机器学习第三次作业
课后习题

2000094627
南宫圣光

1.5 *试述机器学习能在互联网搜索的哪些环节中起什么作用。*

(1) 查询理解（Query Understanding）:通过自然语言处理（NLP）技术识别关键词、实体、情感和语义关系。还使用分类模型预测用户的查询类型（如导航型、信息型、交易型）。纠正拼写错误或模糊查询的歧义。

(2) 文档检索（Document Retrieval）：使用机器学习模型（如向量搜索或深度学习模型）将文档和查询转化为向量表示，通过相似性计算进行检索。还利用预训练语言模型（如 BERT）匹配查询和文档，提高相关性。优化索引结构，提升检索效率。

(3) 排序（Ranking）：使用学习排序算法（Learning to Rank），如 LambdaMART 或神经网络模型，综合多种特征进行文档排序。还骑着结合点击率、停留时间等用户行为数据，优化排序效果。实现个性化推荐，根据用户兴趣和历史行为调整排序。

(4)结果生成（Result Generation）：使用生成式机器学习模型（如 GPT 系列）生成摘要、回答问题或推荐内容。自动提取网页中的重要信息（如标题、摘要、元数据）。还根据不同设备（如手机、PC）优化结果的格式和内容。

5. 用户行为分析（User Behavior Analysis）：使用聚类或分类算法分析用户的搜索路径和点击模式。通过强化学习优化搜索结果展示顺序，以最大化用户满意度。预测用户可能感兴趣的内容，提前推荐。

2.6 *试说错误率与ROC曲线的关系*

（1） FPR（假正率）是错误率的一部分，在 ROC 曲线中对应于 X 轴。

（2） 错误率衡量模型所有类型的预测错误，但 ROC 曲线通过显示敏感度与特异度之间的平衡来更详细地分析模型性能。

（3） ROC 曲线的面积（AUC, Area Under Curve）表示模型整体性能，错误率越低，AUC 值越大。

3.6 *线性判别分析仅在线性可分数据上能获得理想结果，试设计一个改进方法，使其能较好地用于非线性可分数据*

（1） **核方法（Kernel Method）**：通过将原始数据映射到高维特征空间，使非线性问题在高维空间中变为线性可分。具体方法:

使用核函数（如高斯核、多项式核）将原始特征映射到高维空间：$\phi(x): x \to \mathbb{R}^d$ 。此后在高维空间中应用线性判别分析。

（2） **非线性特征扩展**：在原始特征中添加多项式、交互项、或其他非线性特征。例如，将二维特征(x1,x2)扩展为 (x1,x2,x1^2,x2^2,x1x2)，然后在扩展后的特征上应用线性判别方法。

（3） **使用降维方法（如 t-SNE 或 PCA）**：在使用线性判别方法之前，先使用降维方法将数据分布的非线性结构揭示出来。通过降维后得到的特征输入线性判别方法模型。

（4） **集成方法**：结合多种分类器，例如结合决策树或支持向量机（SVM），处理线性判别方法无法处理的复杂数据分布。通过对线性判别方法和其他非线性分类器的结果进行加权融合提升性能。

（6） **深度学习方法**：使用深度神经网络提取数据的高维非线性特征，将这些特征传递给线性判别方法进行分类。深度学习模型如自动编码器（Autoencoder）可以有效处理非线性关系。

4.2 *试分析使用"最小训练误差"作为决策树划分选择准则的缺陷。*

（1） **容易导致过拟合**：因为训练误差的最小化会倾向于对训练数据进行过于复杂的划分，决策树可能会过度拟合训练数据中的噪声和异常值。因此导致虽然在训练数据上的误差很低，但在测试数据上的表现可能很差，导致泛化能力不足。

（2） **缺乏对模型复杂度的控制**：因为最小化训练误差只关注当前节点的划分，而忽略了树的深度和复杂度问题。所以模型可能生成非常深的树，增加了计算成本，并导致模型复杂度过高。

