction of the gradient of the cost function, so as to gradually approach the optimal solution. Specifically, gradient descent calculates the update step size based on the gradient direction of the current parameter and updates the parameter along the negative gradient direction.)

梯度下降有两种常见的变体：批量梯度下降（Batch Gradient Descent）和随机梯度下降（Stochastic Gradient Descent）。

- 批量梯度下降：在每次迭代中，批量梯度下降使用整个训练集计算代价函数关于参数的梯度，然后根据梯度大小和学习率来更新参数。这种方法通常需要更多的计算资源，但可能更稳定且更容易收敛到全局最优解。(Batch gradient descent: In each iteration, batch gradient descent uses the entire training set to calculate the gradient of the cost function with respect to the parameter, and then updates the parameter according to the gradient

size and learning rate. This approach usually requires more computing resources, but may be more stable and converge to a global optimal solution more easily.）

- 随机梯度下降：与批量梯度下降不同，随机梯度下降在每次迭代中只使用一个样本计算代价函数的梯度，并更新参数。这种方法的计算速度较快，但可能在参数更新过程中产生较大的波动，导致训练过程不太稳定。为了解决这个问题，还有一种改进的方法叫做小批量梯度下降，它在每次迭代中使用一小批样本计算梯度。（Random gradient descent: Unlike batch gradient descent, random gradient descent calculates the gradient of the cost function using only one sample in each iteration and updates the parameters. The calculation speed of this method is fast, but it may produce large fluctuations in the process of parameter updating, which leads to the instability of the training process. To solve this problem, there is an improved approach called small-lot gradient descent, which uses a small batch of samples to calculate the gradient in each iteration.）
- 小批量梯度下降是梯度下降的一种变体，介于批量梯度下降和随机梯度下降之间。它在每次迭代中使用一小批样本来计算代价函数的梯度，并更新模型的参数。与批量梯度下降相比，小批量梯度下降只使用一小批样本来计算梯度，而不是整个训练集。这样做的好处是在一定程度上减少了计算开销，特别是在处理大型数据集时更加高效。同时，相比于随机梯度下降，小批量梯度下降在更新参数时引入了一定的样本批量信息，从而更稳定地朝着最优解收敛。（Small batch gradient descent is a variation of gradient descent, between batch gradient descent and random gradient descent. It uses a small batch of samples in each iteration to calculate the gradient of the cost function and update the parameters of the model. In contrast to batch gradient descent, small batch gradient descent uses only a small batch of samples to calculate the gradient, rather than the entire training set. This has the advantage of reducing the computational overhead to some extent, especially when dealing with large data sets. Meanwhile, compared with random gradient descent, small-batch gradient descent introduces certain sample batch information when updating parameters, so that it converges towards the optimal solution more stably.）

梯度下降的性能受到学习率的影响。学习率控制了参数更新的步长，如果学习率过小，收敛速度可能会很慢；而如果学习率过大，可能会导致在最小值附近震荡或无法收敛。因此，选择合适的学习率对梯度下降算法的效果至关重要。

三、多项式回归（polynomial regression）

多项式回归的基本思想是在线性回归模型的基础上引入高阶项，通过添加自变量的高次幂和交互项来拟合数据中的非线性关系。例如，对于一个一元回归问题，多项式回归可以表示为 y = β0 + β1*x + β2*x^2 + ... + βn*x^n，其中 x 是自变量，y 是因变量，β0、β1、β2 等是模型的系数，n 是多项式的阶数。(The basic idea of polynomial regression is to introduce higher-order terms on the basis of the linear regression model, and to fit the nonlinear relation in the data by adding higher powers and interaction terms of independent variables. For example, for a unary regression problem, polynomial regression can be expressed as y = β0 + β1x + β2x^2 +... + βn*x^n, where x is the independent variable, y is the dependent variable, β0, β1, β2, etc. are the coefficients of the model, and n is the order of the polynomial.）

多项式回归可以灵活地拟合各种形状的曲线，使模型更能适应非线性关系的数据。通过选择合适的多项式阶数，可以在保持模型的灵活性的同时避免过拟合。然而，需要注意的是，随着多项式阶数的增加，模型复杂度也会增加，可能导致过拟合的风险。（Polynomial regression can fit curves of various shapes flexibly, making the model more adaptable to the data of nonlinear relationship. By choosing appropriate polynomial order, the flexibility of the model can be maintained while avoiding overfitting. However, it is important to note that as the order of the polynomial increases, so does the complexity of the model, which may lead to the risk of overfitting.）

在实际应用中，多项式回归可以用于探索和建模数据中的非线性关系，提供更准确的预测或描述。然而，需要根据具体问题和数据的特点来选择合适的多项式阶数，以及进行模型评估和调优，以避免过拟合并获得良好的预测性能。

学习曲线

学习曲线是一种用于评估机器学习算法性能的工具，它可以帮助我们理解模型在不同训练集大小或训练迭代次数下的表现情况。学习曲线通常是绘制训练集和验证集上模型性能指标随训练样本数量或训练迭代次数的变化曲线。（Learning curve is a tool used to evaluate the performance of machine learning algorithms. It can help us understand how a model

performs under different training set sizes or training iterations. The learning curve is usually the curve that draws the model performance index changes with the number of training samples or the number of training iterations on the training set and verification set.）

三、逻辑回归

逻辑回归是一种用于解决分类问题的机器学习算法。与线性回归不同，逻辑回归的输出是一个概率值，表示样本属于某个特定类别的概率。逻辑回归常用于二分类问题，但也可以扩展到多分类问题。（Logistic regression is a machine learning algorithm used to solve classification problems. Unlike linear regression, the output of logistic regression is a probability value representing the probability that the sample belongs to a particular category. Logistic regression is often used for binary classification problems, but can also be extended to multiclassification problems.）

逻辑回归的核心思想是使用一个逻辑函数（也称为 sigmoid 函数）将线性模型的输出转化为概率值。（The core idea of logistic regression is to use a logical function (also known as sigmoid function) to convert the output of a linear model into a probability value.）

逻辑回归的训练过程是通过最大似然估计来优化模型参数。我们希望最大化给定训练样本的条件概率，即最大化似然函数。为了简化计算，通常使用对数似然函数进行优化。最终的优化目标是找到最优的参数θ，使得对数似然函数最大化。

在实际应用中，逻辑回归可以使用梯度下降等优化算法进行参数估计。通过反复迭代更新参数，逐步逼近最优解。优化过程中的损失函数可以选择使用对数损失函数（也称为逻辑损失函数）来度量模型的预测误差。

Softmax 回归，也称为多类别逻辑回归或多项逻辑回归，是一种用于多类别分类的模型。它是逻辑回归在多类别情况下的推广。（Softmax regression, also known as multicategory logistic regression or multinomial logistic regression, is a model used for multicategory classification. It is a generalization of logistic regression in the case of multiple categories.）

Softmax 回归的目标是将输入样本分到多个不同的类别中。与二元逻辑回归不同，Softmax 回归使用了一种称为 Softmax 函数的激活函数，将输出转化为各个类别的概率分布。Softmax 函数可以将输入向量映射到一个概率分布上，使得所有类别的概率之和为 1。（The goal of Softmax regression is to divide the input samples into a number of different categories. Unlike binary logistic regression, Softmax regression uses an activation function called the Softmax function to convert the output into a probability distribution for each category. The Softmax function maps the input vector to a probability distribution such that the sum of the probabilities for all categories is 1.）

Lecture5:降维（Dimensionality Reduction）

一、介绍

降维（dimensionality reduction）是指将高维数据映射到低维空间的过程。在机器学习和数据分析中，降维是一种常用的技术，它有助于减少特征的数量，消除冗余信息，提高模型的效率和泛化能力，并可视化高维数据。（dimensionality reduction refers to the process of mapping higher-dimensional data to lower-dimensional Spaces. Dimension reduction is a common technique used in machine learning and data analysis to help reduce the number of features, eliminate redundant information, improve model efficiency and generalization, and visualize high-dimensional data.）

降维方法可以分为两大类：特征选择（Feature Selection）和特征提取（Feature Extraction）。

特征选择是指选择原始特征中最具有代表性和相关性的子集，以保留原始数据的主要信息。常用的特征选择方法有过滤法、包装法和嵌入法等。（Feature selection refers to the selection of the most representative and relevant subset of the original features to retain the main information of the original data. The commonly used feature selection methods include filtering, packaging and embedding.）

特征提取则是通过线性或非线性变换，将原始高维特征转换为低维特征表示。常见的特征提取方法包括主成分分析（Principal Component Analysis, PCA）、线性判别分析（Linear Discriminant Analysis, LDA）、非负矩阵分解（Non-negative Matrix Factorization, NMF）等。（Feature extraction is to transform the original high-dimensional feature into low-dimensional feature representation through linear or nonlinear transformation. Common feature extraction methods include Principal Component Analysis (PCA), Linear Discriminant Analysis (Linear Discriminant Analysis), LDA), Non-negative Matrix Factorization (NMF), etc.）

降维的好处包括减少存储空间需求、加快模型训练和预测的速度、降低维度灾难带来的问题、提高可视化和解释性等。（The benefits of dimensionality reduction include reduced storage requirements, faster model training and prediction, reduced problems with dimensional disasters, improved visualization and interpretation, and more.）

选择适当的降维方法需要考虑数据的特点、任务需求和算法的适用性。同时，降维过程中需要注意保留数据的重要信息，避免丢失关键特征或引入较大的信息损失。（The selection of appropriate dimension reduction method needs to consider the characteristics of data, task requirements and the applicability of the algorithm. At the same time, important data information should be retained during dimension reduction to avoid loss of key features or information loss.）

二、投影（projection）

在降维中，投影（projection）是一种常用的方法，它将高维数据映射到低维空间中的子空间。投影可以通过线性变换实现，将数据点投影到一个低维的子空间上。

在投影中，我们选择一个投影矩阵（或称为映射矩阵），它将原始数据的每个样本点映射到一个低维空间上的向量。这个投影矩阵的选择是通过优化问题或某种约束条件来确定的，以便在降维过程中最大程度地保留数据的结构和信息。

在投影过程中，我们通常会选择使得投影后数据的方差最大的投影方向，这样可以尽可能保留数据的方差信息。这种方法称为主成分分析（Principal Component Analysis, PCA），它通过找到数据的主要方向（即主成分），实现数据的降维。

在降维中，我们希望通过投影将高维数据映射到低维空间，同时尽可能地保留数据的信息。其中一个常用的目标是最大化投影后数据的方差。

数据的方差表示数据的变化程度，它是描述数据分布的一个重要统计量。如果我们在投影后的低维空间中能够保留较高的方差，意味着投影后的数据点之间的差异性较大，数据的信息较丰富。相反，如果投影后的数据方差较小，数据点之间的差异性较小，数据的信息丢失较多。（The variance of data represents the degree of variation of data, which is an important statistic to describe the distribution of data. If we can retain a high variance in the low-dimensional space after projection, it means that the difference between the projected data points is large and the data information is rich. On the contrary, if the variance of the projected data is small, the difference between data points is small, and the data will lose more information.）

因此，当我们选择投影方向时，通常会选择使得投影后数据的方差最大的投影方向。这样做的目的是尽可能保留原始数据的变化程度，以便在降维后的空间中仍能够较好地区分数据样本。(When we choose the projection direction, we usually choose the projection direction that maximizes the variance of the projected data. The purpose of this is to preserve the

degree of variation of the original data as much as possible so that the data sample can still be better located in the space after dimensionality reduction.）

具体来说，在主成分分析（PCA）中，我们通过计算数据的协方差矩阵，找到使得投影后数据方差最大的特征向量，作为投影方向。这些特征向量称为主成分，它们对应着数据中最显著的方差方向。通过选择较多的主成分，我们可以保留更多的数据方差，从而更好地保留数据的信息。

总之，通过选择投影方向使得投影后数据的方差最大，我们可以尽可能保留数据的变化程度和信息，从而实现有效的降维。

三、主成分分析（PCA, principal component analysis）

1. 步骤：
    1. 数据标准化：对原始数据进行标准化处理，使得每个特征的均值为 0，标准差为 1。这样可以消除不同特征之间的量纲差异。（Data standardization: The original data is standardized so that the mean value of each feature is 0 and the standard deviation is 1. This eliminates dimensional differences between different features.）
    2. 计算协方差矩阵：根据标准化后的数据，计算其协方差矩阵。协方差矩阵反映了不同特征之间的相关性。（Calculate the covariance matrix: calculate the covariance matrix according to the standardized data. The covariance matrix reflects the correlation between different features.）
    3. 特征值分解：对协方差矩阵进行特征值分解，得到特征值和对应的特征向量。特征值表示每个主成分所解释的方差比例，特征向量表示每个主成分的方向。（Eigenvalue decomposition: The eigenvalue decomposition is carried out on the covariance matrix to obtain the eigenvalue and corresponding eigenvector. The eigenvalue represents the proportion of variance explained by each principal component, and the eigenvector represents the direction of each principal component.）
    4. 特征值排序：将特征值按降序排列，选择前 k 个特征值对应的特征向量作为主成分。（Eigenvalue sorting: The eigenvalues are sorted in descending order, and the eigenvectors corresponding to the first k eigenvalues are selected as the main components.）特征值越大对应的主成分越重要
    5. 投影数据：将原始数据投影到选定的主成分上，得到降维后的数据。投影过程就是将原始数据点在主成分方向上的投影长度作为新的特征值。（Projection data: The original data is projected onto the selected principal component to obtain the data after dimensionality reduction. In the projection process, the projection length of the original data point in the principal component direction is taken as the new eigenvalue.）

    通过主成分分析，我们可以选择保留较高方差比例的主成分，实现数据的降维，并且尽可能地保留原始数据的信息。这有助于减少数据的维度，并可用于可视化、特征选择、数据压缩等应用场景。（Through principal component analysis, we can choose to retain the principal component with a higher variance ratio, achieve data dimension reduction, and retain the information of the original data as much as possible. This helps reduce the dimensionality of the data and can be used in visualization, feature selection, data compression, and other application scenarios.）
2. 数据标准化和中心化

a) 数据标准化和数据中心化是常见的数据预处理步骤，通常在数据分析和机器学习任务中使用。

b) 数据标准化（Normalization）是指将数据转换为特定范围或分布的过程，以便使得不同特征具有相同的尺度。常见的标准化方法包括将数据转换为均值为 0，标准差为 1 的标准正态分布，或者将数据缩放到特定的最小值和最大值范围之间。标准化可以消除特征之间的量纲差异，使得不同特征在模型训练过程中对结果的贡献更加均衡。（The Normalization of data is the process of converting data into specific ranges or distributions so that different characteristics have the same scale. Common normalization methods include converting the data to a standard normal distribution with a mean of 0 and a standard deviation of 1, or scaling the data to a specific range between minimum and maximum values. Standardization can eliminate dimensional differences among features and make the contribution of different features to the results more balanced in the process of model training.）

c) 数据中心化（Centering）是指将数据的均值移动到原点（或其他中心位置）的过程。它通过减去数据的均值来实现，使得数据的中心位置为 0。数据中心化有助于消除特征之间的偏差，使得模型更加关注特征的相对变化而不是

绝对值。(Data centralization (Centering) is the process of moving the mean of data to the origin (or other central location). It does this by subtracting the mean of the data so that the center of the data is zero. Data centralization helps eliminate bias between features, making the model more concerned with relative changes in features rather than absolute values.)

d) 在某些情况下，数据标准化和数据中心化可以同时进行，即先对数据进行中心化，然后再进行标准化。这样可以确保数据在各个特征上具有相同的尺度，并且均值为 0。这种预处理可以帮助模型更好地捕捉到数据的结构和变化模式，并提高模型的稳定性和性能。(In some cases, data standardization and data centralization can go hand in hand, where data is first centralized and then standardized. This ensures that the data has the same scale across the features and has a mean of 0. This preprocessing can help the model capture the structure and change pattern of data better, and improve the stability and performance of the model.)

e) 总之，数据标准化和数据中心化是常用的数据预处理技术，用于消除特征之间的量纲差异和偏差，以提高数据分析和机器学习模型的效果

PCA 进行标准化和中心化的原因

1. 消除量纲差异：不同特征可能具有不同的尺度和单位，如果不进行标准化，那些具有较大尺度的特征将会对 PCA 的结果产生更大的影响。标准化可以将特征转换为具有相同尺度的形式，使得它们在 PCA 计算中的贡献更加均衡。
2. 中心化数据：中心化是指通过减去数据的均值来将数据的中心位置移动到原点或其他中心位置。中心化数据可以消除特征之间的偏差，使得 PCA 能够更好地捕捉到数据的相对变化模式。如果数据没有经过中心化处理，那么 PCA 的主成分可能会受到数据均值的影响，导致结果不准确或不稳定。

通过标准化和中心化数据，可以确保在进行 PCA 分析时各个特征具有相同的尺度，并且数据的中心位置在原点附近。这样可以使得 PCA 计算更加准确和稳定，确保主成分能够更好地捕捉到数据的方差和结构，从而提高降维结果的可靠性和解释性。(By standardizing and centralizing the data, it is possible to ensure that each feature has the same scale when PCA is performed and that the center of the data is located near the origin. This can make PCA calculation more accurate and stable, and ensure that the principal component can better capture the variance and structure of the data, so as to improve the reliability and interpretation of the dimension reduction results.)

