## 题库

### **第五章 数据故事化的方法与技术**

**一、选择题**

**1. 可视故事化是指什么？**

A. 以可视化方法为主要叙事手段的故事创作过程。

B. 以文字为主要叙事手段的故事创作过程。

C. 以语音为主要叙事手段的故事创作过程。

D. 以视频为主要叙事手段的故事创作过程。

答案：A

解析：根据给定文本，可视故事化是指以可视化方法为主要叙事手段的故事创作过程。

1. **可视故事通常可以采用以下哪些形式进行展示？**

A. 统计图表

B. 照片

C. 插图

D. 视频

E. 文字

F. 语音

答案：A、B、C、D、E、F

解析：根据给定文本，可视故事及其故事点可以为统计图表、照片、插图、视频，并且还可以采用文字、语音和视频来增强可视故事的效果。

1. **在可视故事化中，数据故事化是什么？**

A. 目的

B. 手段

C. 结果

D. 应用

答案：A

解析：根据给定文本，可视故事化中，数据故事化是目的，而数据可视化是手段。

1. **可视故事化的常见类型有几种？**

A. 5种

B. 6种

C. 7种

D. 8种

答案：C

解析：根据给定文本，可视故事的常见类型有7种，包括杂志、流程图、连环画、标注图表、分区海报、幻灯片、电影/视频/动画。

1. **可视故事化与数据可视化的区别是什么？**

A. 可视化侧重的是数据要素的可视化，而可视故事化侧重的是如何将数据要素做成数据产品——数据故事。

B. 可视化侧重的是数据要素的组织，而可视故事化侧重的是数据要素的可视化。

C. 可视化是数据故事化的一种类型。

D. 可视化和可视故事化没有本质区别。

答案：A

解析：根据给定文本，可视故事化与数据可视化的区别在于，可视化侧重的是数据要素的可视化，而可视故事化侧重的是如何将数据要素做成数据产品——数据故事。

1. **在Tableau中，可视故事由什么组成？**

A. 故事点

B. 故事线

C. 图表

D. 照片

答案：A

解析：根据给定文本，在Tableau中，可视故事（Visual Story）由若干个故事点（Story Points）组成，而每个故事点包含一个或多个可视化图表。

1. **条形图主要用于表示什么类型的数据之间的比较？**

A. 分类型数据之间的比较

B. 连续数据之间的比较

C. 时间序列数据之间的比较

D. 地理空间数据之间的比较

答案：A

解析：根据给定文本，条形图显示的是分类型数据（categorical data）之间的比较。

1. **饼图主要表示什么之间的比例关系？**

A. 整体与部分之间的比例关系

B. 连续数据之间的比例关系

C. 分类型数据之间的比例关系

D. 时间序列数据之间的比例关系

答案：A

解析：根据给定文本，饼图（Pie Chart）主要表示整体与部分之间的比例关系。

1. **为什么有些评论家反对使用饼图？**

A. 饼图对数据的要求高，可视化表达能力有限，容易被读者误读。

B. 饼图难以绘制，需要专业技能。

C. 饼图占用过多的空间。

D. 饼图在统计学中没有实际意义。

答案：A

解析：根据给定文本，有些评论家反对使用饼图是因为饼图对数据的要求高，可视化表达能力有限，容易被读者误读。

1. **箱线图主要用于可视化什么类型的数据分布？**

A. 分类型数据的分布

B. 连续数据的分布

C. 时间序列数据的分布

D. 地理空间数据的分布

答案：B

解析：根据给定文本，箱线图主要用于可视化数据分布，表示连续数据的分布情况。

1. **散点图主要用于显示什么数据的分布情况？**

A. 分类型数据的分布

B. 连续数据的分布

C. 时间序列数据的分布

D. 地理空间数据的分布

答案：B

解析：根据给定文本，散点图主要用于显示数据点在笛卡尔坐标系统中的分布情况，每个点所对应的横/纵坐标代表的是该数据在对应维度上的属性值。

1. **维恩图主要用于表示什么数据集合运算的关系？**

A. 分类型数据的运算关系

B. 连续数据的运算关系

C. 时间序列数据的运算关系

D. 地理空间数据的运算关系

答案：A

解析：根据给定文本，维恩图（Venn Diagram）用于表示数据的集合运算的关系，使用环图形元素之间的重叠关系表示数据集合的并与交等集合运算。

1. **ALE图用于显示什么对机器学习模型预测结果的影响方式？**

A. 特征之间的相互影响方式

B. 样本数据的影响方式

C. 数据分布的影响方式

D. 特征对预测结果的影响方式

答案：D

解析：根据给定文本，ALE图（Accumulated Local Effects Plot）用于显示某一变量在整个值域上对机器学习模型的预测结果的影响——累积局部效应。