（3） **无法体现信息增益或纯度改进**：决策树的核心目标是通过划分最大化信息增益或最小化纯度的混杂程度（如基尼系数或熵），但"最小训练误差"只关注当前划分后样本分类的错误率，无法准确反映划分质量。这种方法可能会选择并非最佳的信息增益划分点，导致决策树结构欠佳。

（4） **缺乏对数据整体的优化**：仅以训练误差为目标的划分策略无法保证决策树对整个数据集的优化，容易忽视后续节点之间的相互影响。局部优化的划分方式可能会导致整棵树的全局性能下降。

5.1 *试述将线性函数 f(x) = w^T * x用作神经元激活函数的缺陷。*

（1） **表达能力有限**：线性函数的输出是输入的线性组合，整个神经网络无论多少层都只能表示一个线性变换。因此无论网络的深度如何，使用线性激活函数的神经网络本质上相当于单层线性模型，无法学习到复杂的非线性关系。

（2） **无法处理非线性数据**：线性激活函数无法将数据映射到非线性特征空间。因此对于复杂的分类问题（如非线性可分数据），网络无法找到合适的决策边界。

（3） **激活值不受限制**：线性函数的输出范围是整个实数集，没有任何限制或归一化。对于深层神经网络，可能导致权重更新幅度过大或梯度爆炸，训练过程变得不稳定。输出范围无限，可能不适合用于概率预测等任务。

（4） **缺乏激励效果**：线性函数的梯度是一个常数，与输入无关。因此梯度不会随输入值变化，这会使权重更新始终是固定的，限制了模型的学习能力。无法有效利用梯度下降算法进行复杂的优化。

6.6 *试析 SVM 对噪声敏感的原因。*

（1） **对边界点高度依赖**：因为SVM通过支持向量（距离分类边界最近的点）来决定分类边界的位置。所以如果数据中存在噪声点（即错误标注或异常值），这些点可能会被错误地选为支持向量，从而大幅度影响决策边界。即使是少量的噪声，也可能显著降低模型的分类性能。

（2） **使用硬间隔时对异常值的敏感性**：在硬间隔（Hard Margin）条件下，SVM严格要求数据点与分类边界的距离必须满足某些条件。因此噪声点可能无法满足硬间隔的约束，导致模型过于复杂，甚至无法收敛。

（3） **核函数的影响**：SVM依赖于核函数将数据映射到高维特征空间。在高维空间中，噪声点可能与正常点的分布更加混杂。因此由于核方法将数据映射到更高维度，噪声点的影响可能被放大，导致分类器的性能下降。

（4） **惩罚系数C的影响**：因为SVM的目标函数中包含一个惩罚系数 C（用于平衡间隔的大小和误分类点的惩罚。）如果C值过大，SVM会试图严格分类所有训练样本，包括噪声点，导致过拟合。所以如果数据中噪声比例较高，分类边界会受到显著干扰。

7.2 试证明：条件独立性假设不成立时，朴素贝叶斯分类器仍有可能产生最优贝叶斯分类器。

朴素贝叶斯分类器（Naive Bayes Classifier，NBC）：

假设特征之间条件独立，即：

$$P(X_1, X_2, \ldots, X_n \mid Y) = \prod_{i=1}^{n} P(X_i \mid Y)$$

基于此简化假设计算类别的后验概率 P(Y｜X)，然后选取概率最大的类别作为预测结果。

最优贝叶斯分类器（Optimal Bayesian Classifier，OBC）：

最优贝叶斯分类器通过真实的联合概率分布P(X1，X2，……，Xn｜Y) 完全计算后验概率 P(Y｜X)，然后进行分类。假设条件独立性不成立时，朴素贝叶斯分类器的简化可能会导致分类结果与最优贝叶斯分类器不同。