3. 奇异值分解（**SVD(singular value decomposition)**）

是一种常用的矩阵分解方法，将一个矩阵分解为三个矩阵的乘积。它在数据分析和降维技术中有广泛的应用。

SVD 将一个矩阵分解为三个矩阵：U、Σ 和 V^T。其中，U 和 V 是正交矩阵，Σ 是一个对角矩阵，对角线上的元素称为奇异值。奇异值表示了矩阵在每个特征向量方向上的重要性，类似于 PCA 中的特征值。(SVD decomposes a matrix into three matrices: U, Σ, and V^T. Where U and V are orthogonal matrices, sigma is a diagonal matrix, and the elements on the diagonal are called singular values. Singular values represent the importance of the matrix in the direction of each eigenvector, similar to eigenvalues in PCA.)

4. 解释方差比

方差比（Variance Ratio）是一种用于衡量数据集中各个主成分的解释能力的指标。在主成分分析（PCA）中，方差比可以帮助我们理解每个主成分对总方差的贡献程度。

方差比是每个主成分的方差与总方差之比。主成分的方差表示了该主成分所包含的数据变异程度，而总方差则表示了原始数据集中的总变异程度。方差比越大，说明相应的主成分能够解释更多的数据变异性。

通过观察方差比，我们可以确定需要保留的主成分数量。通常情况下，我们希望保留那些能够解释大部分数据变异性的主成分，而忽略那些贡献较小的主成分。方差比可以帮助我们确定一个合适的主成分数目，以便在降维过程中平衡数据的信息保留和维度的减少。

方差比可以以累计的方式进行计算，即将每个主成分的方差逐步累加，并计算累计方差与总方差之比。通过绘制累计方差比曲线，我们可以直观地观察到主成分数量与累计方差比之间的关系，进一步确定适当的主成分数目。

四、 独立成分分析(Independent Component Analysis (ICA))

独立成分分析（Independent Component Analysis，ICA）是一种用于从混合信号中提取独立成分的统计方法。它的目标是将观测到的多维信号分解为相互独立的成分，这些成分在原始信号中具有最大的非高斯性。

ICA 假设原始信号是通过线性组合的方式得到的，并且成分之间是相互独立的。通过对观测到的混合信号进行处理，ICA 可以估计出原始信号的独立成分和相应的混合系数。

ICA 的核心思想是通过最大化成分的非高斯性来解耦混合信号。非高斯性指的是数据分布的非对称性或非高斯分布特性，而高斯分布的数据在线性组合后仍然保持高斯性。因此，通过寻找使成分最非高斯的投影方向，ICA 可以将混合信号分解成相互独立的成分。

ICA 假设原始信号是通过线性混合得到的，且成分之间是相互独立的。因此，ICA 的任务是将混合信号分解为相互独立的成分，并找到用于分离这些成分的线性变换。通过分离出独立成分，我们可以更好地理解信号的组成部分，以及它们在混合过程中的贡献和相互关系。ICA 在信号处理、图像处理、语音处理等领域具有广泛的应用，可以帮助我们揭示数据中隐藏的独立成分信息。（ICA assumes that the original signal is obtained by linear mixing and that the components are independent of each other. The ICA's task, therefore, is to decompose the mixed signal into its independent components and find the linear transformation used to separate these components. By isolating the independent components, we can better understand the components of the signal and their contributions and interrelationships in the mixing process. ICA has a wide range of applications in signal processing, image processing, speech processing and other fields, which can help us reveal the independent component information hidden in the data.）

五、各种降维中的方法

1. 非负矩阵分解 NMF（Non-Negative Matrix Factorization）是一种矩阵分解方法，用于将一个非负矩阵分解为两个非负矩阵的乘积。NMF 的目标是找到两个矩阵，使得它们的乘积能够近似原始矩阵，并且所有的元素都是非负的。

NMF 在特征提取和降维中具有广泛的应用。通过将原始数据表示为非负矩阵的乘积形式，NMF 能够捕捉到数据中的潜在特征和模式，同时还具有稀疏性和解释性。NMF 适用于处理非负数据，例如图像、文本、音频等领域，可以用于图像分析、文本挖掘、信号处理等任务。NMF 的优点包括数据解释性强、对异常值和噪声具有鲁棒性等。

2. Manifold learning（流型学习）

Manifold learning（流形学习）是一类用于非线性降维和数据可视化的方法，它试图从数据中发现潜在的低维流形结构。以下是一些常见的流形学习方法：

1. 多维尺度变换（MDS）：通过计算数据点之间的距离或相似性，将高维数据映射到低维空间，保持数据点之间的距离关系。
2. 等距映射（Isomap）：基于图论的方法，利用数据点之间的测地距离来构建数据的近邻图，然后通过最小化低维空间中的距离与原始数据的测地距离之间的差异，将数据映射到低维空间。
3. 局部线性嵌入（LLE）：通过局部线性逼近来描述数据点之间的关系，将每个数据点表示为其最近邻数据点的线性组合，然后在低维空间中保持这些线性关系。
4. 流形正则化（Manifold Regularization）：结合了图方法和正则化技术，通过在降维过程中约束数据点之间的平滑性，实现对流形结构的保持。
5. 流形哈希（Locality-sensitive Hashing）：通过将数据点映射到二进制码的方式来实现高效的近似最近邻搜索，常用于大规模数据集的降维和相似性搜索。

这些方法在处理高维数据、非线性结构的数据或需要可视化数据时具有重要的应用价值，能够帮助我们理解数据中的潜在结构和关系。

Manifold learning 方法的应用广泛，包括图像处理、语音识别、文本分析等领域。它可以帮助我们理解数据的内在特征、减少数据维度、提取有用的特征以及可视化高维数据，从而为后续的数据分析和机器学习任务提供更好的基础。

Lecture6:支持向量机（SVM）

一、SVM 介绍

支持向量机（Support Vector Machine，SVM）是一种用于分类和回归分析的机器学习算法。它的目标是通过在特征空间中找到一个最优的超平面，将不同类别的样本点尽可能地分开。（Support Vector Machine (SVM) is a machine learning algorithm for classification and regression analysis. Its goal is to separate the sample points of different classes as far as possible by finding an optimal hyperplane in the feature space.）

支持向量机的关键思想是将样本点映射到高维特征空间，使得在该空间中可以线性可分或近似线性可分。然后，在特征空间中寻找一个超平面，使得离超平面最近的样本点（即支持向量）到超平面的距离最大化。这个最大化间隔的超平面被称为最优分割超平面，可以有效地进行分类。（The key idea of support vector machine is to map the sample points to a high-dimensional feature space so that it can be linearly separable or nearly linearly separable in this space. Then, a hyperplane is found in the feature space so that the distance from the nearest sample point (i.e. support vector) to the hyperplane is maximized. This hyperplane with maximum spacing is called the optimally segmented hyperplane and can be classified efficiently.）

支持向量机的优势在于它不仅仅局限于线性分割，还可以通过使用不同的核函数来处理非线性问题。常用的核函数包括线性核、多项式核和高斯核等。这些核函数能够将样本点映射到高维特征空间，从而在高维空间中实现非线性分割。(The advantage of support vector machines is that they are not limited to linear segmentation, but can also handle nonlinear problems by using different kernel functions. Common kernel functions include linear kernel, polynomial kernel and Gaussian kernel. These kernel functions can map sample points to high-dimensional feature space, so as to realize nonlinear segmentation in high-dimensional space.）

支持向量机还具有较好的泛化能力和对于噪声的鲁棒性。它可以有效地处理高维数据和小样本数据，并且在处理二分类和多分类问题时都具有良好的性能。(Support vector machine also has good generalization ability and robustness to noise. It can deal with high dimensional data and small sample data effectively, and has good performance when dealing with binary classification and multi-classification problems.)

总的来说，支持向量机是一种强大的分类和回归算法，它通过寻找最优分割超平面在特征空间中最大化间隔，从而实现有效的模式识别和预测。（Support vector machine (SVM) is a powerful classification and regression algorithm that maximizes the spacing in feature space by finding the optimal segmented hyperplane to achieve efficient pattern recognition and prediction.）

基本过程：

1. 数据准备：首先，准备带有标签的训练数据集，其中包含特征和相应的类别标签。
2. 特征转换：根据具体问题和数据特点，将原始特征转换为高维特征空间。这可以通过使用不同的核函数来实现，例如线性核、多项式核或高斯核等。
3. 定义超平面：在转换后的特征空间中，寻找一个超平面来最佳地分割不同类别的样本点。超平面的选择是通过最大化间隔来进行的，即使得离超平面最近的支持向量到超平面的距离最大化。
4. 训练模型：使用训练数据集对支持向量机模型进行训练，以找到最优的超平面参数。训练过程通常涉及最小化损失函数，并且可能涉及正则化和其他优化技术。
5. 模型评估：使用测试数据集对训练得到的模型进行评估，计算分类的准确率、精确度、召回率等指标，以评估模型的性能。
6. 模型预测：使用训练好的支持向量机模型对新的未标记样本进行分类预测。

二、特征尺度 Feature Scales

SVM 对于特征尺度是敏感的，因此在使用 SVM 之前，通常需要对特征进行尺度调整。这是因为 SVM 的决策边界是基于特征之间的距离计算的，如果不同特征的尺度差异较大，可能会导致某些特征在决策中起到更大的作用，而忽略了其他特征。（SVM is sensitive to feature scale, so it is usually necessary to scale features before using SVM. This is because the decision boundary of SVM is calculated based on the distance between features. If the scale difference of different features is large, some features may play a greater role in the decision, while other features are ignored.）

常见的特征尺度调整方法包括：

1. 标准化：对每个特征进行标准化处理，使其均值为 0，标准差为 1。这可以通过减去特征均值并除以特征标准差来实现。（Standardization: Each feature is standardized so that its mean is 0 and its standard deviation is 1. This can be done by subtracting the characteristic mean and dividing by the characteristic standard deviation.）
2. 归一化：将特征缩放到一个固定的范围，通常是[0, 1]或[-1, 1]。这可以通过将特征值减去最小值并除以特征值范围来实现。（Normalization: Scaling the feature to a fixed range, usually [0, 1] or [-1, 1]. This can be done by subtracting the eigenvalue from the minimum value and dividing by the range of eigenvalues.）

通过对特征进行尺度调整，可以确保所有特征在相同的尺度上进行比较，避免尺度差异对 SVM 的性能产生影响。这样可以提高 SVM 的稳定性和准确性，并确保每个特征都能够对分类起到合适的作用。（By scaling features, we can ensure that all features are compared on the same scale to avoid the impact of scale differences on SVM performance. This can improve the stability and accuracy of SVM, and ensure that each feature can play a proper role in classification.）

三、硬边缘分类 hard margin classification

硬边缘分类是支持向量机（SVM）中的一种分类方式，其目标是在完全分离两个类别的情况下构建最优的决策边界。具体来说，硬边缘分类要求训练样本被正确地分类，并且要求决策边界与最近的训练样本之间的间隔最大化。(Hard margin classification is a classification method in support vector machine (SVM), whose goal is to construct the optimal decision boundary when two categories are completely separated. Specifically, hard margin classification requires that the training samples be correctly classified and that the spacing between the decision boundary and the nearest training sample be maximized.)

在硬边缘分类中，SVM 会尝试找到一个最优的超平面来将不同类别的样本分开，使得两个类别的样本完全分离，并且使得超平面到最近的训练样本的距离最大化。这样做的目的是为了使分类结果更加鲁棒和准确。(In hard edge classification, SVM will try to find an optimal hyperplane to separate samples of different classes, so that samples of two classes are completely separated, and maximize the distance between the hyperplane and the nearest training sample. The purpose of this is to make the classification results more robust and accurate.)

局限性：

1. 对噪声和异常值敏感：Hard margin classification 要求数据是线性可分的，对于包含噪声或异常值的数据，可能会导致无法找到一个完全分开两个类别的超平面。(Sensitivity to noise and outliers: Hard margin classification requires that the data be linearly separable and, for data containing noise or outliers, may result in an inability to find a hyperplane that completely separates the two categories.)
2. 对数据的尺度和分布敏感：Hard margin classification 对于数据的尺度和分布敏感。如果数据的尺度差异很大，或者类别之间存在重叠，可能会导致分类结果不准确。（Sensitive to data scale and distribution: Hard margin classification is sensitive to data scale and distribution. If the scale of the data is very different, or there is overlap between categories, the classification results may be inaccurate.）
3. 需要线性可分的数据：Hard margin classification 要求数据是线性可分的，即存在一个超平面可以完全分开两个类别的样本。然而，在现实世界的许多情况下，数据是非线性可分的，这就需要采用其他方法或技巧来处理非线性情况。（Requires linearly separable data: Hard margin classification requires that the data be linearly separable, i.e. there is a hyperplane that can completely separate the two categories of samples. However, in many real-world situations, data is nonlinearly separable, which requires other methods or techniques to deal with nonlinearities.）
4. 模型复杂度高：在处理大规模数据集时，Hard margin classification 的模型复杂度较高，需要计算和存储大量的支持向量，导致计算和存储资源的消耗较大。(High model complexity: When processing large-scale data sets, the Hard margin classification model has high complexity, which requires calculation and storage of a large number of support vectors, resulting in a large consumption of computing and storage resources.)

四、软边界分类 soft margin classification

软边界分类是指在支持向量机（SVM）中允许存在一定程度的分类错误，即容忍一些样本点位于超平面的错误一侧。相

比于硬边界分类，软边界分类更加灵活，能够处理一些数据集中存在噪声或重叠的情况。(Soft boundary classification means that a certain degree of classification error is allowed in support vector machine (SVM), that is, some sample points are tolerated on the wrong side of the hyperplane. Compared with hard boundary classification, soft boundary classification is more flexible and can deal with noise or overlap in some data sets.)

1. 数据预处理：与硬边界分类相同，首先对输入数据进行预处理，包括特征缩放和标准化等操作。
2. 构建目标函数：目标函数由两部分组成：第一部分是最小化分类错误和松弛变量的总和，用于确保分类的准确性；第二部分是正则化项，用于控制超平面的复杂度。
3. 设置软边界参数：通过引入松弛变量，可以灵活地控制容忍错误的程度。通常使用一个正则化参数 C 来调节松弛变量的惩罚力度，较小的 C 值表示更多容忍错误，较大的 C 值表示更严格要求分类准确性。
4. 求解优化问题：使用优化算法（如凸优化算法）求解目标函数，找到最优的超平面和松弛变量。
5. 核函数的应用（可选）：与硬边界分类相同，如果数据不是线性可分的，可以使用核函数将输入空间映射到一个高维特征空间，从而在高维空间中寻找一个线性可分的超平面。
6. 预测和分类：训练完成后，使用得到的超平面对新样本进行预测和分类，根据样本点与超平面的距离确定其所属类别。

软边界分类允许一定的错误分类，使得模型更具有鲁棒性和泛化能力，适用于处理一些复杂的实际问题。然而，设置合适的正则化参数 C 是关键，过小或过大的 C 值都可能导致不理想的分类结果。因此，对于软边界分类，需要进行适当的参数调优和交叉验证来选择最优的 C 值。(Soft boundary classification allows some wrong classification, which makes the model more robust and generalization ability, and is suitable for dealing with some complex practical problems. However, setting the proper regularization parameter C is the key. Too small or too large a value of C may lead to unsatisfactory classification results. Therefore, for soft boundary classification, appropriate parameter tuning and cross-validation are needed to select the optimal C value.)

