习题集 2024

1 机器学习概述

1. 监督学习中的任务一般可分为分类问题和回归问题。请简述这两种任务的定义与区别。
   答案： 分类问题：给定训练样本集 $\{(x_i, y_i)\}_{i=1,2,...,n}$, 学习 x 到 y 的映射，其中 y 是离散值。
   回归问题：给定训练样本集 $\{(x_i, y_i)\}_{i=1,2,...,n}$, 学习 x 到 y 的映射，其中 y 是连续值。
   区别：分类强调依据类别标签 y 对样本 x 空间的划分，回归强调 x 与回归值的拟合。

2. 下面属于分类任务的是 (C)

   (A) 股票预测

   (B) 房价预测

   (C) 目标检测

   (D) 西瓜的糖含量预测

3. 请简述聚类任务的定义。
   答案： 给定训练样本集 $\{x_i, i = 1, .., n\}$, 学习 x 到类簇 y 的映射。

2 Bayes 学习

1. 请简述什么是朴素贝叶斯分类器，并给其公式。
   答案： 朴素贝叶斯分类器是一种基于贝叶斯定理的分类算法，其特点在于采用了“属性条件独立性假设”。
   公式如下：
   $P(c \mid x) = \frac{P(c)P(x|c)}{P(x)} \propto P(c)P(x \mid c) = P(c)\prod_{i=1}^{d} P(x_i \mid c)$

2. 下列哪个选项为朴素贝叶斯公式 (B)

   (A) $P(c \mid x) \propto P(c)P(x \mid c) = P(c)\prod_{i=1}^{d} P(x_i \mid c, x_i^p)$

   (B) $P(c \mid x) \propto P(c)P(x \mid c) = P(c)\prod_{i=1}^{d} P(x_i \mid c)$

   (C) $P(c \mid x) \propto P(c)P(x \mid c) = \sum_{i=1}^{d} P(c, x_i)\prod_{j=1}^{d} P(x_j \mid c, x_i)$

   (D) $P(c \mid x) = \sum_{i=1}^{d} P(c)P(x_i \mid c)$

3. 贝叶斯二分类问题中的各类样本均为多维正态分布，在 (C) 条件下，决策界为线性决策界。

(A) 各类多维正态分布的均值相等

(B) 各类多维正态分布的方差相等

(C) 各类多维正态分布的协方差矩阵相等

(D) 各类多维正态分布的峰值相等

4. 已知贝叶斯公式 $P(M \mid D) = \frac{P(D|M)P(M)}{P(D)}$，下列选项属于极大似然估计的为 (A)

(A) $\hat{M} = \arg\max_M P(D \mid M)$

(B) $\hat{M} = \arg\max_M P(M \mid D) \propto P(M)P(D \mid M)$

(C) $\hat{M} = E(P(M \mid D)$

(D) $\hat{M} = \arg\max_M E(P(M))$

5. 已知贝叶斯公式 $P(M \mid D) = \frac{P(D|M)P(M)}{P(D)}$，下列选项属于最大后验估计的为 (B)

(A) $\hat{M} = \arg\max_M P(D \mid M)$

(B) $\hat{M} = \arg\max_M P(M \mid D) \propto P(M)P(D \mid M)$

(C) $\hat{M} = E(P(M \mid D)$

(D) $\hat{M} = \arg\max_M E(P(M))$

3 线性分类

1. 请简述向量内积运算法则和几何意义。

答案:

对于向量 $\mathbf{a}, \mathbf{b}$:

$$a = [a_1, a_2, ..., a_n]$$

$$b = [b_1, b_2, ..., b_n]$$

a 和 b 的内积公式为:

$$a^T b = a_1 b_1 + a_2 b_2 + \ldots + a_n b_n$$

内积的几何意义为: 可以用来表征或计算两个向量之间的夹角，以及在 b 向量在 a 向量方向上的投影。

2. 向量的 P 范数距离公式? 2 范数公式? 1 范数公式? 无穷范数公式?

