半监督学习

1. 什么是半监督学习？请简要描述其定义和主要思想。它与监督学习和无监督学习有什么区别？

半监督的学习：有少量训练样本，学习机以从训练样本获得的知识为基础，结合测试样本的分布情况逐步修正已有知识，并判断测试样本的类别。

有监督的学习：学习器通过对大量有标记的训练例进行学习，从而建立模型用于预测未见示例的标记，很难获得大量的标记样本。

无监督的学习：无训练样本，仅根据测试样本的在特征空间分布情况来进行标记，准确性差。

1. 解释图半监督学习（graph-based semi-supervised learning）的基本思想和过程。

基于相似的样本具有相同标签的平滑假设，构建图结构，利用节点标签扩散，将已标记样本根据样本特征的相似性进行扩散，获得未标记样本的标记。过程包括：图表示、初始化标签、标签传播、学习模型、分类预测。

1. 什么是半监督聚类？它与传统聚类方法有何不同？

半监督聚类在拥有部分额外监督信息时，可利用监督信息改善聚类效果，结合了有标签的数据和未标签的数据。传统聚类方法通常只使用无标签数据进行聚类。

1. 什么是“伪标签”？它在半监督学习中的作用是什么？

"伪标签"是一种半监督学习中的策略，它通过使用已训练模型对未标记数据进行预测，并将这些预测结果作为虚拟的标签（伪标签）来扩充训练数据。可以扩充训练数据和引入额外信息，但是需要注意误差传播的问题。

迁移学习

1. 什么是迁移学习？请简要描述其定义和主要思想。

迁移学习指利用数据、任务、或模型之间的相似性，将在旧领域学习过的模型，应用于新领域的一种学习过程。迁移学习提供了一种基于大数据“预训练”的模型在自己的特定数据集上进行“微调”的技术。核心思路是找到源域和目标域之间的差异性，并将整体工作归结为两方面：一是度量两个领域的相似性，不仅定性地告诉我们它们是否相似，更定量地给出相似程度；二是以度量为准则，通过我们所要采用的学习手段，增大两个领域之间的相似性，完成迁移学习。

1. 请解释“领域”在迁移学习中的概念。

领域是机器学习的主体，主要由两个部分构成：数据和生成这些数据的概率分布。通常用D表示一个领域，可被表示为D= {X,Y,P(x,y)}。其中X,Y,分别为数据所处的特征空间和标签空间，P为数据服从的概率分布。

1. 什么是领域自适应？请简要描述其基本原理。

领域自适应是迁移学习中的一个重要概念，它旨在解决在源领域上训练的模型在目标领域上性能下降的问题。在领域自适应中，源领域和目标领域通常具有不同的分布或特性，例如，它们可能来自不同的传感器、环境或任务设置。

基本原理包括：

1. 源领域训练：在源领域上，我们使用标记数据来训练一个模型。这个模型在源领域上表现良好，因为它是在这个领域上训练的。
2. 领域间差异：源领域和目标领域之间存在领域间的差异，这些差异可能包括数据分布的不同、特征的变化等。由于这些差异，直接将源领域上训练的模型应用到目标领域上可能导致性能下降。
3. 域适应方法：领域自适应的目标是通过调整模型，使其适应目标领域的特点，从而提高在目标领域上的性能。有许多领域自适应方法，其中一些常见的包括：
4. 特征层面的适应：调整模型的特征提取层，以减小源领域和目标领域之间的特征差异。
5. 实例重权：调整训练样本的权重，使目标领域中的样本在训练中更加重要。
6. 对抗训练：引入对抗性训练，通过最小化源领域和目标领域之间的领域差异来提高泛化性能。
7. 联合训练：在源领域和目标领域上联合训练，以共同优化模型，使其能够适应目标领域。
8. 评估：在目标领域上评估调整后的模型性能，根据评估结果进一步调整和改进领域自适应方法。
9. 什么是源领域和目标领域？什么是领域间距离，如何度量？

源领域：源领域是模型在训练阶段接触到的领域。在源领域中，模型接收到标记数据，通过学习这些数据来训练模型。源领域通常具有相对丰富的标记数据。

目标领域：目标领域是模型在测试或应用阶段面对的领域。在目标领域中，模型通常面临的是没有或很少有标记的数据。迁移学习的目标是通过从源领域学到的知识来提高模型在目标领域上的性能。

将源域和目标域的差异性视为两个域内样本数据的概率分布的差异性进行研究，根据采用显式或隐式的距离度量分为两类：

1. 显式度量：由预定义好的距离公式产生的度量，具有特定形式，常用的距离度量有欧式距离、闵可夫斯基距离、马氏隐式度量：并非预先定义好的，而是可以在数据中动态学习的、更适合数据分布的度量距离、余弦相似度等
2. 迁移学习的基本类型有哪些，请简要介绍，并分别列举典型算法
3. 基于样本的迁移学习：根据源领域和目标领域的相似度来学习源领域样本的权重
4. 基于特征的迁移学习：学习一个特征变换T来减小迁移正则化项R(·,· )
5. 基于模型的迁移学习：学习如何将源领域的判别函数f(·,· )对目标领域数据进行正则化和微调

小样本学习

1. 什么是相似度函数，相似度函数学习的基本思想是什么？

相似度函数（Similarity Function）是用于度量两个对象之间相似程度的函数。在机器学习和数据挖掘中，相似度函数通常用于比较两个样本、文档、图像或其他数据表示的对象之间的相似性。相似度函数的设计直接影响了模型的性能，因为它定义了模型对相似性的理解。在迁移学习中，相似度函数可以用于度量源领域和目标领域之间的相似性，或者在特定任务中度量不同任务之间的相似性。

基本思想是从大规模训练数据集中学习一个相似度函数（元学习），将相似度函数用于预测，通过比对查询样本与支持集每一个样本的相似度，找到得分最高的那个。

1. 请介绍相似度函数学习的流程，并解释孪生网络和三元组损失
   1. 在大规模数据集上训练一个孪生网络
   2. 给定一个k-way，n-shot的支持集
   3. 给定查询样本，预测其属于的分类

孪生网络是一种特殊的神经网络结构，有两个相同的子网络，它们共享相同的权重。这两个子网络分别处理输入样本，然后通过一个比较层（如欧氏距离或余弦相似度）将两个子网络的输出进行比较。孪生网络的训练目标是使得相似的样本在比较层中的距离较小，而不相似的样本的距离较大。对比损失是孪生网络中常用的损失函数。

三元组损失是一种用于训练相似度学习模型的损失函数，通常与三元组网络一起使用。在一个三元组中，包含一个锚点样本和两个样本，一个是正样本，另一个是负样本。三元组损失的目标是通过拉近锚点样本和正样本之间的距离，并推远锚点样本和负样本之间的距离，从而学到一个能够准确衡量相似性的模型。