软边界分类的目的：

1. 容忍噪声和异常值：在现实数据中，可能存在噪声或异常值，这些异常样本可能对模型的训练和预测产生不良影响。软边界分类允许一定程度上的错分，可以更好地容忍这些异常样本，减少对其的过度敏感性。(Tolerance of noise and outliers: In real-world data, there may be noise or outliers, which may adversely affect the training and prediction of the model. Soft boundary classification allows a certain degree of misclassification, which can better tolerate these abnormal samples and reduce over-sensitivity to them.) ---提高模型的鲁棒性（对噪声或异常点的容忍程度，也就是模型的抗干扰能力）
2. 提高模型泛化能力：硬边界分类在训练集上表现良好，但可能在新样本上过度拟合，导致泛化能力较差。软边界分类允许一定程度上的重叠，能够更好地适应未见过的数据，并提高模型在新样本上的预测能力。(Improve model generalization ability: Hard boundary classification performs well on training sets, but may overfit on new samples, resulting in poor generalization ability. Soft boundary classification allows a degree of overlap, better ADAPTS to previously unseen data, and improves the model's predictive power on new samples.) ---提高模型的泛化能力

五、评估模型鲁棒性的方法

1. 交叉验证：使用交叉验证来评估模型的性能。通过将数据集划分为训练集和测试集，多次重复训练和测试过程，可以获得模型在不同数据集上的表现。交叉验证可以帮助检测模型对于不同数据集的稳定性。
2. 留一验证：特别适用于小样本数据集。留一验证是一种极端的交叉验证方法，每次只保留一个样本作为测试集，其余样本作为训练集。通过多次迭代，计算模型在每个样本上的预测误差，可以评估模型对于个别样本的鲁棒性。
3. 自助采样法（Bootstrap）：自助采样是一种有放回的随机采样方法，可以从原始数据集中重复抽样产生多个大小相等的训练集。通过在多个自助样本上训练模型，并在原始数据集上进行测试，可以评估模型的稳定性和鲁棒性。
4. 异常值检测：检测和处理数据中的异常值，这些异常值可能会对模型的训练和预测产生不良影响。常用的方法

包括箱线图、Z-Score 和基于模型的异常值检测方法。

5. 对抗性样本攻击：通过引入对抗性样本，即针对模型进行故意设计的具有一定扰动的样本，来评估模型的鲁棒性。通过检测模型对于对抗性样本的识别和预测能力，可以评估模型在面对意外输入时的表现。（Adversarial sample attack: By introducing adversarial samples, that is, samples with certain disturbance designed deliberately for the model, to evaluate the robustness of the model. The model's performance in the face of unexpected inputs can be evaluated by testing its ability to recognize and predict adversarial samples.）

六、SVM 的核函数

SVM（支持向量机）通过使用核函数来进行非线性分类或回归任务。核函数是一种将输入特征映射到高维特征空间的函数，它可以将原始数据从低维空间转换到高维空间，从而使线性不可分的数据在高维空间中变得线性可分。(SVM (support vector machine) performs non-linear classification or regression tasks by using kernel functions. Kernel function is a kind of function that maps the input feature to the high-dimensional feature space. It can convert the original data from the low-dimensional space to the high-dimensional space, so that the linearly indivisible data becomes linearly divisible in the high-dimensional space.)

1.多项式核函数

多项式核函数是支持向量机（SVM）中常用的核函数之一。它将输入样本映射到高维特征空间，并使用多项式函数计算样本之间的内积。

多项式核函数的定义如下： $K(x, y) = (\gamma x^T y + r)^d$

其中，x 和 y 是输入样本，γ 是缩放因子，r 是常数项，d 是多项式的阶数。

多项式核函数通过计算样本在高维空间中的多项式内积，可以将线性不可分的样本在高维空间中变得线性可分。通过增加多项式的阶数，可以增加模型的复杂度，从而提高对非线性数据的拟合能力。

选择合适的多项式核函数的参数（如 γ、r 和 d）是 SVM 中的关键任务。较大的 γ 值会使样本之间的相似度下降得更快，较小的 γ 值则会使相似度下降得更慢。常数项 r 可以用来控制多项式的偏移量。多项式的阶数 d 决定了决策边界的复杂度，较高的阶数能够更好地适应复杂的数据分布，但也容易导致过拟合。

2.高斯核函数

高斯核函数，也称为径向基函数（Radial Basis Function, RBF），是支持向量机（SVM）中常用的核函数之一。它通过计算样本之间的高斯相似度来将输入样本映射到高维特征空间。

高斯核函数的定义如下： $K(x, y) = \exp(-\gamma \|x - y\|^2)$

其中，x 和 y 是输入样本，γ 是控制高斯核函数宽度的参数。

高斯核函数通过计算样本之间的欧氏距离的平方，并使用指数函数对其进行变换，得到样本在高维空间中的相似度。较小的 γ 值会使高斯核函数的宽度变大，样本之间的相似度下降得更慢，而较大的 γ 值会使高斯核函数的宽度变小，样本之间的相似度下降得更快。

使用高斯核函数可以将非线性可分的样本在高维空间中变得线性可分，从而提高了 SVM 对复杂数据的拟合能力。通过调整 γ 的取值，可以控制模型的平滑性和复杂度，较小的 γ 值会产生较平滑的决策边界，较大的 γ 值则会产生更复杂的决策边界。

当 γ 值较大时，高斯核函数会将距离较近的样本点视为更重要的邻居，而距离较远的样本点则会被较少考虑。这意味着分类器会更加关注局部的细节和小尺度的特征变化，从而使决策边界适应更多的样本点，包括噪声点。这使得决策边界可以更好地拟合训练数据，但也可能导致过拟合，即对训练数据过度敏感而在新数据上表现较差。(When the value of γ is high, the Gaussian kernel will treat the closer sample points as more important neighbors, while the farther sample points will be less considered. This means that the classifier will pay more attention to local details and small-scale feature changes, so that the decision boundary can adapt to more sample points, including noise points. This allows the decision boundary to better fit the training data, but it can also lead to overfitting, i.e. being overly sensitive to the training data and performing poorly on the new data.)

相反，当 γ 值较小时，高斯核函数的影响范围扩大。这意味着更多的样本点会被视为邻居，并且分类器更加关注全局的数据分布。这可能导致决策边界更平滑和简单，减少过拟合的风险。但较小的 γ 值也可能导致决策边界无法完全适

应训练数据中复杂的局部特征。(On the contrary, when γ value is small, the influence range of Gaussian kernel function expands. This means that more sample points will be regarded as neighbors, and the classifier pays more attention to the global data distribution. This may result in smoother and simpler decision boundaries, reducing the risk of overfitting. However, the small γ value may also cause the decision boundary to be unable to fully adapt to the complex local features in the training data.)

七、SVM regression（SVM 解决回归问题）

SVM 回归是一种基于超平面的回归方法，它通过最小化预测误差和控制离群值的影响来拟合数据。通过调整 ε 和选择合适的核函数，可以调节模型的灵活性和对非线性关系的适应能力。

在 SVM 回归中，目标是找到一个函数，该函数能够将输入数据映射到连续的输出变量。与分类问题不同，SVM 回归的目标是尽可能使大部分训练样本落在预测函数的 ε-tube（带状区域）内，同时保持尽可能少的训练样本位于 ε-tube 之外。

SVM 回归的基本思想是在特征空间中寻找一个最优的超平面，该超平面尽可能地拟合训练样本，同时允许一定的误差。与分类问题不同，SVM 回归使用 ε（epsilon）参数来控制对训练样本的容忍度。具体来说，对于每个训练样本，我们允许其预测值与真实值之间的差异在 ε 范围内。

SVM 回归的优化目标是最小化损失函数，其中损失函数由两部分组成：第一部分是预测误差的平方，第二部分是对预测值超出 ε 范围的惩罚。通过调整 ε 的值，可以控制模型的灵活性和容忍度。

与 SVM 分类器类似，SVM 回归也可以使用不同的核函数来处理非线性关系，例如多项式核函数和高斯核函数。这些核函数可以通过内积计算将输入特征映射到高维空间，从而使模型能够更好地拟合非线性关系。

在 SVM 回归中，我们引入一个容错参数 ε（epsilon），它定义了一个范围，表示我们允许预测值与真实值之间的最大差异。具体而言，我们希望大部分的训练样本都落在一个带状区域内，该区域的宽度由 ε 决定，同时我们希望尽量少的样本位于带状区域之外。

为了找到最佳的拟合曲线，SVM 回归的目标是最小化一个损失函数。该损失函数由两个部分组成：第一个部分是预测误差的平方，表示预测值与真实值之间的差异；第二个部分是对预测值超出 ε 范围的惩罚项。这样，模型就可以在保持大部分样本落在带状区域内的同时，尽量减少样本在带状区域之外的情况。

增大 ε 的效果是放宽对预测误差的限制，使得模型更容忍与真实值之间的差异。这意味着拟合曲线可以更灵活地适应数据，允许一些样本落在容错范围之外。因此，较大的 ε 可以导致模型更加宽松，具有更高的容忍度。

C 的值越小，表示对误分类的容忍度越高，模型更倾向于选择较大的间隔，可能会容忍更多的训练误差。这可能导致模型在训练数据上的拟合效果较差，但可能具有更好的泛化能力。

相反，C 的值越大，表示对误分类的容忍度越低，模型更倾向于选择较小的间隔，更严格地拟合训练数据。这可能导致模型在训练数据上的拟合效果较好，但可能对噪声或离群值更敏感，并且可能导致过拟合的风险增加。

Lecture7:朴素贝叶斯算法（Naive Bayes）

一、朴素贝叶斯定理

朴素贝叶斯定理是基于贝叶斯定理推导出的一种分类方法。它表达了在已知特征条件下，计算类别的后验概率的关系。根据朴素贝叶斯定理，给定一个观察到的特征向量 x 和类别变量 y，后验概率 P(y|x) 可以通过先验概率 P(y) 和条件概率 P(x|y) 计算得出。具体公式为：

P(y|x) = (P(x|y) * P(y)) / P(x)

其中，P(x|y) 是在类别 y 下观察到特征向量 x 的条件概率，P(y) 是类别 y 的先验概率，P(x) 是特征向量 x 的边缘概率。朴素贝叶斯分类器通过计算不同类别的后验概率，选择具有最高后验概率的类别作为预测结果。这一定理假设了特征之间是相互独立的，即给定类别的情况下，特征之间的条件独立性。尽管这个假设在现实中很难完全成立，但朴素贝叶斯分类器仍然在实际应用中表现良好。

二、朴素贝叶斯分类器

朴素贝叶斯分类器是一种基于贝叶斯定理和特征条件独立性假设的概率分类算法。它通过计算给定特征的条件概率，以及类别的先验概率，来预测新样本属于不同类别的后验概率，并选择具有最高后验概率的类别作为预测结果。（Naive Bayes classifier is a probabilistic classification algorithm based on Bayes theorem and characteristic condition independence assumption. By calculating the conditional probability of a given feature and the prior probability of a class, it predicts the posterior probability of a new sample belonging to a different class, and selects the category with the highest posterior probability as the prediction result.）

Lecture8:决策树与 KNN

一、参数化方法 parametric methods（有事先确定好的假设函数只是通过不断地对参数进行调整来对给定的数据进行拟合）

在参数化方法中，模型的形式和复杂度是事先确定的，但其中的参数是可以通过训练数据来学习和调整的。（In the parameterized method, the form and complexity of the model are determined in advance, but the parameters can be learned and adjusted through training data.）

参数化方法是一种统计建模技术，它对数据的函数形式或分布做出具体的假设。这些方法旨在从给定的数据中估计所假设模型的参数。参数化方法的特点是对整个数据总体使用固定数量的参数进行描述。（Parametric method is a statistical modeling technique that makes specific assumptions about the functional form or distribution of data. These methods aim to estimate the parameters of a hypothetical model from a given set of data. The parameterization method is characterized by the use of a fixed number of parameters to describe the whole data.）

在参数化方法中，模型由一组参数定义，目标是根据可用的数据估计这些参数。一旦参数估计完成，就可以利用模型对数据进行预测或进行推断。（In a parameterized approach, the model is defined by a set of parameters, and the goal is to estimate these parameters based on the available data. Once the parameter estimates are complete, the model can be used to predict or extrapolate the data.）

在实际应用中选择适当的模型是具有挑战性的。由于缺乏先验知识或真实模型的了解，我们需要根据现有数据和问题的特点来选择适合的模型。这也突出了参数化方法的一种实际情况，即模型的选择是一个具有挑战性的任务，需要根据具体情况进行权衡和判断。（Choosing the right model for a practical application can be challenging. Due to the lack of prior knowledge or understanding of the real model, we need to choose the appropriate model according to the existing data and characteristics of the problem. This also highlights one of the practical aspects of the parameterization approach, that is, model selection is a challenging task that requires trade-offs and judgments on a case-by-case basis.）

参数化方法假设数据遵循特定的概率分布，并且通过估计该分布的参数来建立模型。这意味着在参数化方法中，模型的形式和复杂度是事先确定的，通常由一组固定的参数决定。常见的参数化方法包括线性回归、逻辑回归和高斯混合模型等。参数化方法通常具有计算效率高和解释性强的优点，但对数据分布的假设限制了模型的灵活性和适应性。（The parameterized approach assumes that the data follows a specific probability distribution and builds the model by estimating the parameters of that distribution. This means that in a parameterized approach, the form and complexity of

the model are determined in advance, usually by a fixed set of parameters. Common parameterization methods include linear regression, logistic regression and the Gaussian mixture model. Parameterized methods usually have the advantages of high computational efficiency and strong interpretation, but the assumption of data distribution limits the flexibility and adaptability of the model.)

二、 非参数化方法 non-parametric methods（不需要假设函数）

非参数化方法是一种灵活的统计建模方法，它不依赖于事先对数据分布或模型形式的假设。相反，非参数化方法允许模型从数据中学习和适应数据的复杂性。(The nonparametric method is a flexible statistical modeling method that does not rely on prior assumptions about data distribution or model form. In contrast, a non-parametric approach allows the model to learn from and adapt to the complexity of the data.)

非参数化方法通常基于基本的数学工具和技术，如核密度估计、最近邻方法、决策树和基于树的方法（如随机森林和梯度提升树）等。这些方法能够根据数据的特点自动调整模型的复杂度，从而更好地适应数据的分布和结构。（Nonparametric methods are usually based on basic mathematical tools and techniques such as kernel density estimation, nearest neighbor methods, decision trees, and tree-based methods such as random forests and gradient lift trees. These methods can automatically adjust the complexity of the model according to the characteristics of the data so as to better adapt to the distribution and structure of the data.)

非参数化方法的优点在于其灵活性和适应性。它们能够适应各种类型的数据分布，包括非线性、多模态和异态数据。此外，非参数化方法通常不受特定参数数量的限制，因此可以更好地适应复杂的数据模式。(The advantage of the nonparametric method lies in its flexibility and adaptability. They can adapt to various types of data distribution, including nonlinear, multimodal, and heterogeneous data. In addition, non-parameterized methods are generally not limited by the number of specific parameters, so they can be better adapted to complex data schemas.)

非参数化方法的缺点在于计算复杂度较高，并且对数据量较大的情况可能存在较高的计算成本。此外，由于非参数化方法对数据的自由度较高，过拟合的风险也较大。因此，在应用非参数化方法时，需要谨慎地选择合适的模型和调整参数，以避免过度拟合和模型复杂度不当的问题。(The disadvantage of the non-parametric method lies in its high computational complexity and high computational cost when the data volume is large. In addition, due to the high degree of freedom of data for non-parametric methods, the risk of overfitting is also greater. Therefore, when applying the non-parametric method, it is necessary to carefully select the appropriate model and adjust the parameters to avoid the problems of over-fitting and improper model complexity.)

非参数化方法则不对数据分布做出明确的假设，而是通过从数据中学习和适应模型的复杂性来建立模型。非参数化方法通常基于基本的数学工具和技术，如核密度估计、最近邻方法和决策树等。这使得非参数化方法更加灵活，能够适应各种类型的数据分布和结构。非参数化方法通常具有更高的计算复杂度和较大的自由度，但也更容易受到过拟合的影响。（Non-parametric methods make no explicit assumptions about data distribution, but build models by learning from the data and adapting to the complexity of the model. Nonparametric methods are usually based on basic mathematical tools and techniques, such as kernel density estimation, nearest neighbor methods, and decision trees. This makes non-parametric methods more flexible and adaptable to various types of data distribution and structure. Non-parametric methods usually have higher computational complexity and greater degrees of freedom, but are also more susceptible to overfitting.)