答案:

L1 范数

$$\|x\|_1 = \sum_i |x_i|$$

L2 范数

$$\|x\|_2 = \sqrt{\sum_i x_i^2}$$

无穷范数

$$\|x\|_\infty = \max(|x_i|)$$

Lp 范数, 描述为

$$L_p = \sqrt[p]{\sum_{i=1}^{n} |x_i|^p}$$

当 p 取 0，1，2 时即对应着 L0, L1 L2 范数。

3. 超平面的表达式和线性判别函数分别是什么，并阐述线性判别函数值的几何含义。

答案:

超平面表示为:

$$w^T x + b = 0$$

其中 w 是权重向量，x 是特征向量，b 是偏置项。

线性判别函数:

$$f(x) = w^T x + b$$

线性判别函数值的几何含义是点 x 到超平面的有符号距离，正值表示在超平面法向量一侧，负值表示在超平面另一侧，为零表示在超平面上。

4. 余弦相似性的公式是什么?

答案:

对于两个向量 $\vec{a}, \vec{b}$ 余弦相似度公式为:

$$\cos(\vec{a}, \vec{b}) = \frac{\vec{a} \cdot \vec{b}}{|\vec{a}||\vec{b}|}$$

5. 对于一个向量 a=[1,2,3], 请计算它的 L1 范数 L2 范数，并计算 a=[1,2,3] 与向量 b=[-1,3,4] 的余弦相似性。

答案:

1. L1 范数:

$$\|a\|_1 = |1| + |2| + |3| = 6$$

2. L2 范数:

$$\|a\|_2 = \sqrt{1^2 + 2^2 + 3^2} = \sqrt{14}$$

3. 余弦相似性:

$$\text{cosine similarity } = \frac{a^T b}{\|a\|_2 \cdot \|b\|_2} = \frac{17}{\sqrt{364}}$$

其中, $a^T b = 1 \cdot (-1) + 2 \cdot 3 + 3 \cdot 4$ 。

6. 假设我们有两维的输入数据, 感知机的线性判别函数为 $f(x) = w^T x + b$, 初始参数为 $w = [0,0]$, 阈值（偏置）$b = 0$。感知机优化的目标函数为 $J(w) = \sum_{x_i \in E} -w^T x_i y_i$, 其中 E 为错误分类样本集，优化过程的学习率（步长）为 $\eta = 1$ 。

训练样本如下: 1. 输入 $[1,2]$, 标签 $y=1$; 2. 输入 $[2,3]$, 标签 $y=-1$; 3. 输入 $[-1,-1]$, 标签 $y=1$ 。现在我们按照随机梯度下降的方式进行模型学习, 迭代更新参数。迭代更新的规则为:

$$w \leftarrow w - \eta \nabla J(w)$$
$$b \leftarrow b - \eta \nabla J(b)$$

请给出 6 步迭代后的线性判别函数。

答案:

第一步迭代:

- 输入 $[1,2]$, 标签 $y=1$
- 预测: $f([1,2]) = [0,0]\cdot[1,2]+0=0$
- 预测错误, 更新参数: $w \leftarrow [0,0]+1\cdot 1\cdot[1,2]=[1,2], b \leftarrow 0+1\cdot 1=1$

第二步迭代:

- 输入 $[2,3]$, 标签 $y=-1$
- 预测: $f([2,3]) = [1,2]\cdot[2,3]+1=9>0$
- 预测错误, 更新参数: $w \leftarrow [1,2]+1\cdot(-1)\cdot[2,3]=[-1,-1], b \leftarrow 1+1\cdot(-1)=0$

第三步迭代:

- 输入 $[-1,-1]$, 标签 $y=1$
- 预测: $f([-1,-1]) = [-1,-1]\cdot[-1,-1]+0=2>0$
- 由于预测正确, 参数不需要更新。

第四步迭代:

- 输入 $[1,2]$, 标签 $y=1$
- 预测: $f([1,2]) = [-1,-1]\cdot[1,2]+0=-3<0$
- 预测错误, 更新参数: $w \leftarrow [-1,-1]+1\cdot 1\cdot[1,2]=[0,1], b \leftarrow 0+1\cdot 1=1$

第五步迭代:

- 输入 $[2,3]$, 标签 $y=-1$
- 预测: $f([2,3]) = [0,1]\cdot[2,3]+1=4>0$
- 预测错误, 更新参数: $w \leftarrow [0,1]+1\cdot(-1)\cdot[2,3]=[-2,-2], b \leftarrow 1+1\cdot(-1)=0$

第六步迭代:

- 输入 $[-1,-1]$, 标签 $y=1$
- 预测: $f([-1,-1]) = [-2,-2]\cdot[-1,-1]+0=4>0$
- 由于预测正确, 参数不需要更新。

最终的线性判别函数为 $f(x)=[-2,-2]\cdot x$ 。

7. Fisher 鉴别分析中类内散度矩阵的刻画了什么统计信息?

答案:

类内散度矩阵刻画了类别内部样本的松散程度，主要用于度量同一类别内样本点的离散程度。类内散度矩阵描述了类别内样本点之间的协方差关系，散度矩阵的特征向量代表了数据在特征方向上的方差。

8. Logistic 模型的目标函数是什么估计？(C)

   (A) 均方误差估计

   (B) 平均绝对误差估计

   (C) 最大似然估计

   (D) 最小二乘法估计

9. 线性分类器的任务是, 给定训练样本集 $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}$, 其中 $(x_i, y_i) \in \mathbb{R}^d \times \mathbb{R}, y_i$ 是类别标签, 目标是学习 $w_0, \mathbf{w}$ 。以下哪一项表示线性分类器分类错误 (D)。

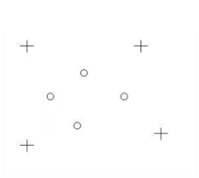
   (A) $w^T x_i + w_0 \geq 0, For\ all\ i, such\ that\ y_i = +1$

   (B) $w^T x_i + w_0 \leq 0, For\ all\ i, such\ that\ y_i = -1$

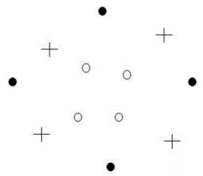
   (C) $Together\ y_i(w^T x_i + w_0) > 0$

   (D) $Together\ y_i(w^T x_i + w_0) < 0$

10. 下面哪些是线性不可分问题？(AB)



A



B

C

D

4 非线性分类

1. 决策树模型在特征空间中的决策界通常是由什么样的几何形式构成的（C）

   (A) 由圆形边界围成

   (B) 由多边形边界围成

   (C) 由垂直于坐标轴的超平面围成

   (D) 由曲线边界围成

2. 以下哪些量化指标常用作决策树中的节点选择依据（ABC）

(A) 信息增益

(B) 信息增益率

(C) 基尼指数

(D) 线性回归系数

3. 在决策树剪枝后，剪枝节点的类别非纯度相比剪枝前该分枝上各个节点的平均非纯度通常会怎样变化（A）

(A) 显著增加

(B) 显著减少

(C) 保持不变

(D) 无法预测

4. 在决策树模型中，给定一个数据集，其中每个属性的取值数量分别为 $V_1, V_2, \ldots, V_n$，其中 n 是属性的总数。如果以属性为问题，那么决策树中候选问题的总数是多少（A）

(A) n

(B) $V_1 + V_2 + \ldots + V_n$

(C) $V_1 \times V_2 \times \ldots \times V_n$

(D) $\max(V_1, V_2, \ldots, V_n)$

5. 假设有以下 4 个训练样本点：

样本点	x_1	x_2	类别
1	1	2	A
2	2	3	A
3	4	2	B
4	5	3	B

现有一个测试样本点 (3, 2)。使用 KNN 算法对其进行分类，假设 $K = 3$，请计算该测试样本点的类别。

计算过程和答案：

首先，计算测试样本点与每个训练样本点之间的欧氏距离 $d(x,y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}$。然后，选出距离最小的 3 个邻居，并通过投票确定测试样本点的类别。

样本点之间的距离计算如下表所示：

样本点	距离
1	$\sqrt{(3-1)^2 + (2-2)^2} = 2$
2	$\sqrt{(3-2)^2 + (2-3)^2} = \sqrt{2}$
3	$\sqrt{(3-4)^2 + (2-2)^2} = 1$
4	$\sqrt{(3-5)^2 + (2-3)^2} = \sqrt{5}$

选出的最近的 3 个邻居为样本点 1、2 和 3。其中，样本点 1 和 2 属于类别 A，样本点 3 属于类别 B。因此，测试样本点的类别为 A。

6. 列举 3 种以上集成学习的集成策略
   答案： 几何平均、算数平均、bagging、boosting

7. 以下关于支持向量机（SVM）表述正确的是（ACD）

   (A) SVM 基于最大间隔思想，旨在找到最优的决策边界，使得不同类别之间的间隔最大化。

   (B) SVM 仅适用于线性可分的数据集，不能处理非线性数据。

   (C) SVM 使用核函数原理将数据映射到更高维空间以解决非线性问题。

   (D) SVM 基于经验风险最小化原则，优化模型以减少训练集上的分类错误。

8. Kullback-Leibler 散度（KL 散度）衡量两个概率分布 $p(x)$ 和 $g(x)$ 之间的相似性，给出其数学表达式。
   答案：
   KL 散度的数学表达式为：

$$D_{KL}(p \parallel g) = \sum_x p(x) \log\left(\frac{p(x)}{g(x)}\right)$$
$$D_{KL}(g \parallel p) = \sum_x g(x) \log\left(\frac{g(x)}{p(x)}\right)$$