虽然条件独立性假设不成立，但朴素贝叶斯分类器仍可能与最优贝叶斯分类器一致，其原因包括以下两种情况：

(1) **后验概率排序一致性**：朴素贝叶斯分类器只需确保对每个类别Y的后验概率排序与最优贝叶斯分类器一致即可保证分类结果一致。即使条件独立性假设不成立，只要：

$$\arg\max_y P(Y = y \mid X) = \arg\max_y \prod_{i=1}^{n} P(X_i \mid Y = y) P(Y = y)$$

成立，朴素贝叶斯分类器仍能产生与最优贝叶斯分类器相同的分类结果。

(2) **特征之间的相关性对分类无关紧要**：如果特征之间的相关性不会改变不同类别之间的相对后验概率，则条件独立性假设的违背不会影响分类结果。例如，在某些数据分布下，特征的相关性可能对 P(Y｜X) 的分类决策没有实质性影响。

所以朴素贝叶斯分类器可能最优的实际情况可以归纳为：

（1） **特征相关性在类别间一致**：如果所有类别下特征间的相关性模式相同，则条件独立性假设的违背不会影响分类。

（2） **数据特征对后验概率贡献较小**：某些情况下，特征的相关性对分类决策的影响很小，朴素贝叶斯分类器仍可以获得近似最优的结果。

用数学证明：

假设有两个类别 $Y \in \{C1, C2\}$，特征为$X1, X2$。实际联合分布为：$P(X1, X2 \mid Y)$

朴素贝叶斯的简化假设为：$P(X1, X2 \mid Y) \approx P(X1 \mid Y) P(X2 \mid Y)$

即使上述近似不成立，但如果：

$$\arg\max_Y P(Y) P(X_1 \mid Y) P(X_2 \mid Y) = \arg\max_Y P(Y) P(X_1, X_2 \mid Y)$$

朴素贝叶斯分类器的分类结果仍然是最优的。

8.4 *GradientBoosting [Friedman, 2001] 是一种常用的 Boosting 算法，试分析其与 AdaBoost 的异同。*

(1) **损失函数的优化方式**:

AdaBoost：优化指数损失函数，通过提高错误分类样本的权重来引导新的弱学习器更好地关注这些样本。

Gradient Boosting：采用梯度下降法直接最小化用户指定的损失函数（如平方误差、对数损失），每轮训练的新模型用于修正当前模型的残差。

(2) **样本权重更新**:

AdaBoost：错误分类的样本权重会被放大，正确分类的样本权重会被减小。

Gradient Boosting：没有显式调整样本权重，而是根据当前模型的负梯度（残差）更新。

(3) **异常值的敏感性**:

AdaBoost：异常值会导致权重异常增加，可能过度拟合这些噪声数据。

Gradient Boosting：通过逐步优化残差，对异常值的鲁棒性更强。

(4) **适用范围**:

AdaBoost：更适合分类问题（默认使用指数损失）。

Gradient Boosting：可以灵活定义损失函数，广泛用于分类和回归任务。

(5) **模型复杂性控制**:

AdaBoost：模型复杂度控制主要依赖于弱学习器的数量。

Gradient Boosting：可以通过学习率、树的深度等参数精细控制模型复杂性。

9.6 *试析 AGNES 算法使用最小距离和最大距离的区别。*

特性	最小距离	最大距离
距离定义	两簇中最近点的距离	两簇中最远点的距离
适合的簇形状	非凸形状簇	紧密且均匀的簇
对噪声点的敏感性	对噪声点较敏感	对噪声点鲁棒性较高
链式效应	容易发生链式效应	不容易发生链式效应
计算复杂度	较低	较高

11.6 *试分析岭回归与支持向量机的联系。*

(1) 相似性：岭回归与支持向量机都采用了$L2$正则化，通过限制参数的大小提升模型的鲁棒性，目标函数由正则化项和损失函数组成，且均是凸优化问题。

(2) 差异性：岭回归关注的是全局的平方误差最小化，而支持向量机特别强调边界间隔的最大化，并采用不同的损失函数（如ϵ-不敏感损失）以处理不同的任务需求。