三、多模态分布

多模态分布（Multimodal Distribution）指的是一个随机变量的概率分布具有多个显著的峰值或模态。每个峰值对应着分布中的一个集中区域，表示该区域内数据出现的概率较高。这些峰值之间可以存在空隙或低概率区域。

与单一模态的分布相比，多模态分布具有更多的局部特征，能够更好地反映数据中的多样性和复杂性。多模态分布经常出现在实际问题中，例如人群的身高分布、商品价格的分布等。每个峰值对应着分布中的一个集中区域，可能代表了不同的数据子群体或现象。(A Multimodal Distribution refers to a probability distribution of a random variable with multiple significant peaks or modes. Each peak corresponds to a concentrated region in the distribution, indicating a high probability of data occurrence in this region. There can be gaps or areas of low probability between these peaks. Compared

with the single mode of distribution, multimodal distribution with more local characteristics, can better reflect the diversity and complexity of the data. Multimodal distribution often appears in practical problems, such as the height distribution of people and the distribution of commodity prices. Each peak corresponds to a concentrated area in the distribution and may represent a different data subpopulation or phenomenon.）

四、决策树

决策树是一种基于树状结构进行分类和回归的机器学习方法。它通过对输入特征的一系列判断来逐步分割数据集，并生成一棵树形结构，其中每个内部节点表示一个特征判断，每个叶节点表示一个类别或回归值。（Decision tree is a kind of machine learning method for classification and regression based on tree structure. It gradually segments the data set by making a series of judgments on input features and generates a tree-like structure in which each internal node represents a feature judgment and each leaf node represents a category or regression value.）

1. 构建步骤

a) 特征选择：根据某个度量指标（如信息增益、基尼不纯度等），选择最佳的特征作为当前节点的判断条件。（Feature selection: According to some metric index (such as information gain, Gini impurity, etc.), the best feature is selected as the judgment condition of the current node.）

b) 样本划分：根据选定的特征将数据集划分为不同的子集，每个子集对应一个分支路径。（Sample partitioning: The data set is divided into different subsets according to the selected characteristics, and each subset corresponds to a branch path.）

c) 递归构建：对每个子集，重复步骤 1 和步骤 2，直到满足停止条件（如达到最大树深度、节点包含的样本数小于阈值等）。（Recursive construction: For each subset, repeat steps 1 and 2 until stopping conditions are met (e.g., the maximum tree depth is reached, the number of samples the node contains is less than the threshold, etc.).）

d) 叶节点标记：将最后生成的叶节点标记为相应的类别或回归值。（Leaf node labeling: The last generated leaf node is labeled with the corresponding category or regression value.）

2. 特征选择：

基尼不纯度：基尼不纯度是一种衡量节点纯度的指标，用于评估决策树节点的划分能力。

计算方法

1. 假设有 K 个类别，节点中包含的样本总数为 N。
2. 对于节点上的每个类别 i，计算其在节点中的概率（即该类别样本数除以总样本数），表示为 p(i)。
3. 基尼不纯度的计算公式为：Gini = 1 - Σ(p(i)^2)，其中 Σ 表示对所有类别求和。
4. 最终得到的基尼不纯度值越小，表示节点的纯度越高。

当基尼不纯度较小时，节点中的样本更倾向于属于同一类别，这意味着节点的划分更加准确，纯度更高。在决策树的构建过程中，选择基尼不纯度最小的特征进行划分可以使得子节点的纯度最大化，进而提高模型的准确性。（When Gini impurity is low, the samples in the nodes are more likely to belong to the same category, which means the nodes are more accurately classified and have higher purity. In the process of decision tree construction, choosing the features with the least Gini impurity to divide can maximize the purity of the child nodes and improve the accuracy of the model.）

基尼不纯度的最小值为 0，表示节点中的样本全部属于同一类别，即节点完全纯净。当基尼不纯度较小但不为 0 时，节点中可能存在一定比例的混杂样本，但整体上节点的纯度仍然较高。(The minimum Gini impurity value is 0, indicating that all the samples in the node belong to the same category, that is, the node is completely pure. When the Gini impurity is small but not zero, there may be a certain percentage of mixed samples in the node, but the purity of the node is still high on the whole.)

基尼不纯度越小越纯，决策树算法的目标是通过选择最小化基尼不纯度的划分来构建纯度更高的节点和子节点，从而提高模型的分类准确性。（The smaller the Gini impurity, the purest. The goal of the decision tree algorithm is to construct nodes and child nodes with higher purity by selecting the division that minimizes Gini impurity, so as to improve the classification accuracy of the model.）

熵（Entropy）是信息论中用来衡量不确定性的概念。在决策树中，熵用来衡量节点中样本的混乱程度或不确定性。（Entropy

is a concept used to measure uncertainty in information theory. In decision trees, entropy is used to measure the degree of confusion or uncertainty of samples in nodes)

熵的计算公式为：

$E = -\sum(p_i * \log2(p_i))$

其中，p_i 表示样本属于第 i 个类别的概率。熵的取值范围是[0, 1]，当样本属于同一类别时，熵为 0，表示节点纯度最高；当样本均匀分布在不同类别上时，熵为 1，表示节点纯度最低。

在决策树的节点特征选择中，通常使用信息增益（Information Gain）或基尼指数（Gini Index）来衡量特征的重要性，它们都与熵有关。信息增益是指特征划分前后熵的变化，选择信息增益最大的特征作为节点划分的依据；基尼指数是指节点的基尼不纯度，选择基尼指数最小的特征作为节点划分的依据。这些指标的计算都涉及到熵的概念，通过最小化熵或最大化信息增益或基尼指数，决策树可以构建出更加纯净的节点和更好的划分规则。

信息增益（Information Gain）是用于决策树节点特征选择的一个指标，它衡量了在使用某个特征进行划分后，数据集的熵减少了多少，即对分类任务的贡献程度。(Information Gain is an index used for feature selection of decision tree nodes. It measures how much entropy of data set is reduced after a certain feature is used for division, that is, how much contribution it makes to classification tasks.)

信息增益的计算公式为：

$\text{Information Gain}(D, A) = \text{Entropy}(D) - \sum((|D_v|/|D|) * \text{Entropy}(D_v))$

其中，D 表示当前节点的数据集，A 表示待选择的特征，D_v 表示在特征 A 上取值为 v 的子数据集，|D|表示数据集 D 的样本总数，|D_v|表示子数据集 D_v 的样本总数，Entropy(D)表示数据集 D 的熵，Entropy(D_v)表示子数据集 D_v 的熵。信息增益的值越大，表示特征 A 对分类任务的贡献越大，选择具有最大信息增益的特征作为节点的划分依据，能够使得决策树更快地将样本分成不同的类别。信息增益方法更偏向于选择具有较多取值的特征，可能会对具有较少取值但分类能力较强的特征不够敏感。因此，在实际应用中，可以结合其他特征选择的方法来综合考虑特征的重要性。

3. 从一个决策树中，我们可以获得以下信息：
    1. 分类规则：决策树通过一系列的节点和分支构建了一棵树形结构，每个节点代表一个特征，每个分支代表该特征的取值，最终的叶节点代表一个分类标签。通过遵循树的路径，我们可以根据输入的特征值来确定样本的分类。(Classification rule: The decision tree constructs a tree-like structure through a series of nodes and branches. Each node represents a feature, each branch represents the value of the feature, and the final leaf node represents a classification label. By following the path of the tree, we can determine the classification of samples based on the input eigenvalues.)
    2. 特征重要性：决策树可以计算特征的重要性，即通过特征在树中的使用频率或其对分类结果的贡献程度来衡量。较高的重要性表示该特征对于分类决策具有较大的影响力。(Feature importance: Decision trees can calculate the importance of features, which can be measured by the frequency of feature used in the tree or its contribution to classification results. A higher importance indicates that the feature has a greater influence on classification decision.)
    3. 决策路径：对于一个输入样本，我们可以通过决策树的分支路径来查看它是如何被分类的。从根节点开始，按照特征的取值逐步向下遍历树的分支，直到到达叶节点得到最终的分类结果。(Decision path: For an input sample, we can look at the branch path of the decision tree to see how it is classified. Starting from the root node, it gradually traverses the branches of the tree according to the value of features until it reaches the leaf node to get the final classification result.)
    4. 样本划分：决策树根据特征的取值将样本划分为不同的子集，每个子集对应一个叶节点。通过观察样本在不同叶节点上的分布，我们可以了解哪些特征值更能够将样本划分为不同的类别。(Sample partition: The decision tree divides the sample into different subsets according to the value of the feature, and each subset corresponds to a leaf node. By observing the distribution of samples on different leaf nodes, we can learn which eigenvalues are better able to classify samples into different categories.)
    5. 树的结构：决策树以树形结构表示了分类规则和特征之间的关系。我们可以查看树的层级结构、节点的连接关

系以及分支的条件，以深入理解分类的过程和决策的依据。(Tree structure: Decision tree represents the relationship between classification rules and features with tree structure. We can look at the hierarchy of trees, the connection relationships of nodes, and the conditions of branches to gain insight into the classification process and the basis for decisions.)

4. CART（Classification and Regression Trees）

是一种基于决策树的机器学习算法，既可以用于分类问题，也可以用于回归问题。CART 算法通过递归地将数据集划分为更小的子集，并在每个子集上构建一个决策树节点，从而生成一棵二叉树。(Is a machine learning algorithm based on decision tree, which can be used for both classification problems and regression problems. CART algorithms generate a binary tree by recursively dividing the data set into smaller subsets and building a decision tree node on each subset.)

5.决策树的超参数（防止过拟合）

决策树的超参数是在构建决策树时需要手动设置的参数，它们控制着决策树的结构和拟合程度。以下是一些常见的决策树的超参数：

a) 最大深度（max_depth）：决策树允许的最大深度，用于控制决策树的复杂度和过拟合风险。(The maximum allowable depth of the decision tree, which is used to control the complexity and overfitting risk of the decision tree.)

b) 最小样本拆分（min_samples_split）：一个内部节点在拆分之前所需的最小样本数。控制决策树的生长方式，避免过度细分。(The minimum number of samples required for an internal node before splitting. Control the way the decision tree grows to avoid excessive subdivision.)

c) 最小样本叶节点（min_samples_leaf）：一个叶节点所需的最小样本数。控制叶节点的最小大小，避免过拟合和生成过多的叶节点。(The minimum number of samples required for a leaf node. Control the minimum size of leaf nodes to avoid overfitting and generating too many leaf nodes.)

d) 特征选择标准（criterion）：用于衡量特征的质量和选择最佳拆分的标准。常见的标准包括基尼系数（gini）和信息增益（entropy）。(Used to measure the quality of features and to select the best split criteria. Common criteria include the gini coefficient and entropy.)

e) 分割特征选择数量（max_features）：用于拆分的特征数量。可以指定具体数量或使用默认值（如平方根或对数）。(The number of features used for splitting. You can specify an amount or use a default value (such as square root or logarithm).)

f) 类别权重（class_weight）：用于处理类别不平衡问题，给予不同类别不同的权重。(It is used to deal with the imbalance of categories and give different weights to different categories.)

6. 决策树的不稳定性 instability

a) 数据的微小变化可能导致不同的划分：决策树的构建过程中，根据选择的特征和划分标准来进行节点的划分。然而，数据的微小变化可能导致不同的划分结果，尤其是当数据特征较多或决策树较深时。(Small changes in data may lead to different divisions: During the construction of decision trees, nodes are divided according to selected characteristics and division criteria. However, small changes in the data can lead to different partitioning results, especially if the data has more characteristics or the decision tree is deep.)

b) 数据中的噪声和离群点：决策树对于噪声和离群点比较敏感。在数据中存在噪声或离群点时，决策树可能会受到这些异常数据的影响，导致不稳定的模型结果。(Noise and outliers in data: Decision trees are sensitive to noise and outliers. When there is noise or outliers in the data, the decision tree may be affected by these abnormal data, resulting in unstable model results.)

c) 训练集的变化：决策树的构建是基于训练集的，如果训练集的样本或特征发生变化，决策树的结构和预测结果可能会发生改变。(The construction of the decision tree is based on the training set. If the samples or features of the training set change, the structure and prediction results of the decision tree may change.)

d) 特征选择的随机性：决策树在每个节点选择划分特征时，通常会考虑多个候选特征，并基于某种准则选择最佳特征。由于特征选择的随机性，即使在相同的数据集上多次训练，也可能得到不同的决策树结构。(When

choosing features for each node, the decision tree usually considers multiple candidate features and selects the best features based on some criteria. Because of the randomness of feature selection, different decision tree structures may be obtained even after repeated training on the same data set.)

7. 决策树解决回归问题

特征分割是决策树构建过程中的关键步骤，它决定了每个节点如何根据特征的取值进行划分。(Feature segmentation is a key step in decision tree construction, which determines how to divide each node according to the value of feature.)

下面是进行特征分割的一般步骤：

1. 选择划分准则：首先需要选择一个合适的划分准则来衡量特征的分割效果。常见的划分准则包括基尼不纯度（Gini Impurity）和信息增益（Information Gain）。(Selection of partitioning criteria: Firstly, an appropriate partitioning criterion should be selected to measure the segmentation effect of features. Common zoning criteria include Gini Impurity and Information Gain.)
2. 计算划分准则的值：针对每个特征，根据选择的划分准则计算其在当前节点的值。基尼不纯度可以通过计算每个类别的概率的平方和的补（1 减去平方和）来获得，而信息增益则是当前节点的熵减去根据特征进行划分后的加权平均熵。(Calculate the value of partitioning criteria: For each feature, calculate its value at the current node according to the selected partitioning criteria. Gini impurity can be obtained by calculating the complement of the sum of squares (1 minus the sum of squares) of the probabilities for each category, and the information gain is the entropy of the current node minus the weighted average entropy divided by features.
3. 尝试所有可能的特征分割：对于每个特征，将其可能的取值作为划分点，计算划分准则的值。对于离散特征，可以尝试每个取值作为划分点；对于连续特征，可以尝试不同的阈值作为划分点。(Try all possible feature segmentation: for each feature, take its possible value as the partition point and calculate the value of the partition criterion. For discrete features, each value can be tried as a partition point. For continuous features, different thresholds can be tried as partition points.)
4. 选择最佳的特征分割：根据划分准则的值，选择使得划分准则最小或最大化的特征分割点。具体的选择方法取决于所使用的划分准则和目标。(Select the best feature segmentation: Select the feature segmentation points that minimize or maximize the partitioning criteria according to the value of the partitioning criteria. The exact selection method depends on the partitioning criteria and objectives used.)
5. 执行特征分割：根据选择的最佳特征分割点，将当前节点的数据集划分成不同的子集。每个子集将成为下一层决策树的节点，继续进行后续的特征分割过程。(Perform feature segmentation: According to the selected best feature segmentation point, the data set of the current node is divided into different subsets. Each subset will become a node of the decision tree at the next level and continue the subsequent feature segmentation process.)

五、KNN（K 近邻算法）

对于分类问题，给定一个未知样本，KNN 算法会计算该样本与训练集中所有样本之间的距离，并选择与该样本最近的 K 个训练样本。然后，通过多数表决的方式，将这 K 个训练样本中出现最频繁的类别作为该未知样本的预测类别。(For the classification problem, given an unknown sample, the KNN algorithm will calculate the distance between the sample and all samples in the training set, and select the K training samples closest to the sample. Then, by majority vote, the category that appears most frequently among the K training samples is taken as the prediction category of the unknown sample.

对于回归问题，KNN 算法会计算未知样本与训练集中所有样本之间的距离，并选择与该样本最近的 K 个训练样本。然后，根据这 K 个训练样本的标签值，通过平均值或加权平均值的方式来预测未知样本的目标值。(For regression problems, KNN algorithm will calculate the distance between the unknown sample and all samples in the training set, and select the K training samples closest to the sample. Then, according to the label values of the K training samples, the target values of unknown samples are predicted by means of average or weighted average.)

KNN 算法的关键参数是 K 值，它决定了邻居的数量。较小的 K 值会使模型更加敏感，容易受到噪声的影响，而较大的 K 值会使模型更加平滑，可能忽略掉类别或目标值的局部特征。(The key parameter of KNN algorithm is K value, which determines the number of neighbors. A smaller K value will make the model more sensitive and susceptible to noise, while

a larger K value will make the model smoother and may ignore local features of the category or target value.）

K 值选择方法

1. 经验法则：根据经验规则选择一个合适的 K 值。一般来说，较小的 K 值会使模型更加敏感，适用于复杂的数据集，而较大的 K 值会使模型更加平滑，适用于简单的数据集。
2. 交叉验证：使用交叉验证来评估不同 K 值下模型的性能，并选择性能最好的 K 值。常用的交叉验证方法包括 K 折交叉验证和留一交叉验证。
3. 网格搜索：通过遍历一定范围内的 K 值，利用交叉验证或其他性能评估指标来选择最佳的 K 值。网格搜索可以通过自动化地尝试多个 K 值来寻找最优解。
4. 考虑样本分布：根据样本的分布情况来选择 K 值。如果样本分布均匀，则选择较小的 K 值可能更合适；如果样本分布不均匀，则选择较大的 K 值可能更合适。

选择 K 值时需要综合考虑数据集的特点和任务需求。较小的 K 值可以捕捉到局部特征，但可能对噪声敏感；较大的 K 值可以平滑预测结果，但可能忽略掉局部特征。

Lecture9: 集成学习（Ensemble learning）

一、投票分类器(Voting Classifier)

概念：投票分类器（Voting Classifier）是一种集成学习方法，它将多个独立的分类器组合起来进行预测。投票分类器属于集成学习的一种，旨在通过聚合多个模型的预测结果来提高整体性能。(Voting Classifier is an ensemble learning method, which combines multiple independent classifiers to make predictions. Voting classifier is a kind of ensemble learning, which aims to improve the overall performance by aggregating the predicted results of multiple models.)

在投票分类器中，每个独立的分类器，也称为基分类器，独立地对输入数据进行分类预测。然后，投票分类器使用一种投票策略将这些预测结果组合起来，确定最终的预测结果。(In a voting classifier, each independent classifier, also called a base classifier, independently makes classification predictions on the input data. The vote classifier then combines these predictions using a voting strategy to determine the final prediction.)

二、投票策略

1. 硬投票：在硬投票中，基分类器的预测类别标签进行统计，选择出现次数最多的类别标签作为最终预测结果。适用于类别标签是离散的、无需概率估计的分类问题。(Hard voting: In hard voting, the prediction category labels of the base classifier are counted, and the category labels that occur the most times are selected as the final prediction result. It is applicable to the classification problem that the category labels are discrete and do not require probability estimation.)
2. 软投票：在软投票中，不仅考虑类别标签，还考虑基分类器对于每个类别的预测概率或置信度。将基分类器对每个类别的概率进行平均或组合，选择平均概率或置信度最高的类别作为最终预测结果。软投票适用于基分类器能够提供概率估计的情况。(Soft voting: In soft voting, not only the category label is considered, but also the predicted probability or confidence of the base classifier for each category. The probability of each category is averaged or combined by the base classifier, and the category with the highest average probability or confidence is selected as the final prediction result. Soft voting is suitable for situations where the base classifier can provide a probability estimate.)