1. **什么是SHAP的主要理论依据？**

A. Shapley值

B. 博弈论

C. 机器学习算法

D. 数据可视化

答案：A

解析：根据给定文本，SHAP（SHapley Additive exPlanations）的主要理论依据为博弈论中的Shapley值（Shapley Values）。

1. **SHAP方法可以解释哪些机器学习模型的输出？**

A. 仅适用于决策树模型

B. 适用于任何机器学习模型的输出

C. 适用于神经网络模型的输出

D. 适用于线性回归模型的输出

答案：B

解析：根据给定文本，SHAP（SHapley Additive exPlanations）可以解释任何机器学习模型的输出，为每个特征分配一个特定预测的重要性值。

1. **Facets是谷歌公司为支持哪个项目发布的开源可视化工具？**

A. PAIR

B. TensorFlow

C. DeepMind

D. Google Brain

答案: A

解析: 文中明确提到Facets是谷歌公司为支持PAIR项目发布的可视化工具。

1. **Facets包括哪两个可视化部分？**

A. facet overview和facet dive

B. data analysis和feature distribution

C. data visualization和model exploration

D. statistical summary和interactive interface

答案: A

解析: 文中明确提到Facets包括facet overview和facet dive这两个可视化部分。

1. **facet overview通过什么方式对特征进行排序？**

A. 字母顺序

B. 特征重要性

C. 分布距离

D. 数据大小

答案: C

解析: 文中提到facet overview根据分布距离对特征进行排序。

1. **Facets Dive用于探索数据集中不同特性数据点之间的什么关系？**

A. 线性关系

B. 关联关系

C. 相似关系

D. 偏差关系

答案: B

解析: 文中明确提到Facets Dive用于探索数据集中不同特性数据点之间的关系。

1. **反事实解释的目标是找到改变给定预测所需的最小扰动。这里的反事实是什么？**

A. 对已知事实的重新假设

B. 对未知事实的假设

C. 对实际事实的改变

D. 对预测结果的猜测

答案: A. 对已知事实的重新假设

解析: 文中明确提到反事实是对既定事实重新假设来估计某一因素的影响概率。

1. **反事实解释方法的一个特征是什么？**

A. 对模型内部逻辑的解释

B. 最小化解释的内容

C. 提供可解释的特征

D. 满足XAI的法律义务

答案: B

解析: 文中明确提到反事实解释方法需要最小化需要解释的内容，以避免泄露商业秘密和侵犯个人隐私信息。

1. **反事实解释方法的应用可以支持用户的哪项权利？**

A. 拒绝模型的输出

B. 获得解释并提出质疑

C. 修改模型的预测结果

D. 请求模型的训练数据

答案: B

解析: 文中提到反事实解释方法有助于保障用户的知情权，为算法决策的合法性和合理性提供支持。

1. **反事实解释方法的应用流程中包括哪四个方面？**

A. 信任、真相、证明、敏感性

B. 信任、准确性、证据、敏感性

C. 信任、真相、证据、敏感性

D. 可信度、准确性、证明、敏感性

答案: A

解析: 文中明确提到反事实解释方法应用流程中包括信任、真相、证明、敏感性这四个方面。

1. **LIME是一种用于解释什么类型的模型的可解释性算法？**

A. 图像分类模型

B. 文本处理模型

C. 回归模型

D. 所有类型的模型

答案: D

解析: 文中明确提到LIME是一种模式无关的、局部解释的事后解释技术，可支持任何分类模型和回归模型的解释。

1. **LIME算法的解释过程中涉及哪两个模型？**

A. 复杂模型和线性模型

B. 决策树模型和逻辑回归模型

C. 随机森林模型和支持向量机模型

D. 神经网络模型和贝叶斯模型

答案: A

解析: 文中提到LIME算法通过一个复杂模型和一个线性模型来进行解释。

1. **LIME算法中，如何选择用于解释的特征？**