投票分类器可以使用各种类型的基分类器构建，例如决策树、逻辑回归、支持向量机或神经网络。通过组合多个分类器的预测结果，投票分类器可以利用各个模型的差异性和互补优势，提高整体性能并得到更可靠的预测结果。(Voting classifiers can be built using various types of base classifiers, such as decision trees, logistic regression, support vector machines, or neural networks. By combining the prediction results of multiple classifiers, voting classifiers can take advantage of the differences and complementary advantages of each model to improve the overall performance and obtain more reliable prediction results.)

值得注意的是，投票分类器的有效性取决于基分类器的多样性和准确性。当基分类器具有多样性且表现良好时，投票分类器有潜力优于任何单个基分类器，并提供更可靠的预测结果。(It is important to note that the effectiveness of voting classifiers depends on the diversity and accuracy of base classifiers. When there is a diversity of base classifiers and they perform well, voting classifiers have the potential to outperform any single base classifier and provide more reliable predictions.)

三、 其他合并分类器的方法（集成学习方法）

- 平均法（Averaging）：将多个分类器的预测结果进行平均，可以是硬平均（对分类标签进行统计平均）或软平均（对概率或置信度进行平均）。平均法适用于基分类器相对独立且性能相近的情况。(Averaging: Averaging the predictions of multiple classifiers, either hard averaging (statistical averaging of classification labels) or soft averaging (Averaging probability or confidence). The average method is applicable when the base classifier is relatively independent and has similar performance.)

- 加权平均法（Weighted Averaging）：与平均法类似，但对不同的基分类器赋予不同的权重。权重可以根据基分类器的准确性或其他评估指标进行设置，以更加重视性能较好的分类器。(Weighted Averaging: similar to the averaging method, but assigns different weights to different base classifiers. The weights can be set

according to the accuracy of the base classifier or other evaluation indicators to pay more attention to the classifiers with better performance.)

- 投票加权法（Voting with Weights）：类似于硬投票，但为每个基分类器赋予不同的权重。权重可以根据基分类器的准确性或其他评估指标进行设置，以更有针对性地考虑性能较好的分类器。(Voting with Weights: similar to hard voting, but different weights are assigned to each base classifier. The weights can be set according to the accuracy of the base classifier or other evaluation indicators to consider classifiers with better performance more specifically.)

- 堆叠法（Stacking）：堆叠法通过构建一个元分类器来整合多个基分类器的预测结果。首先，基分类器对数据进行预测，然后将这些预测结果作为输入特征，与原始特征一起输入给元分类器进行最终的预测。（Stacking method (Stacking): Stacking integrates the predictions of multiple base classifiers by constructing a meta-classifier. First, the base classifier makes predictions on the data, and then feeds these predictions as input features, together with the original features, to the meta-classifier for final predictions.)

- 提升法（Boosting）：提升法通过串行训练多个基分类器，每个分类器都尝试纠正前一个分类器的错误。最终的预测结果是基于所有分类器的加权组合。著名的提升算法包括 Adaboost 和 Gradient Boosting。（Boosting: Boosting trains multiple base classifiers in serial, each trying to correct the errors of the previous one. The final prediction is based on a weighted combination of all classifiers. Well-known lifting algorithms include Adaboost and Gradient Boosting.)

四、 Bagging 和 Pasting（都是用于集成学习的方法，提高模型的泛化能力）

共同点：

它们都是通过生成多个基学习器，它们使用的算法和参数设置通常是相同的，每个基学习器使用独立的采样集进行训练，并最后组合它们的预测结果来得到最终的集成预测结果。两种方法都旨在减少模型的方差，提高模型的泛化能力和稳定性。（They all work by generating multiple base learners, often using the same algorithm and parameter Settings, each of which is trained using an independent set of samples, and finally combining their predictions to get the final integrated prediction. Both methods aim to reduce the variance of the model and improve the generalization ability and stability of the model.)

总结起来，Bagging 和 Pasting 的主要区别在于样本选择的方式：Bagging 使用自助采样，样本有放回地被选择；而 Pasting 使用无放回采样，每个样本只会被选择一次。(To sum up, the main difference between Bagging and Pasting lies in the way samples are selected. Bagging uses self-sampling, in which samples are selected in return. Pasting uses no-place sampling, in which each sample is selected only once.)

1. Bagging 使用自助采样（bootstrap sampling）的方法，从原始数据集中有放回地随机采样生成多个采样集。然后，使用每个采样集独立训练一个基学习器，最后将它们的预测结果进行平均或投票，得到最终的集成预测结果。Bagging 的优势在于减少方差、提高模型的稳定性和泛化能力，尤其适用于大型数据集。（Bagging uses the method of bootstrap sampling to generate multiple sample sets from the original data set by randomly sampling them back and forth. Then, each sample set is used to train a base learner independently. Finally, their prediction results are averaged or voted to get the final integrated prediction results. Bagging has the advantage of reducing variance and improving model stability and generalization, especially for large data sets.)
2. Pasting 使用无放回采样（without replacement sampling）的方式从原始数据集中随机选择样本生成多个采样集。每个采样集都用于训练一个基学习器，最后将它们的预测结果进行平均或投票来生成最终的集成预测结

果。Pasting 适用于数据集较小的情况，能够充分利用有限的样本数据。(Pasting generates multiple sampling sets by randomly selecting samples from the original data set without replacement sampling. Each sample set is used to train a base learner, and their predictions are then averaged or voted to produce the final integrated prediction. Pasting is suitable for small data sets and can make full use of limited sample data.)

3. 区别
   - Bagging（自助聚集法）：Bagging 使用自助采样（bootstrap sampling）的方式选择样本。在每次采样中，样本有放回地被选择，因此同一个样本可能在多个采样集中出现。这意味着某些样本可能被重复选取，而其他样本可能在某些采样集中没有被选择到。(Bagging: Bagging uses bootstrap sampling to select samples. In each sample, the sample is selected in the back, so the same sample may appear in multiple sample sets. This means that some samples may be selected repeatedly, while others may not be selected in some sample sets.)

     优点：自助采样的主要优点是可以利用采样集进行模型训练和性能评估，无需额外的验证集。由于每个采样集都是从原始数据集中独立随机抽取的，因此可以生成多个训练集和验证集的组合，用于训练多个模型并评估它们的性能。此外，袋外样本可以用作对模型的泛化能力进行估计。(The main advantage of self-sampling is that the sample set can be used for model training and performance evaluation without additional verification set. Because each sample set is independently and randomly drawn from the original data set, a combination of multiple training sets and validation sets can be generated for training multiple models and evaluating their performance. In addition, out-of-pocket samples can be used to estimate the generalization ability of the model.)

   - Pasting（粘贴法）：Pasting 使用无放回采样（without replacement sampling）的方式选择样本。在每次采样中，样本被选择后不放回，因此每个样本在所有采样集中只出现一次，没有重复选择的情况。(Pasting: Pasting selects samples without replacement sampling. In each sampling, samples are not returned after being selected, so each sample only appears once in all sampling sets without repeated selection.)

     优点：无放回采样的优点在于确保每个样本只出现一次，避免了样本重复和样本选择偏差。它还可以更好地反映真实数据的分布特征，并提供更可靠的模型评估结果。(The advantage of sampling without retracting is to ensure that each sample occurs only once, avoiding sample duplication and sample selection bias. It can also better reflect the distribution characteristics of real data and provide more reliable model evaluation results.)

     缺点：然而，与自助采样相比，无放回采样需要更大的样本量才能充分利用数据。由于每次采样后样本会被移除，因此采样集的样本数较少，可能导致训练的模型方差较高。(However, sampling without retracting requires a larger sample size than self-sampling to make full use of the data. Since the samples will be removed after each sampling, the sample number of the sample set is small, which may lead to a high variance of the trained model.)

4. 应用/优点：
   - 减少过拟合：通过训练多个基学习器并将它们的预测结果进行集成，可以减少模型的方差，降低过拟合的风险，提高模型的泛化能力。(Reduce overfitting: By training multiple base learners and integrating their predicted results, the variance of the model can be reduced, the risk of overfitting can be reduced, and the generalization ability of the model can be improved.)
   - 提高模型稳定性：集成多个基学习器的预测结果可以降低模型的方差，使整体模型对于输入数据的变化具有更好的稳定性。(Improve model stability: Integrating the prediction results of multiple base learners can reduce the variance of the model and make the overall model have better stability against the changes of input data.)

- 改善分类或回归的准确性：通过结合多个基学习器的预测结果，集成模型能够在分类或回归任务中取得更准确的结果，尤其在处理复杂问题或存在噪声的数据时效果更显著。（Improved accuracy of classification or regression: By combining predictions from multiple base learners, integrated models can achieve more accurate results in classification or regression tasks, especially when dealing with complex problems or noisy data.）
- 处理大规模数据集：Bagging 和 Pasting 都能够并行地训练多个基学习器，适用于处理大规模数据集的场景。（Processing large data sets: Bagging and Pasting are both capable of training multiple base learners in parallel and are suitable for scenarios dealing with large data sets.）

5. 它们可以得到更合理决策边界的原因

Bagging 和 Pasting 可以得到更合理的决策边界，这是因为它们通过采样集的多样性和集成预测的方式来减少过拟合风险。采样集的多样性使得每个基学习器学到不同的特征和模式，产生不同的决策边界，而集成预测可以融合这些边界，得到更综合和合理的结果。此外，通过投票或平均基学习器的预测结果，可以平衡各自的决策边界，提高整体模型的鲁棒性和泛化能力。因此， Bagging 和 Pasting 相较于单一分类器能够产生更合理的决策边界。（Bagging and Pasting can lead to more reasonable decision boundaries because they reduce the risk of overfitting by diversifying the sample set and integrating predictions. The diversity of the sample set makes each base learner learn different features and patterns and produce different decision boundaries. Ensemble prediction can fuse these boundaries to get more comprehensive and reasonable results. In addition, the predicted results of voting or average-based learning can balance the respective decision boundaries and improve the robustness and generalization ability of the overall model. Therefore, Bagging and Pasting can produce more reasonable decision boundaries than a single classifier.）

五、 袋外估计（out-of-bag evaluation, OOB）

OOB 评估是一种有效且方便的评估随机森林性能的方法，通过使用袋外样本作为验证集，可以获得对模型泛化能力的可靠估计。（OOB evaluation is an effective and convenient method to evaluate the performance of random forests. By using out-of-pocket samples as verification sets, reliable estimates of the model's generalization ability can be obtained.）

袋外样本：袋外样本（Out-of-Bag Samples）是指在自助采样（Bootstrap Sampling）中，未被任何一个采样集选中的样本。由于自助采样是有放回的随机采样，每次采样中大约有 36.8%的样本未被选中，这些未被选中的样本就是袋外样本。（Out-of-Bag Samples refer to the samples that are not selected by any Sampling set in Bootstrap Sampling. Since self-sampling is random sampling with retractions, about 36.8% of samples in each sampling are not selected, and these unselected samples are out-of-pocket samples.）

每个采样集都是从原始数据集中独立随机抽取的，因此每个样本有大约 63.2%的概率被选中至少一次，即成为某个采样集的一部分。因此，剩余约 36.8%的样本未被任何一个采样集选中，这些样本就是袋外样本。（Each sample set is selected independently and randomly from the original data set, so each sample has an approximate 63.2% chance of being selected at least once, that is, becoming part of a sample set. Therefore, about 36.8% of the remaining samples are not selected by any sample set, and these samples are out-of-pocket samples.）

- OOB（Out-of-bag）评估是随机森林中一种特殊的交叉验证技术。在随机森林中，每个决策树的训练样本是通过自助采样（bootstrap sampling）得到的，即从原始训练数据中有放回地随机抽取样本。由于自助采样的性质，部分样本在每棵树的训练集中没有被选中，这些未被选中的样本就构成了该树的袋外样本。(OOB (Out-of-bag) evaluation is a special cross-validation technique in random forest. In the random forest, the training samples of each decision tree are obtained by bootstrap sampling, that is, samples are randomly selected from the original training data. Due to the nature of self-sampling, some samples are not selected in the training set of each tree, and these unselected samples constitute the out-of-pocket samples of the tree.)

- 在训练每棵树时，可以使用袋外样本作为验证集来评估树的性能。这样，每棵树都有一个与其独立的验证集，

可以用来测量它对未见过的样本的预测准确性。在每棵树的训练过程中，使用袋外样本对树进行验证并计算其准确率或其他性能指标。(As each tree is trained, out-of-pocket samples can be used as verification sets to evaluate the performance of the tree. In this way, each tree has a separate verification set that can be used to measure its prediction accuracy for samples it has not seen before. During the training of each tree, out-of-pocket samples were used to verify the tree and calculate its accuracy or other performance indicators.)

- 最终，随机森林的整体性能可以通过集成所有树的袋外样本的预测结果来评估。对于分类问题，可以通过投票或取平均的方式获得最终的集成预测结果。对于回归问题，可以将每棵树的预测结果平均得到最终的预测值。(Ultimately, the overall performance of the random forest can be evaluated by integrating the predicted results of an out-of-pocket sample of all the trees. For classification problems, the final integrated prediction results can be obtained by voting or averaging. For regression problems, you can average the predicted results of each tree to get the final predicted value.)

- OOB 评估的优点在于它提供了对集成模型性能的无偏估计，无需额外的验证集或交叉验证。通过利用袋外样本作为验证集，可以充分利用数据并避免了验证集的选择和划分过程。此外， OOB 评估还可以用来进行模型选择，例如选择最优的树个数或其他超参数。(The advantage of OOB evaluation is that it provides an unbiased estimate of the performance of the integration model without additional verification sets or cross-validations. By using out-of-pocket samples as verification sets, the data can be fully utilized and the selection and division of verification sets can be avoided. In addition, OOB evaluation can be used for model selection, such as selecting the optimal number of trees or other hyperparameters.)

六、 Random pitches and Random subspaces

Random Patches 和 Random Subspaces 都是通过引入随机性来构建多个基分类器，以增加模型的多样性和泛化能力。它们的核心思想是通过限制训练数据的子集或特征子集来构建基分类器，从而避免过拟合并提高模型的稳定性和性能。(Both Random Patches and Random Subspaces are designed to build multiple base classifiers by introducing randomness, so as to increase the diversity and generalization ability of models. Their core idea is to build a base classifier by limiting the subset or feature subset of training data, so as to avoid over-fitting and improve the stability and performance of the model.)

- Random Patch 是一种基于随机采样的技术，它在训练过程中随机选择一部分特征和样本，构建基分类器。这意味着每个基分类器仅使用了原始特征空间和样本空间的一部分，而不是使用全部特征和样本。通过在随机选择的特征和样本子集上训练多个基分类器，Random Patch 可以引入多样性，从而提高集成模型的性能。(Random Patch is a technology based on random sampling. It randomly selects some features and samples in the training process to build a base classifier. This means that each base classifier uses only a portion of the original feature space and sample space, rather than all features and samples. By training multiple base classifiers on randomly selected features and sample subsets, Random Patch can introduce diversity, thus improving the performance of the integrated model.)
- Random Subspace 是一种基于随机特征选择的技术，它在训练过程中随机选择一部分特征，构建基分类器。与 Random Patch 不同的是， Random Subspace 仅针对特征空间进行随机选择，而保持全部样本。每个基分类器使用不同的特征子集进行训练，从而引入多样性。在测试时，所有基分类器都对样本进行预测，并通过投票或取平均值的方式得出最终的集成结果。(Random Subspace is a technology based on random feature selection. It randomly selects some features in the training process to build a base classifier. Different from Random Patch, Random Subspace only randomly select feature space and keeps all samples. Each base classifier is trained with a different subset of features to introduce diversity. In the test, all base classifiers predict the sample and get the final integration result by voting or averaging.)

七、 随机森林（Random Forest）

随机森林（Random Forest）是一种集成学习方法，用于构建一个强大的分类或回归模型。它结合了决策树的预测能力和随机性的特点。（Random Forest is an ensemble learning method used to build a powerful classification or regression model. It combines the prediction ability and randomness of the decision tree.）

随机森林的构建过程：

- 首先，从原始数据集中进行有放回的随机抽样（自助采样）得到一个采样集；然后，在每个节点划分时随机选择一部分特征子集；接下来，使用采样集和选定的特征子集构建决策树模型；重复以上步骤多次，构建多个决策树；最后，在分类任务中通过投票的方式，每个决策树投票选择最终的类别；在回归任务中通过取平均值的方式，每个决策树给出一个预测值，最终取平均得到集成的预测结果。通过自助采样和随机特征选择的随机性，随机森林能够提高模型的多样性和泛化能力，具有良好的性能和鲁棒性。（First, a sampling set is obtained by random sampling with retractions (self-sampling) from the original data set. Then, a subset of features is randomly selected during each node division. Next, a decision tree model is constructed by sampling set and selected feature subset. Repeat the above steps several times to construct multiple decision trees; Finally, in the classification task, each decision tree votes to select the final category. In the regression task, by means of averaging, each decision tree gives a predicted value, and finally, the integrated prediction result is obtained by averaging. Through self-sampling and randomness of random feature selection, the random forest can improve the diversity and generalization ability of the model, with good performance and robustness.）

随机森林集成多个分类器的过程：

- 随机森林集成多个分类器的过程包括从原始训练数据集中进行自助采样得到多个采样集，然后在每个采样集上随机选择特征子集构建决策树模型。每个决策树模型独立地对输入样本进行预测，并通过投票或平均的方式获得最终的分类结果。通过集成多个分类器，随机森林能够充分利用每个分类器的独立性和多样性，提高整体分类性能，减少过拟合，并具备更好的泛化能力。（The process of integrating multiple classifiers in the random forest includes self-sampling from the original training data set to obtain multiple sample sets, and then randomly selecting feature subsets on each sample set to build a decision tree model. Each decision tree model independently predicts the input samples and obtains the final classification result by voting or averaging. By integrating multiple classifiers, the random forest can make full use of the independence and diversity of each classifier, improve the overall classification performance, reduce overfitting, and have better generalization ability.）

八、 特征重要性

1. 特征重要性可以通过多种方式计算，其中一种常用的方法是基于随机森林的特征重要性评估。（Feature importance can be calculated in a variety of ways, one of the common methods is the random forest-based feature importance assessment.）

   特征重要性的计算步骤：

   a) 在构建随机森林时，记录每个特征在各个决策树中被用作划分的次数。（When building a random forest, record the number of times each feature is used as a partition in each decision tree.）
   b) 对于每个特征，将其在所有决策树中被用作划分的次数累加起来，得到该特征的总划分次数。（For each feature, the number of times it is used as partition in all decision trees is added up to get the total number of partition of the feature.）
   c) 将每个特征的总划分次数归一化，得到特征的相对重要性。（The relative importance of each feature is obtained by normalizing the total number of partition times of each feature.）

   越好的特征能够将数据集划分成越纯的子集，也就说明该特征越重要。通过计算特征在个个决策树中被用作划分的次数可以衡量各个特征的重要性。（The better the feature, the purer the subset of the data set, the more important the feature. The importance of each feature can be measured by counting the number of times the

feature is used as a partition in the individual decision tree.）

2. 衡量特征重要性的方法

- 随机森林中的特征重要性：在随机森林算法中，可以通过统计特征在决策树中被用作划分的次数来评估特征的重要性。这种方法基于特征被选为划分特征的频率，认为频繁被选中的特征对模型的预测能力更为关键。（Importance of features in the random forest: In random forest algorithms, the importance of features can be assessed by counting the number of times they are used as partitions in the decision tree. This method is based on the frequency of features being selected as partitioning features, and the frequently selected features are considered to be more critical to the predictive ability of the model.）
- 基于模型权重或系数的特征重要性：在一些线性模型或正则化模型中，特征的权重或系数可以用来衡量其重要性。较大的权重或系数表示特征在模型中对输出结果的影响更大，因此可以被视为重要特征。（Feature importance based on model weight or coefficient: In some linear or regularized models, the weight or coefficient of a feature can be used to measure its importance. A larger weight or coefficient indicates that features in the model have a greater impact on the output results, so they can be regarded as important features.）
- 方差相关性：特征的方差可以用来评估其重要性。如果某个特征的方差较大，说明该特征在数据中具有较大的变化范围，可能对预测结果有更大的贡献。（Variance correlation: The variance of a feature can be used to assess its importance. If the variance of a feature is large, it indicates that the feature has a large range of variation in the data and may contribute more to the prediction results.）
- 互信息或信息增益：这些指标衡量特征与目标变量之间的相关性或信息增益。特征与目标变量之间的高相关性或较大的信息增益表示该特征对于目标变量的解释能力较强，因此可以被认为是重要特征。（Mutual information or information gain: These measures measure the correlation or information gain between a feature and a target variable. The high correlation or large information gain between the feature and the target variable indicates that the feature has a strong explanatory ability for the target variable, so it can be considered an important feature.）
- 基于特征选择算法：特征选择算法可以根据不同的准则或算法选择最佳的特征子集。这些算法可能考虑特征之间的相关性、冗余性以及与目标变量之间的相关性，从而确定特征的重要性。（Feature selection algorithm: The feature selection algorithm can select the best feature subset according to different criteria or algorithms. These algorithms may consider the correlation between features, redundancy, and correlation with target variables to determine the importance of features.）

九、 Boost（集成学习中的提升法）中的 Adaboost 和 Gradient Boosting

1. Adaboost：Adaboost（AdaBoost）是一种集成学习（ensemble learning）方法，用于提升分类算法的性能。它通过反复训练一系列弱分类器（weak classifier）并将它们组合成一个强分类器（strong classifier）来提高分类准确率。（Adaboost (AdaBoost) is an ensemble learning method used to improve the performance of classification algorithms. It improves classification accuracy by repeatedly training a series of weak classifiers and combining them into a strong classifier.）

主要步骤：

a) 初始化样本权重：对于有 N 个样本的训练集，开始时给每个样本赋予相等的权重，即 1/N。（Initialize sample weight: For a training set with N samples, equal weight is assigned to each sample at the beginning, that is, 1/N.）

b) 迭代训练弱分类器：通过多次迭代训练，每一次迭代都会生成一个弱分类器。每个弱分类器根据当前样本权重进行训练，其中权重高的样本会得到更多的关注。（Iteration training weak classifier: Through multiple iteration training, each iteration will generate a weak classifier. Each weak classifier is trained according to the current sample weight, among which the samples with high weight will get more attention.）

c) 更新样本权重：根据上一轮的分类结果，将被错误分类的样本的权重增加，而被正确分类的样本的权重减少。

（Update the sample weight: According to the classification results of the last round, the weight of the wrongly classified samples will be increased, while that of the correctly classified samples will be reduced.）

d) 组合弱分类器：将所有的弱分类器按照一定的权重组合成一个强分类器。弱分类器的权重取决于其在训练过程中的准确率。（Combination of weak classifiers: All the weak classifiers according to a certain weight recombination into a strong classifier. The weight of weak classifier depends on its accuracy in the training process.）

e) 重复步骤 2 至步骤 4，直到达到预定的迭代次数或达到一定的性能指标。（Repeat Steps 2 through 4 until a predetermined number of iterations is reached or a certain performance target is reached.）

简要步骤:

Adaboost 是一种集成学习方法，通过反复训练一系列弱分类器，并根据它们的表现进行加权组合，构建一个强分类器。它通过调整样本权重来重点关注被错误分类的样本，以提高整体的分类性能。（Adaboost is an ensemble learning method that builds a strong classifier by repeatedly training a series of weak classifiers and combining them weighted according to their performance. It focuses on the misclassified samples by adjusting sample weights to improve the overall classification performance.）

Adaboost 通过将多个弱分类器的结果进行加权组合，生成一个能够更好地进行分类的强分类器。在每一轮迭代中，Adaboost 会调整样本的权重，使得下一轮迭代更关注被错误分类的样本，从而提高整体的分类性能。（Adaboost generates a strong classifier capable of better classification by combining the results of multiple weak classifiers in a weighted way. In each iteration, Adaboost will adjust the weight of samples, making the next iteration pay more attention to the misclassified samples, so as to improve the overall classification performance.）

Adaboost 的优点包括可以提高分类准确率，对于处理复杂的数据集和噪声数据具有较好的鲁棒性。然而，它也对异常值敏感，并且对于数据集中的噪声和错误标签可能会导致模型的过拟合。(The advantages of Adaboost include improved classification accuracy and good robustness for processing complex data sets and noisy data. However, it is also sensitive to outliers, and noise and mislabeling in the data set can lead to overfitting of the model.)

2. Gradient Boosting

Gradient Boosting 是一种集成学习方法，用于构建强大的预测模型。它通过迭代地训练一系列弱预测模型（通常是决策树），每一次迭代都根据前一次迭代的残差来调整模型的参数。（Gradient Boosting is an ensemble learning method used to build a strong prediction model. It trains a series of weak prediction models (usually decision trees) iteratively, with each iteration adjusting the model's parameters according to the residual of the previous iteration.）

Residual（残差）：预测值与真实值之间的差值（The difference between the predicted value and the true value.）

主要步骤:

a) 初始化模型：使用一个简单的预测模型（如常数）作为初始模型。(Initialize the model: Use a simple prediction model (such as a constant) as the initial model.)

b) 计算残差：用初始模型预测样本，并计算实际值与预测值之间的残差（差异）。(Calculate residuals: Predict the sample with the initial model and calculate the residuals (differences) between the actual and predicted values.)

c) 训练弱模型：使用残差作为目标变量，训练一个新的弱预测模型，它试图捕捉先前模型未能解释的残差部分。（Training Weak models: Using residuals as the target variable, training a new weak prediction model that attempts to capture parts of residuals that previous models failed to account for.）

d) 更新模型：将新的弱模型与之前的模型进行加权组合，产生一个更强大的模型。(Update model: The new weak model is weighted with the previous model to produce a more powerful model.)

e) 重复步骤 2 至步骤 4，直到达到预定的迭代次数或满足性能指标。（Repeat Steps 2 through 4 until the desired number of iterations is reached or performance indicators are met.）

最终，Gradient Boosting 通过迭代训练一系列弱模型，并将它们组合成一个强模型，以逐步减小预测误差。每一次迭代都试图修正前一次模型的错误，从而提高整体模型的性能。（Finally, Gradient Boosting iteratively trains a

series of weak models and combines them into a strong model to gradually reduce prediction errors. Each iteration attempts to correct the errors of the previous model to improve the performance of the overall model.）

Gradient Boosting 的优点包括能够处理复杂的非线性关系，对异常值具有较好的鲁棒性，以及能够自动进行特征选择。然而，它也对数据中的噪声和过拟合敏感，需要适当的参数调整和正则化来避免过拟合的问题。（The advantages of Gradient Boosting include being able to handle complex nonlinear relations, better robustness to outliers, and automatic feature selection. However, it is also sensitive to noise and overfitting in the data and requires proper parameter adjustment and regularization to avoid the problem of overfitting.）

GBRT(Gradient Boosted Regression Trees) GBRT 是 Gradient Boosting 在回归问题上的应用。

十、学习率 learning rate

学习率（Learning Rate）是机器学习算法中的一个重要超参数，它决定了每一步迭代中模型参数更新的幅度或步长大小。学习率的选择对于模型的收敛性和性能有重要影响。（Learning Rate is an important super parameter in machine learning algorithm, which determines the updating amplitude or step size of model parameters in each iteration. The choice of learning rate has an important effect on the convergence and performance of the model.）

较高的学习率可以加快模型的收敛速度，但可能会导致模型在最优点附近震荡或无法收敛。此时，模型可能会在参数空间中跳过最优点，并且可能无法取得最佳性能。（A higher learning rate can accelerate the convergence rate of the model, but it may cause the model to oscillate around the optimal point or fail to converge. At this point, the model may skip the best advantage in the parameter space and may not achieve the best performance.）

较低的学习率可以提高模型的稳定性，但可能需要更多的迭代次数才能达到收敛。如果学习率过低，模型可能会收敛到一个次优解。（A lower learning rate can improve the stability of the model, but it may require more iterations to reach convergence. If the learning rate is too low, the model may converge to a suboptimal solution.）

十一、堆叠法（Stacking）

1.Stacking 是一种集成学习方法，通过将多个基础模型（也称为初级学习器）的预测结果作为输入，再训练一个元模型（也称为次级学习器）来进行最终的预测。(Stacking is an integrated learning method that takes the predictions of multiple foundation models (also known as primary learners) as inputs and trains a metametter (also known as secondary learners) for final prediction.）（用训练集训练多个不同的基础模型，以不同基础模型在测试集上的预测结果作为元数据集特征，使用元数据集以及对应的真实标签对元元模型进行训练，最终元模型对测试集进行预测）

Stacking 的步骤如下：

a) 准备数据集：将原始训练数据集划分为多个不重叠的子集。(Prepare the data set: Divide the original training data set into multiple non-overlapping subsets.)
b) 构建基础模型：对于每个子集，训练多个不同的基础模型，可以使用不同的算法或参数配置。(Build the base model: For each subset, train multiple different base models, perhaps using different algorithms or parameter configurations.)
c) 生成预测结果：使用训练好的基础模型对剩余的未参与训练的数据进行预测。(Generate prediction results: Use the trained basic model to predict the remaining data not involved in training)
d) 构建元模型：将基础模型的预测结果作为输入特征，原始训练数据集的真实标签作为目标变量，训练一个元模型。(Building a metamodel: A metamodel is trained by taking the prediction result of the basic model as the input feature and the real label of the original training data set as the target variable.)
e) 预测：使用训练好的元模型对新样本进行预测。(Prediction: Use a trained metamodel to predict new samples.)

优缺点：

Stacking 是一种强大的集成学习方法，通过结合多个基础模型的预测结果，可以显著提高预测性能。它引入了模型多样性，增加了模型的泛化能力和抗过拟合能力。同时，它灵活地结合各种算法，适用于不同的任务和数据类型。此外，通过分析每个模型的贡献程度，可以提供更好的可解释性。(Stacking is a powerful integrated learning method that significantly improves the forecasting performance by combining the forecasting results of multiple base models. It introduces model diversity and increases model generalization ability and overfitting resistance. At the same time,

it flexibly combines various algorithms and is suitable for different tasks and data types. In addition, better interpretability can be provided by analyzing the extent to which each model contributes.)

使用 Stacking 方法需要考虑计算复杂度较高的问题，因为涉及多个层级的模型训练和预测。此外，使用第一层模型的预测结果作为输入特征可能存在数据泄露问题，需要注意过拟合。同时，模型选择和调参也变得更加困难，需要仔细选择合适的模型和参数进行调整。(The use of the Stacking method requires high computational complexity because it involves model training and forecasting at multiple levels. In addition, using the predicted results of the first layer model as input features may cause data leakage problems, and overfitting should be paid attention to. At the same time, it is more difficult to select and adjust the model, which requires careful selection of appropriate models and parameters.)

2. 多层 stacking

简单步骤:

1. 第一层模型训练：使用原始训练数据集，在第一层训练多个基础模型，每个模型使用不同的特征子集或算法参数。每个基础模型都针对相同的目标进行预测。
2. 第一层模型预测：使用第一层训练好的模型对原始训练数据进行预测，并将预测结果作为新的特征。
3. 第二层模型训练：将第一层的预测结果和原始特征结合，作为新的训练数据集。在第二层训练一个或多个模型，使用这些新的特征进行预测。
4. 第二层模型预测：使用第二层训练好的模型对原始测试数据进行预测。

通过多层 Stacking，可以在不同的层级上组合多个模型，利用它们之间的相互学习和协作来提高整体预测性能。每个层级的模型可以捕捉不同层次的特征表达和关系，并通过堆叠的方式将这些信息融合起来，以得到更准确的预测结果。(With multi-stacking, multiple models can be combined at different levels and their mutual learning and collaboration can be used to improve overall forecasting performance. The models at each level capture feature representations and relationships at different levels and combine this information in a stacking manner to get more accurate predictions.)

多层 Stacking 的优势在于它的灵活性和强大的表达能力，能够通过组合多个模型的优点来弥补各个模型的缺点，提高整体预测的准确性。然而，需要注意的是，多层 Stacking 也可能面临过拟合的问题，需要进行适当的调参和模型选择来平衡模型复杂度和泛化能力。(The advantages of multi-stacking are its flexibility and strong expression ability. It can combine the advantages of multiple Stacking models to compensate for the shortcomings of each model and improve the accuracy of the overall stacking. However, it is important to note that multi-stacking may also face the problem of overfitting, and appropriate parameters and model selection are required to balance model complexity and generalization capabilities.)

Lecture10: 聚类算法（K-Means DBSCAN Hierarchical clustering）

一、 小批量梯度下降：小批量梯度下降不是基于整个训练集的梯度计算，而是将训练集分为小批量（mini-batch）数据，每个批量数据用于计算梯度和更新参数。(Small batch gradient descent: Small batch gradient descent is not based on the gradient calculation of the entire training set. Instead, the training set is divided into mini-batch data. Each batch of data is used to calculate the gradient and update parameters.)