A. 基于特征重要性进行选择

B. 随机选择一定数量的特征

C. 由用户指定需要解释的特征

D. 使用所有可用特征进行解释

答案: C

解析: 文中没有明确提到具体的特征选择方式，但可以推断用户可以根据需要指定需要解释的特征。

1. **LIME算法通过什么方式对新生成的样本进行加权？**

A. 根据样本与预测点的距离

B. 根据特征重要性

C. 根据预测结果的置信度

D. 根据模型的准确性

答案: A

解析: 文中提到LIME算法通过计算新生成的样本与预测点之间的距离，并根据距离进行加权。

1. **LIME算法中的线性模型的权值用来做什么？**

A. 作为特征的重要性指标

B. 作为解释模型的参数

C. 作为样本的权重

D. 作为模型的预测结果

答案: B

解析: 文中提到LIME算法使用线性模型的权值作为决策解释。

1. **LIME算法对图像识别的解释中使用了什么技术对图像进行处理？**

A. 超像素分割

B. 特征提取

C. 数据采样

D. 模型训练

答案: A

解析: 文中提到LIME算法在图像识别中使用超像素分割对图像进行处理。

1. **LIME算法在解释黑盒模型时，解释的内容是针对什么进行排序和解释？**

A. 数据样本

B. 特征重要性

C. 模型预测结果

D. 训练数据集

答案: A

解析: 文中提到LIME算法对于黑盒模型进行解释时，是针对数据样本进行排序和解释。

1. **Anchors是一种基于什么规则的模型无关解释？**

A. if-then规则

B. while循环规则

C. for循环规则

D. switch-case规则

答案: A

解析: 根据提供的内容，Anchors是一种基于if-then规则的模型无关解释。

1. **Anchors的方法基于什么原理？**

A. 特征选择

B. 特征工程

C. 特征子集

D. 特征提取

答案: C. 特征子集

解析: 根据提供的内容，Anchors的方法基于选择最小的特征子集作为规则，作为预测结果的充分条件。

1. **Anchors相比于SHAP，主要具有什么优势？**

A. 更高的计算成本

B. 更广泛的适用性

C. 更高的精度

D. 更好的解释性

答案: B. 更广泛的适用性

解析: 根据提供的内容，Anchors相比于SHAP在计算方面不像SHAP那么费力，并且在解释非线性决策边界和提供易于解释的规则方面具有优势。

1. **反事实解释方法的目标是什么？**

A. 找到最小扰动

B. 解释算法的内部运行逻辑

C. 理解数据集的特征分布情况

D. 比较反例特征

答案: A

解析: 根据之前提供的内容，反事实解释方法的目标是找到改变给定预测所需的最小扰动。

1. **Anchors的应用范围不包括以下哪个领域？**

A. 文本分类

B. 结构化预测

C. 图像分类

D. 无监督学习

答案: D

解析: 根据提供的内容，Anchors在文本分类、结构化预测和图像分类等领域展现了良好的应用性和灵活性，但与无监督学习无关。

1. **CEM是一种模型无关的什么解释技术？**

A. 局部的事后解释技术

B. 全局的事后解释技术

C. 局部的事前解释技术

D. 全局的事前解释技术

答案: A

解析: 根据提供的内容，CEM是一种模型无关的局部的事后解释技术。

1. **CEM的关键思想是找到什么特征集？**

A. 相关正特征集

B. 相关负特征集

C. 假设正特征集

D. 假设负特征集

答案: A

解析: 根据提供的内容，CEM的关键思想是找到能够证明分类结果的合理性的最小且充分的相关正特征集。

1. **CEM引入了什么方法来评估数据流形的接近度？**

A. 卷积自动编码器

B. 循环神经网络

C. 支持向量机

D. 主成分分析

答案: A

解析: 根据提供的内容，CEM引入了卷积自动编码器来评估数据流形的接近度。

1. **A/B测试是一种什么类型的试验？**

A. 无监督学习试验

B. 有监督学习试验

C. 随机对照试验

D. 分层抽样试验

答案: C

解析: 根据提供的内容，A/B测试是一种随机对照试验。

1. **A/B测试的关键在于提出什么？**

A. 假设

B. 结论