梯度下降、小批量梯度下降、随机梯度下降：

1. 数据量：
    - 梯度下降：在每次参数更新时，使用全部的训练数据计算损失函数的梯度。
    - 小批量梯度下降：将训练数据划分为多个小批量数据，在每次参数更新时，使用一个小批量数据计算损失函数的梯度。
    - 随机梯度下降：在每次参数更新时，只使用单个训练样本计算损失函数的梯度。
2. 参数更新：
    - 梯度下降：每次参数更新时，使用全部训练数据的梯度的平均值来更新模型的参数。
    - 小批量梯度下降：每次参数更新时，使用一个小批量数据的梯度的平均值来更新模型的参数。
    - 随机梯度下降：每次参数更新时，使用单个训练样本的梯度来更新模型的参数。
3. 计算效率：
    - 梯度下降：使用全部训练数据计算梯度，计算开销较大，尤其是在大规模数据集上。
    - 小批量梯度下降：只使用小批量数据计算梯度，相对于梯度下降来说，计算开销较小，特别适用于大规模数据集。
    - 随机梯度下降：只使用单个训练样本计算梯度，计算开销最小，尤其适用于大规模数据集和在线学习。

1. Data volume:

- Gradient descent: The gradient of the loss function is calculated using all the training data at each parameter update.
- Small-lot gradient descent: The training data is divided into multiple small-lot data, and the gradient of the loss function is calculated using one small-lot data at each parameter update.
- Random gradient descent: Only a single training sample is used to calculate the gradient of the loss function at each parameter update.

2. Parameter update
    - Gradient descent: The average value of gradients of all training data is used to update model parameters each time parameters are updated.
    - Small-lot gradient descent: The average value of the gradient of a small-lot data is used to update the model's parameters each time the parameters are updated.
    - Random gradient descent: The gradient of a single training sample is used to update the parameters of the model each time the parameters are updated.
3. Computational efficiency:
    - Gradient descent: Using all training data to calculate the gradient is expensive, especially on large data sets.
    - Small batch gradient descent: Only small batch data is used to calculate the gradient. Compared with gradient descent, the calculation cost is low, especially suitable for large data sets.
    - Random gradient descent: Using only a single training sample to calculate the gradient, the computational cost is minimal, especially suitable for large-scale data sets and online learning.

二、K-Means、DBSCAN、Agglomerative Clustering

1. K-Means 简单步骤：K-Means algorithm involves the following steps: randomly initialize K cluster centroids, assign data points to the nearest centroid based on distance, update centroids by taking the mean of assigned data points, repeat the assignment and update steps until convergence, and output the final centroids and cluster assignments. The algorithm

aims to minimize the within-cluster sum of squares by iteratively optimizing centroids. （K- means 算法包括以下步骤:随机初始化 K 个聚类质心，根据距离将数据点分配给最近的质心，通过对分配的数据点取均值来更新质心，重复分配和更新步骤直到收敛，并输出最终的质心和聚类分配。该算法旨在通过迭代优化质心来最小化聚类内平方和。）

2. DBSCAN 简单步骤：Choose an epsilon (ε) neighborhood radius and a minimum number of points (MinPts) required to form a dense region. Iterate through unvisited points, expanding their ε-neighborhoods to identify core points and their reachable points. Mark noise and border points. Output discovered clusters based on core points and their reachable points. DBSCAN identifies dense regions based on connectivity without specifying the number of clusters in advance.（选择一个 ε (ε)邻域半径和形成密集区域所需的最小点数(MinPts)。迭代未访问点，扩展它们的 ε-邻域以确定核心点及其可达点。标记噪声点和边界点。根据核心点及其可达点输出发现的聚类。DBSCAN 根据连通性识别密集区域，而无需事先指定集群的数量。）

3. Agglomerative Clustering is a hierarchical algorithm that starts with individual clusters for each data point and progressively merges the most similar clusters until the desired number of clusters or a similarity threshold is reached. It iteratively merges clusters based on similarity, creating a hierarchical structure. It is flexible, as it does not require a predefined number of clusters, and is applicable to diverse data types and distance measures. (聚集聚类是一种分层算法，它从每个数据点的单个聚类开始，逐步合并最相似的聚类，直到达到所需的聚类数量或相似阈值。它基于相似性迭代地合并集群，创建一个层次结构。它是灵活的，因为它不需要预定义的集群数量，并且适用于不同的数据类型和距离度量。）

三、K-Means

1. 初始化质心（Initialize the center of mass）

- 随机选择：最简单的方法是随机选择 K 个数据点作为初始质心。（Random selection: The simplest method is to randomly select K data points as the initial center of mass.）
- K-Means++：K-Means++是一种改进的质心初始化方法，它尝试选择初始质心使它们相互之间的距离较远。具体步骤包括随机选择一个初始质心，然后迭代选择下一个质心时，以概率方式选择与当前质心距离较远的数据点作为下一个质心，直到选择完所有的质心。(K-means ++ is an improved method of centroid initialization, which tries to select the initial centroid so that they are far away from each other. The specific steps include randomly selecting an initial centroid, then iteratively selecting the next centroid, and probabilistically selecting data points that are far away from the current centroid as the next centroid until all centroids have been selected.)
- 均匀分布：可以在数据集的边界框中均匀地分布 K 个初始质心。(Uniform distribution: K initial centers of mass can be evenly distributed in the bounding box of the data set.)
- 人为设定：根据先验知识或问题的特定要求，手动指定初始质心的位置。(Artificial setting: Manually specify the location of the initial center of mass based on prior knowledge or specific requirements of the problem.)

2. K 值的选择

原因：

a) 选择合适的 K 值可以避免聚类结果过于细致或过于粗糙的情况。当 K 值过大时，可能会出现过拟合，每个簇内的样本数量较少，容易产生噪声和异常值的影响。而当 K 值过小时，可能会出现欠拟合，多个真实簇被合并成一个簇，丢失了簇内部的细节。(Selecting the appropriate K value can avoid the situation that the clustering result is too detailed or too rough. When the K value is too large, overfitting may occur, and the number of samples in each cluster is small, which is easy to produce noise and the influence of outliers. However, when the value of K is too small, underfitting may occur, multiple real clusters are merged into one cluster, and the details inside the cluster are lost.)

b) 较小的 K 值通常意味着更高的计算效率，因为需要处理的数据点较少。选择适当的 K 值可以减少计算复杂度和存储需求，提高算法的效率和可扩展性。(A smaller K value usually means more computational efficiency because there are fewer data points to process. Choosing the appropriate K value can reduce the computational complexity and storage requirements, and improve the efficiency and scalability of the algorithm.)

3. 肘部法则 Elbow rule

根据肘部法则选择的 K 值通常是在增加 K 值时， SSE 减少幅度显著减小的拐点处。这是因为增加 K 值后，每个簇内的样本数量减少， SSE 的改善效果会逐渐减弱，而拐点处的 K 值通常能够平衡聚类的准确性和模型的复杂度。(The K value chosen according to the elbow rule is usually the inflection point at which the SSE reduction decreases significantly when the K value is increased. This is because after increasing K value, the number of samples in each cluster decreases, and the improvement effect of SSE will gradually weaken, and the K value at the inflection point can usually balance the accuracy of clustering and the complexity of the model.)

肘部法则仅作为一种启发式方法，不一定适用于所有情况。在某些数据集中，SSE 曲线可能不具有明显的拐点，或者最佳 K 值可能不止一个。因此，结合领域知识和其他评估方法来选择最佳的 K 值是推荐的做法。(The elbow rule is only used as a heuristic and may not be applicable in all cases. In some data sets, SSE curves may not have obvious inflection points, or there may be more than one optimal K value. Therefore, combining domain knowledge and other evaluation methods to select the best K value is recommended practice.)

4. 轮廓系数（silhouette score）

较高的平均轮廓系数通常表示较好的聚类结果。轮廓系数能够考量聚类的紧密性和分离度，对于评估不同 K 值下的聚类质量和选择最佳 K 值具有一定的指导作用。然而，轮廓系数也有其局限性，对于非凸形状的簇和噪声数据可能产生不准确的结果，因此在使用时需要综合考虑其他评估指标和实际应用场景。(Higher average contour coefficients usually indicate better clustering results. The contour coefficient can be used to measure the tightness and separation degree of clustering and has a certain guiding role in evaluating the clustering quality under different K values and selecting the best K value. However, the contour coefficient also has its limitations, which may produce inaccurate results for non-convex clusters and noisy data. Therefore, other evaluation indexes and practical application scenarios should be considered comprehensively when using the contour coefficient.)

5. K-Means 的缺点

a) 需要预先指定簇的数量（K 值）：K-Means 算法需要提前知道要将数据划分成的簇的数量。然而，在实际应用中，确定最优的 K 值可能是一个挑战，且选择不合适的 K 值可能导致聚类结果不准确或不理想。(Need to specify the number of clusters in advance (K value) : The k-means algorithm needs to know in advance the number of clusters into which the data is to be divided. However, in practical applications, determining the optimal K value can be a challenge, and selecting an inappropriate K value can lead to inaccurate or unsatisfactory clustering results.)

b) 对初始质心敏感：K-Means 算法对初始质心的选择非常敏感。不同的初始质心可能导致不同的聚类结果，可能会陷入局部最优解而无法收敛到全局最优解。因此，质心的初始化方法对 K-Means 算法的性能和结果有很大影响。(Sensitive to initial centroid: The K-Means algorithm is very sensitive to the selection of initial centroid. Different initial centroid may lead to different clustering results and may fall into the local optimal solution and fail to converge to the global optimal solution. Therefore, the initialization method of centroid has a great influence on the performance and results of K-Means algorithm.)

c) 对异常值和噪声敏感：K-Means 算法对异常值和噪声数据敏感。异常值可能会对质心的计算产生较大影响，导致聚类结果不准确。K-Means 算法无法有效处理非球形簇和不均衡大小的簇。(Sensitive to outliers and noise: K-Means algorithm is sensitive to outliers and noise data. Outliers may have great influence on the calculation of centroid and lead to inaccurate clustering results. K-Means algorithm can not effectively deal with non-spherical clusters and clusters of uneven size.)

d) 限制于欧氏距离度量：K-Means 算法使用欧氏距离作为度量样本之间的相似度，但对于非欧氏空间或具有不同度量的数据，K-Means 算法可能不适用。此外，欧氏距离假设各个特征对距离计算的贡献是相等的，对于具有不同重要性的特征， K-Means 算法可能表现不佳。(Restricted to Euclidean distance measures: The K-Means algorithm uses Euclidean distance as a measure of similarity between samples, but for non-Euclidean Spaces or data with different measures, the K-Means algorithm may not be applicable. In addition, Euclidean

distance assumes that each feature contributes equally to distance calculation, and the K-Means algorithm may perform poorly for features of different importance.)

e) 难以处理大规模数据集：K-Means 算法在处理大规模数据集时可能面临计算和存储的挑战。由于需要计算每个数据点与所有质心之间的距离，当数据集很大时，计算复杂度会显著增加。(Difficulty in processing large data sets: K-Means algorithm may face computational and storage challenges when processing large data sets. Due to the need to calculate the distance between each data point and all centers of mass, computational complexity increases significantly when the data set is large.)

6. K-Means 预处理

a) 特征标准化或归一化：将数据特征进行标准化或归一化，使得各个特征具有相似的尺度。这可以避免某些特征对距离计算和聚类结果产生较大影响，提高算法的稳定性和准确性。(Feature standardization or normalization: data features are standardized or normalized so that each feature has a similar scale. This can avoid the great influence of some features on distance calculation and clustering results, and improve the stability and accuracy of the algorithm.)

b) 处理缺失值：如果数据中存在缺失值，需要进行适当的处理。可以选择填充缺失值或使用插值方法进行估计，以确保数据的完整性和准确性。(If there are missing values in the data, appropriate processing is required. You can choose to fill in missing values or estimate using interpolation methods to ensure data integrity and accuracy.)

c) 异常值处理：异常值可能对 K-Means 算法产生较大影响，因此需要进行异常值检测和处理。可以使用统计方法或离群点检测算法识别和处理异常值，例如通过删除异常值或替换为合适的值。(Outlier processing: Outliers may have a great impact on the K-Means algorithm, so outlier detection and processing are required. Outliers can be identified and processed using statistical methods or outlier detection algorithms, for example by removing outliers or replacing them with appropriate values.)

d) 特征选择或降维：如果数据具有高维特征，可以考虑进行特征选择或降维。这可以减少计算复杂度、消除冗余特征，并提高聚类结果的可解释性和稳定性。(Feature selection or dimension reduction: If the data has high dimensional characteristics, you can consider feature selection or dimension reduction. This can reduce computational complexity, eliminate redundant features, and improve the interpretability and stability of clustering results.)

e) 数据平滑：对于具有噪声或离群值的数据，可以考虑使用平滑技术进行数据平滑处理。平滑可以消除噪声和异常值对聚类结果的干扰，提高算法的鲁棒性。(Data smoothing: For data with noise or outliers, data smoothing technology can be considered for data smoothing processing. Smoothing can eliminate the interference of noise and outliers on clustering results and improve the robustness of the algorithm.)

f) 数据集分割：如果数据集非常大，可以考虑将其分割成小批量进行聚类。这可以减少计算和存储的压力，并提高算法的效率。(Data set segmentation: If the data set is very large, it can be considered to divide it into small batches for clustering. This can reduce computation and storage stress and improve the efficiency of the algorithm.)

7. 半监督学习算法来提高 K-Means 算法的性能

简要过程：半监督学习方法可以提高 K-Means 算法的性能。通过将标记信息传播到未标记数据点并更新聚类结果，半监督 K-Means 利用标签信息来约束聚类过程，提高聚类的准确性和稳定性。这种方法特别适用于标记数据较少的情况下，能够更有效地利用有限的标记信息来改善聚类结果。(Semi-supervised learning method can improve the performance of K-Means algorithm. By propagating label information to unlabeled data points and updating clustering results, semi-supervised K-Means uses label information to constrain the clustering process and improve the accuracy and stability of clustering. This method is especially suitable for the case of less labeling data and can make more effective use of limited labeling information to improve clustering results.)

具体步骤：

a) 初始聚类：首先，使用无监督的 K-Means 算法对未标记数据进行聚类，得到初始的簇分配结果。(Initial clustering: First, the unlabeled data is clustered using the unsupervised K-Means algorithm to get the initial cluster allocation result.)

b) 标记传播：利用已标记的一部分数据，将标签信息传播到未标记的数据点。这可以通过标记传播算法来实现，例如标签传播算法。标记传播的目标是根据已标记样本之间的相似性或连接性，将标签信息传递给未标记样本。传播的方式可以根据具体问题选择，例如基于相似度的传播或基于图的传播。(Tag propagation: Using a portion of the tagged data, the tag information is propagated to the untagged data point. This can be done through tag propagation algorithms, such as tag propagation algorithms. The goal of tag propagation is to pass tag information to unlabeled samples based on similarities or connectivity between labeled samples. The mode of propagation can be selected according to the specific problem, such as similarity-based propagation or graph-based propagation.)

c) 重新聚类：使用传播后的标签信息更新初始聚类结果。将已标记的数据点与其传播的标签固定在相应的簇中，然后重新执行 K-Means 算法，使用传播后的标签作为约束条件。这样可以调整簇的分配，并更好地利用标记信息。重新聚类的过程可以通过最小化簇内平方误差或其他适当的聚类评估指标来完成。(Recluster: Use the propagated label information to update the initial clustering results. The labeled data points and their propagated labels are fixed in the corresponding cluster, and then the K-Means algorithm is re-executed, using the propagated labels as constraints. This allows you to adjust cluster allocation and make better use of tag information. The process of reclustering can be accomplished by minimizing the in-cluster squared error or other appropriate clustering evaluation indicators.)

d) 迭代优化：重复执行步骤 2 和步骤 3，直到达到收敛条件或迭代次数达到预定值。每次迭代都会进一步优化簇的分配和标签传播，提高聚类的准确性和一致性。(Iterative optimization: Repeat steps 2 and 3 until convergence conditions are reached or the number of iterations reaches a predetermined value. Each iteration will further optimize cluster distribution and label propagation to improve the accuracy and consistency of clustering.)

Lecuture11: GMM 高斯混合模型

一、高斯混合模型的构建过程：The construction process of a Gaussian Mixture Model (GMM) involves initializing the parameters, using the Expectation-Maximization (EM) algorithm to iteratively update the parameters based on computed responsibilities, checking for convergence, selecting the optimal number of components, and utilizing the trained GMM for tasks such as clustering or density estimation. The goal is to find the best parameter values that maximize the likelihood of the observed data. （高斯混合模型(GMM)的构建过程包括初始化参数，使用期望最大化(EM)算法根据计算的责任迭代更新参数，检查收敛性，选择最优分量数量，并利用训练好的 GMM 进行聚类或密度估计等任务。目标是找到使观测数据的可能性最大化的最佳参数值。）

二、期望最大化（Expectation Maximization，EM）

Expectation Maximization (EM) is an iterative algorithm that is used for parameter estimation in probabilistic models. (EM 是一种迭代算法，用于在概率模型中进行参数估计）

The core idea of EM algorithm is to optimize parameter estimation by iteratively performing steps E and step M. In step E, the expectation of the implied variable, i.e. the likelihood of the observed data given the current model parameters, is obtained by calculating the posterior probability. In the M step, model parameters are updated according to these expectations to make the likelihood function larger. (EM 算法的核心思想是通过迭代地进行 E 步骤和 M 步骤来优化参数估计。在 E 步骤中，通过计算后验概率，获得对隐含变量的期望，即给定当前模型参数下观测数据的似然。在 M 步骤中，根据这些期望更新模型参数，使似然函数增大。）

步骤：

1. Initialize the model parameters.
2. Expectation step (E step) : Calculate a posteriori probability (expectation) of implied variables based on the current model parameters.
3. Maximization step (M step) : Update model parameters to maximize the likelihood function based on the posterior probability of the implied variable.
4. Repeat steps E and M until convergence (a predetermined stop condition is reached).