C. 数据样本

D. 问题定义

答案: A

解析: 根据提供的内容，A/B测试的关键在于提出假设。

1. **混淆矩阵主要用于评估什么类型的算法性能？**

A. 无监督学习算法

B. 强化学习算法

C. 有监督学习算法

D. 半监督学习算法

答案: C

解析: 根据提供的内容，混淆矩阵主要用于评估有监督学习算法的性能。

1. **模型的精度是模型正确预测的样本占样本总数的比例，其计算公式是什么？**

A. TP/((TP+FP))

B. TP/((TP+FN))

C. ((TP+TN))/(（TP+TN+FP+FN）)

D. (2\* Precision\*Recall)/((Precision+Recall))

答案: C

解析: 根据提供的内容，模型的精度是模型正确分类的样本（TP+TN）与样本总数（TP+TN+FP+FN）的比率。

1. **模型的召回率是模型正确预测的正例样本占正例样本总数的比例，其计算公式是什么？**

A. TP/((TP+FP))

B. TP/((TP+FN))

C. ((TP+TN))/(（TP+TN+FP+FN）)

D. (2\* Precision\*Recall)/((Precision+Recall))

答案: B

解析: 根据提供的内容，模型的召回率是模型正确预测的正例样本占正例样本总数的比例。

1. **模型的F1值是精度和召回率的调和平均值，其计算公式是什么？**

A. TP/((TP+FP))

B. TP/((TP+FN))

C. ((TP+TN))/(（TP+TN+FP+FN）)

D. (2\* PrecisionRecall)/((Precision+Recall))

答案: D

解析: 根据提供的内容，模型的F1值是精度和召回率的调和平均值。

1. **AUC面积是ROC曲线围成的面积，用来评估模型性能，AUC面积越大表示什么？**

A. 模型性能越好

B. 模型性能越差

答案: A. 模型性能越好

解析: 根据提供的内容，AUC面积越大表示模型性能越好。

**二、简答题**

1. **什么是可视故事化？它与数据可视化和数据故事化有何区别？**

答案与解析：

可视故事化是指以可视化方法为主要叙事手段的故事创作过程，生成的故事被称为"可视故事"。它是数据可视化和数据故事化的交叉点。在可视故事化中，数据故事化是目的，而数据可视化是手段。数据可视化侧重于将数据要素可视化，而可视故事化更加强调数据产品的增值性和用户体验。

1. **请简要解释Shapley值的概念和作用。**

答案与解析：

Shapley值是博弈论中的概念，用于衡量联盟中每个成员对收益的**边际贡献**。在机器学习中，Shapley值被应用于**解释预测模型的输出**。它可以为每个特征分配一个特定预测的重要性值，帮助理解各个特征对预测结果的影响程度。Shapley值具有有效性、对称性、无效性和线性等特征，通过计算Shapley值，可以定量估算每个特征取值对预测结果的边际贡献。

1. **简要说明如何使用ALE图解释机器学习模型的预测结果。**

答案与解析：

ALE图（Accumulated Local Effects Plot）用于显示某一变量在整个值域上对机器学习模型的预测结果的影响。通过绘制特征值的取值范围与对应的模型预测结果的累积局部效应之间的关系曲线，可以观察到特征对预测结果的整体影响情况。ALE图可以帮助分析特征的非线性影响、交互效应以及特征在不同取值范围上的预测变化情况。通过解释ALE图中的趋势和变化，可以更好地理解特征与模型预测之间的关系。

1. **什么是Facets工具？**

答案: Facets是谷歌公司为支持PAIR项目发布的一个开源可视化工具，用于帮助用户理解和分析机器学习数据集。

解析: 根据提供的文本内容，Facets是一个可视化工具，旨在支持用户理解和分析机器学习数据集。它是谷歌公司为支持PAIR项目而发布的。

1. **Facets包括哪两个可视化部分？**

答案: Facets包括facet overview和facet dive两个可视化部分。

解析: 根据提供的文本内容，Facets由facet overview和facet dive两个可视化部分组成。这两个部分都提供了对数据集的直观概览和深入分析的功能。

1. **Facet overview主要用于什么目的？**

答案: Facet overview提供整个数据集的直观概览和数据每项特征的分布情况，便于用户快速了解数据集的特征分布情况。

解析: 根据提供的文本内容，Facet overview用于提供整个数据集的概览和数据特征的分布情况。它总结了每个特征的统计量，并比较了训练和测试数据集之间的差异。

1. **Facet dive提供了什么功能？**

答案: Facet dive提供一个易于自定义的界面，用于探索数据集中不同特征数据点之间的关系，帮助用户深入理解数据的单个特征。

解析: 根据提供的文本内容，Facet dive提供了一个可自定义的界面，让用户可以探索数据集中不同特征数据点之间的关系。它有助于用户深入理解数据的单个特征，并通过观察单个特征获取更多信息。

1. **反事实解释方法的目标是什么？**

答案: 反事实解释方法的目标是找到改变给定预测所需的最小扰动，通过对比反例特征告知用户哪些特征值影响了最终结果。

解析: 根据提供的文本内容，反事实解释方法的目标是找到改变给定预测所需的最小扰动，并通过对比反例特征告知用户哪些特征值对最终结果产生了影响。它不需要解释算法的内部运行逻辑，而是关注特定实例中的特征值变化。

1. **什么是欠拟合和过拟合？**

答案:欠拟合是指模型**无法捕捉**到数据中的关键特征和模式，导致在训练数据和新数据上表现都较差的情况。过拟合是指模型**过度学习**了训练数据中的噪声和细节，导致在训练数据上表现很好，但在新数据上表现较差的情况。

解析: 欠拟合和过拟合是机器学习中常见的问题。欠拟合表示模型过于简单，无法拟合数据的复杂性，而过拟合表示模型过于复杂，过度拟合了训练数据中的噪声和细节。

1. **请解释交叉验证的作用是什么？**

答案: 交叉验证的作用是评估机器学习模型的性能和泛化能力。它通过将数据集分为多个子集，在不同的子集上进行训练和验证，从而获取模型在不同数据子集上的性能指标，以减少对特定数据集的依赖性。

解析: 交叉验证是一种常用的模型评估方法，可以更好地评估模型在未见过数据上的表现，并帮助选择最佳的模型参数。

1. **什么是特征选择？为什么它在机器学习中很重要？**

答案: 特征选择是从原始特征集中选择出最具有预测能力的特征子集的过程。它在机器学习中很重要，因为：

可以提高模型的性能和泛化能力，减少过拟合风险。

可以降低模型的复杂度和训练时间。

可以提供更好的解释性和可解释性。

解析: 特征选择可以帮助我们减少数据维度和噪声特征的影响，从而提高模型的效果和解释性。

1. **什么是深度学习？它与传统机器学习有什么区别？**

答案: 深度学习是一种机器学习方法，通过构建深层神经网络来学习和表示数据中的复杂模式和特征。与传统机器学习相比，深度学习有以下区别：

深度学习可以自动学习更高级别的特征表示，而传统机器学习通常需要手动设计特征。

深度学习对大量数据的需求更大，而传统机器学习在小数据集上也可以表现良好。

深度学习模型通常具有更多的参数和更复杂的结构，需要更多的计算资源和训练时间。

解析: 深度学习是机器学习领域的一个重要分支，它通过多层神经网络的组合和训练来实现对复杂模式和特征的学习和提取。

1. **什么是生成对抗网络（GAN）？它的工作原理是什么？**

答案: 生成对抗网络（GAN）是一种包含生成器和判别器两个网络的框架，用于生成逼真的新样本。其工作原理如下：

生成器网络接收一个随机噪声向量作为输入，通过反向传播生成一个伪造的样本。

判别器网络接收生成器产生的样本和真实样本，通过训练来区分两者，并输出一个概率值表示样本的真实性。

生成器和判别器相互竞争、对抗，并通过反复训练来提高生成器生成逼真样本的能力，同时判别器也不断提高对真实和伪造样本的辨别能力。

解析: 生成对抗网络是一种创新的深度学习框架，用于生成逼真的数据样本，常应用于图像生成、文本生成等任务。