1. 初始化模型参数。
2. Expectation 步骤（E 步骤）：基于当前模型参数，计算隐含变量的后验概率（期望）。
3. Maximization 步骤（M 步骤）：根据隐含变量的后验概率，更新模型参数以最大化似然函数。
4. 重复执行 E 步骤和 M 步骤，直到收敛（达到预定的停止条件）。

三、中心极限定理(central limit theorem)：

The central limit theorem states that when independent random variables are added together, their sum tends to follow a normal distribution, regardless of the distribution of the individual variables, as long as certain conditions are met.

中心极限定理指出，当独立的随机变量加在一起时，只要满足一定的条件，它们的和倾向于服从正态分布，而不管单个变量的分布如何。

四、GMM 用于样本的异常检测：

Anomaly detection is done by setting a threshold. Samples whose PDF is less than the threshold value are considered abnormal. （通过设定一个阈值来进行异常检测。测试样本中 PDF 小于阈值的样本被视为异常。）

步骤：

1. Data Preparation: Prepare the dataset on which the anomaly detection will be performed.
2. Model Training: Train a Gaussian Mixture Model on the training dataset. In GMM, a mixture of Gaussian distributions is used to model the underlying data distribution. The model learns the parameters of the Gaussian components, including mean and covariance, using an iterative algorithm such as Expectation-Maximization (EM).

3. Probability Estimation: Compute the probability of each data point in the dataset based on the trained GMM model. This is done by evaluating the probability density function (PDF) of the GMM at each data point.
4. Threshold Selection: Determine a suitable threshold for classifying data points as anomalies or normal. This can be done by analyzing the probability distribution of the data points and selecting a threshold that separates normal and anomalous regions.
5. Anomaly Detection: Classify the data points as anomalies or normal based on the selected threshold. Data points with probabilities below the threshold are considered anomalies, while those above the threshold are classified as normal.
6. Evaluation: Evaluate the performance of the anomaly detection algorithm using appropriate metrics such as precision, recall, or the receiver operating characteristic (ROC) curve. This helps assess the effectiveness of the GMM model in detecting anomalies.

1. 数据准备:准备待进行异常检测的数据集。
2. 模型训练:在训练数据集上训练高斯混合模型。在 GMM 中，使用混合高斯分布来模拟底层数据分布。该模型使用期望最大化(EM)等迭代算法学习高斯分量的参数，包括均值和协方差。
3. 概率估计:基于训练好的 GMM 模型计算数据集中每个数据点的概率。这是通过评估 GMM 在每个数据点的概率密度函数(PDF)来完成的。
4. 阈值选择:确定一个合适的阈值，用于将数据点分类为异常或正常。这可以通过分析数据点的概率分布和选择一个区分正常和异常区域的阈值来完成。
5. 异常检测:根据所选阈值对数据点进行异常或正常分类。概率低于阈值的数据点被认为是异常，而高于阈值的数据点被归类为正常。
6. 评估:使用适当的指标评估异常检测算法的性能，如精度、召回率或接收者工作特征(ROC)曲线。这有助于评估 GMM 模型在检测异常方面的有效性。

五、模型的似然（通过最大似然估计找模型的参数）:

计算模型的似然度时，我们需要对数据的分布形式和模型的参数做出假设。似然度通常表示为在给定模型参数的情况下，观测数据点的联合概率密度函数（PDF）或概率质量函数（PMF）。数学上，如果我们用 θ 表示模型参数，用 D 表示观测数据，那么似然函数可以写为 L(D | θ)，其中 L 表示似然度。(When calculating the likelihood of the model, we need to make assumptions about the distribution form of data and the parameters of the model. The likelihood is usually expressed as the joint probability density function (PDF) or probability mass function (PMF) of the observed data points given the model parameters. Mathematically, if we use theta as model parameters and observation data with D, then the likelihood function can be written as L (D | theta), L said the likelihood of degrees.)

我们的目标是找到使似然函数最大化的模型参数值。这个过程被称为最大似然估计（MLE）。通过最大化似然度，我们可以找到使观测数据在模型假设下最有可能出现的参数值。(Our goal is to find model parameter values that maximize the likelihood function. This process is called maximum likelihood estimation (MLE). By maximizing likelihood, we can find the parameter values that make the observed data most likely to occur under the model's assumptions.)

具体来说，最大似然估计（MLE）是一种常用的参数估计方法，通过寻找使似然函数最大化的参数值来确定最优参数。最大化似然函数相当于寻找最能解释观测数据的模型参数组合。(Specifically, maximum likelihood estimation (MLE) is a commonly used parameter estimation method that determines the optimal parameter by finding the parameter values that maximize the likelihood function. Maximizing the likelihood function is equivalent to finding the combination of model parameters that best explain the observed data.)

定义：Model likelihood, also known as the probability of the data given the model, is a measure of how well the model explains or fits the observed data. It quantifies the probability of obtaining the observed data under the assumptions of the model. A higher likelihood indicates a better fit of the model to the data, while a lower likelihood suggests a poorer fit.

模型似然，也称为给定模型的数据的概率，是衡量模型解释或拟合观测数据的程度。它量化了在模型假设下获得观测数据的概率。似然值越高表明模型与数据的拟合越好，而似然值越低则表明模型与数据的拟合越差。

六、信息标准 information criterion

用于选择高斯分量的份数 Used to select the number of Gaussian components（份数过小欠拟合 过大过拟合：The number of copies is too small to underfit and too large to overfit）

定义：**An information criterion is a metric used for model selection, which balances the goodness of fit of the model to the data and the complexity of the model. It provides a quantitative measure to compare different models based on their relative performance.** 信息准则是一种用于模型选择的度量，它平衡了模型与数据的拟合优度和模型的复杂性。它提供了一种定量的方法来比较不同模型的相对性能。

The smaller the information standard is, the better the model fits the data. 越小越好

1. 人工智能的应用

人工智能（Artificial Intelligence，简称 AI）是一项涵盖多个领域的技术和方法，它模拟和实现了人类智能的某些方面。人工智能的应用广泛，涵盖了许多行业和领域。以下是一些常见的人工智能应用示例：（Artificial Intelligence (AI) is a technology and method covering many fields. It simulates and realizes some aspects of human intelligence. The application of artificial intelligence is wide, covering many industries and fields. Here are some examples of common AI applications:）

1. 机器学习（Machine Learning）：机器学习是人工智能的一个重要分支，通过使用算法和模型让计算机从数据中学习和改进性能。机器学习应用广泛，包括图像识别、语音识别、自然语言处理、推荐系统、智能搜索等。（Machine Learning: Machine learning is an important branch of artificial intelligence that uses algorithms and models to enable computers to learn and improve performance from data. Machine learning has a wide range of applications, including image recognition, speech recognition, natural language processing, recommendation systems, intelligent search, etc.）
2. 自动驾驶（Autonomous Driving）：人工智能在汽车领域的应用，使汽车能够感知环境、决策和控制行驶。自动驾驶技术利用传感器、计算机视觉和机器学习等技术，实现车辆的自主导航和交通规划。（Autonomous Driving: The use of artificial intelligence in cars, which allows vehicles to sense their environment, make decisions and control their driving. Autonomous driving technology uses sensors, computer vision and machine learning to enable autonomous navigation and traffic planning in vehicles.）
3. 语音助手（Voice Assistants）：语音助手如 Siri、Google Assistant、Amazon Alexa 等利用自然语言处理和语音识别技术，能够理解和回答用户的问题，执行指令，提供日程安排、天气预报、音乐播放等服务。（Voice Assistants: Using natural language processing and speech recognition technology, voice assistants such as Siri, Google Assistant, and Amazon Alexa will be able to understand and answer user questions, execute instructions, and provide scheduling, weather reports, and music playback.）
4. 智能机器人（Intelligent Robots）：智能机器人利用人工智能技术和机器学习算法，具备感知、认知和决策能力，能够执行各种任务。智能机器人应用于制造业、医疗、服务业等领域，如工厂自动化、手术机器人、家庭服务机器人等。（Intelligent Robots: Using artificial intelligence technology and machine learning algorithms, intelligent robots have the ability of perception, cognition and decision making, and can perform various tasks. Intelligent robots are used in manufacturing, medical, service and other fields, such as factory automation, surgical robots, home service robots, etc.）
5. 金融科技（FinTech）：人工智能在金融领域的应用包括风险评估、欺诈检测、投资分析、智能客服等。机器学习和数据分析技术可以帮助金融机构提高风险管理和决策能力。（Financial technology (FinTech) : Applications of artificial intelligence in the financial sector include risk assessment, fraud detection, investment analysis, intelligent customer service, etc. Machine learning and data analytics can help financial institutions improve their risk management and decision-making capabilities.）
6. 医疗诊断与辅助（Medical Diagnosis and Assistance）：人工智能技术在医疗领域的应用涉及疾病诊断、影像分析、个性化治疗等。机器学习和深度学习算法可以辅助

医生进行疾病诊断和预测，提高医疗效率和准确性。（Medical Diagnosis and Assistance: The application of artificial intelligence technology in the medical field involves disease diagnosis, image analysis, personalized treatment, etc. Machine learning and deep learning algorithms can help doctors diagnose and predict diseases, improving medical efficiency and accuracy.）

7. 智能城市（Smart Cities）：人工智能技术可以应用于城市管理、交通优化、能源管理、环境监测等方面。通过数据分析和智能决策，实现城市资源的有效利用和提升居民生活质量。（Smart Cities: Artificial intelligence technology can be applied to urban management, transportation optimization, energy management, environmental monitoring and other aspects. Through data analysis and intelligent decision-making, urban resources can be effectively used and residents' quality of life can be improved.）
8. 自然语言处理（Natural Language Processing，NLP）：NLP 技术用于理解和处理人类语言，包括机器翻译、文本分析、情感分析、问答系统等。NLP 可以应用于客服、舆情分析、知识图谱构建等领域。（Natural Language Processing (NLP)：NLP technology is used to understand and process human language, including machine translation, text analysis, sentiment analysis, question and answer system, etc. NLP can be applied to customer service, public opinion analysis, knowledge map construction and other fields.）

Machine Listening

Machine listening（机器听觉）是指通过计算机和机器学习技术来模拟人类听觉系统的能力，使计算机能够感知、理解和处理音频信号。

与机器视觉类似，机器听觉旨在让计算机对音频信号进行分析和解释，从中提取有用的信息，并进行相关的决策和应用。机器听觉涉及多个领域的交叉，包括信号处理、机器学习、音频分析、音频合成、语音识别、音乐信息检索等。

以下是一些机器听觉的具体任务和应用：

1. 语音识别（Speech Recognition）：将语音信号转换为文本形式的任务，使计算机能够理解和处理人类语音输入。
2. 声音分类和识别（Sound Classification and Recognition）：对不同类型的声音进行分类和识别，如环境音、乐器音、动物叫声等。
3. 音频分割和分离（Audio Segmentation and Separation）：将复杂的音频信号分割成不同的音频事件或分离出多个声源。
4. 音频合成（Audio Synthesis）：基于给定的输入生成合成的音频信号，如文本到语音合成（Text-to-Speech）。
5. 音频增强（Audio Enhancement）：改善音频质量，降噪、消除回声、增强语音等，以提升音频的清晰度和可理解性。
6. 音频事件检测和识别（Audio Event Detection and Recognition）：检测和识别特定的音频事件，如咳嗽声、汽车鸣笛声等。
7. 音频情感分析（Audio Emotion Analysis）：分析音频信号中蕴含的情感信息，如高兴、悲伤、紧张等。

机器听觉的应用非常广泛，包括语音助手（如 Siri、Alexa）、音频内容识别和推荐、音频分析和监控、音频处理和修复等领域。通过机器听觉的技术，计算机能够更好地理解和处理音频信号，从而实现更多有意义的音频交互和应用。

Acoustic signal processing tasks（声学信号处理任务）是指在声学领域中对声音信号进行处理和分析的一系列任务。声学信号处理任务旨在从声学信号中提取有用的信息，以实现音频处理、语音识别、音频合成、音频增强等应用。
以下是一些常见的声学信号处理任务:

1. 语音识别（Speech Recognition）：将语音信号转换为对应的文本或命令，使计算机能够理解和处理语音输入。
2. 声音分类和识别（Sound Classification and Recognition）：对不同类别的声音进行分类和识别，例如环境音、乐器音、动物叫声等。
3. 声纹识别（Speaker Recognition）：识别和验证个体的身份基于他们的声音特征，用于声纹识别系统和身份验证。
4. 音频合成（Audio Synthesis）：基于给定的输入，生成合成的音频信号，如文本到语音合成（Text-to-Speech）。
5. 音频增强（Audio Enhancement）：改善音频质量，降噪、消除回声、增强语音等，以提升音频的清晰度和可理解性。
6. 声源定位（Sound Source Localization）：确定声音源的位置和方向，用于音频会议、追踪和定位等应用。
7. 音频分割和分离（Audio Segmentation and Separation）：将复杂的音频信号分割成不同的音频事件或分离出多个声源。
8. 音频特征提取（Audio Feature Extraction）：从音频信号中提取出有用的特征，如梅尔频谱系数（Mel Frequency Cepstral Coefficients，MFCC）、功率谱密度等。

这些声学信号处理任务在语音处理、音频处理、声音分析、音乐信息检索等领域中具有重要的应用价值，为我们提供了更好的音频交互和信息处理能力。

machine listening
Machine listening（机器听觉）是指通过计算机和机器学习技术来模拟人类听觉系统的能力，使计算机能够感知、理解和处理音频信号。
与机器视觉类似，机器听觉旨在让计算机对音频信号进行分析和解释，从中提取有用的信息，并进行相关的决策和应用。机器听觉涉及多个领域的交叉，包括信号处理、机器学习、音频分析、音频合成、语音识别、音乐信息检索等。
以下是一些机器听觉的具体任务和应用:

1. 语音识别（Speech Recognition）：将语音信号转换为文本形式的任务，使计算机能够理解和处理人类语音输入。
2. 声音分类和识别（Sound Classification and Recognition）：对不同类型的声音进行分类和识别，如环境音、乐器音、动物叫声等。
3. 音频分割和分离（Audio Segmentation and Separation）：将复杂的音频信号分割成不同的音频事件或分离出多个声源。
4. 音频合成（Audio Synthesis）：基于给定的输入生成合成的音频信号，如文本到语音合成（Text-to-Speech）。
5. 音频增强（Audio Enhancement）：改善音频质量，降噪、消除回声、增强语音等，以提升音频的清晰度和可理解性。
6. 音频事件检测和识别（Audio Event Detection and Recognition）：检测和识别特定的音频事件，如咳嗽声、汽车鸣笛声等。
7. 音频情感分析（Audio Emotion Analysis）：分析音频信号中蕴含的情感信息，如高兴、

悲伤、紧张等。

机器听觉的应用非常广泛，包括语音助手（如 Siri、Alexa）、音频内容识别和推荐、音频分析和监控、音频处理和修复等领域。通过机器听觉的技术，计算机能够更好地理解和处理音频信号，从而实现更多有意义的音频交互和应用。

自然语言处理（Natural Language Processing，简称 NLP）技术是指将人类自然语言与计算机进行交互和处理的一系列技术和方法。NLP 旨在让计算机能够理解、解析、生成和处理人类语言，从而实现对文本和语言数据的自动分析和处理。

以下是一些常见的自然语言处理技术：

1. 分词（Tokenization）：将文本切分成词语、句子或其他更小的单元，为后续处理建立基本单位。
2. 词性标注（Part-of-Speech Tagging）：对文本中的每个词汇标注其词性，如名词、动词、形容词等。
3. 句法分析（Syntax Parsing）：分析句子的结构和语法关系，如依存关系分析和短语结构分析。
4. 实体识别（Named Entity Recognition）：识别文本中的实体，如人名、地名、组织机构名等。
5. 语义角色标注（Semantic Role Labeling）：标注句子中的语义角色，如谓词、施事者、受事者、时间等。
6. 情感分析（Sentiment Analysis）：分析文本中的情感倾向，判断文本的情感极性，如正面、负面、中性。
7. 文本分类（Text Classification）：将文本分到预定义的类别中，如垃圾邮件过滤、情感分类等。
8. 机器翻译（Machine Translation）：将文本从一种语言翻译成另一种语言。
9. 问答系统（Question Answering）：根据用户提出的问题，从文本中找到相关的答案。
10. 自动摘要（Automatic Summarization）：从大篇幅文本中自动生成摘要或概括。
11. 对话系统（Dialogue Systems）：实现与计算机的自然对话，例如智能助理。

这些技术是 NLP 中的一部分，其目标是使计算机能够理解和处理人类语言，实现文本理解、情感分析、文本生成等任务。NLP 技术在机器翻译、信息检索、文本分析、智能对话等领域具有广泛的应用，并在日常生活中逐渐扮演重要角色。