   对于连续变量，求和变为积分：

$$D_{KL}(p \parallel g) = \int p(x) \log\left(\frac{p(x)}{g(x)}\right) dx$$
$$D_{KL}(g \parallel p) = \int g(x) \log\left(\frac{g(x)}{p(x)}\right) dx$$

   其中，

$$D_{KL}(p \parallel g) \neq D_{KL}(g \parallel p)$$

5 回归

1. 下面属于监督模型的有 (BCD)

   (A) 聚类

   (B) 线性回归

   (C) 支持向量机

   (D) 决策树

2. 回归分析中，误差可以分解为 (ABD) 之和

   (A) 偏差

   (B) 方差

(C) 互信息

(D) 噪声

3. 在多元线性回归中，最小二乘估计的参数向量 $\theta = [\theta_0, \theta_1, \theta_2, ..., \theta_m]$ 可以通过以下哪个公式获得? (B)

(A) $\theta = \mathbf{X}^T\mathbf{y}$

(B) $\theta = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$

(C) $\theta = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$

(D) $\theta = \frac{\sum_{i=1}^{n}(y_i-(\theta_0+\theta_1 x_{i1}+\theta_2 x_{i2}+...+\theta_m x_{im}))^2}{n}$

4. 在机器学习中，损失函数通常由两部分组成，一部分是数据拟合项，另一部分是正则项。正则项在损失函数中的作用是 (C)

(A) 提高模型的偏差，使其更好地拟合训练数据。

(B) 帮助模型更好地拟合训练数据，减小方差。

(C) 控制模型的复杂度，防止过拟合。

(D) 提高模型的方差，使其更灵活地适应不同的数据分布。

6 聚类

1. 以下哪种聚类算法可以进行在线学习（C）

(A) K-Means

(B) 层次聚类

(C) 序贯方法

(D) 高斯混合模型

2. 哪种聚类算法的聚类过程呈现层次结构（B）

(A) K-Means

(B) 层次聚类

(C) DBSCAN

(D) 高斯混合模型

3. 在以下聚类算法中，哪一个算法是基于目标函数的方法（A）

(A) K-Means

(B) 层次聚类

(C) DBSCAN

(D) 高斯混合模型

4. 请列举至少两种用于衡量样本和类簇相似性的公式，并简要描述它们。

答案：

集合为离散点集：

1. 样本到集合最远点距离: $d(x,C) = \max_{y\in C} d(x,y)$
2. 样本到集合最近点距离: $d(x,C) = \min_{y\in C} d(x,y)$
3. 样本到集合平均点距离: $d(x,C) = \frac{1}{|C|}\sum_{y\in C} d(x,y)$

集合为连续区域：

1. 集合为平面：$d(x,H) = \min_{z\in H} d(x,z)$
2. 集合为圆：$d(x,Q) = \min_{z\in Q} d(x,z)$

5. 请列举至少两种用于衡量类簇间相似性的公式。

答案：

1. 类簇间最大距离

$$d_{\max}(C_i,C_j) = \max_{x\in C_i, y\in C_j} d(x,y)$$

2. 类簇间最小距离

$$d_{\min}(C_i,C_j) = \min_{x\in C_i, y\in C_j} d(x,y)$$

3. 类簇间平均距离

$$d_{\text{avg}}(C_i,C_j) = \frac{1}{|C_i|\cdot|C_j|}\sum_{x\in C_i, y\in C_j} d(x,y)$$

4. 类簇间表征点间距离

$$d_{\text{cen}}(C_i,C_j) = d(\mu_i,\mu_j)$$

μ 代表簇 C 的中心点 $\mu = \frac{1}{|C|}\sum_{1\leqslant i\leqslant |C|} x_i$

7 特征选择与降维

1. 描述机器学习中的前向序贯特征选择 (Sequential forward selection) 和后向序贯特征选择 (Sequential Backward Selection) 的过程并说明两者的区别。

答案：

前向序贯：每次加入一个特征，该特征使得新的特征组合最优。一旦增加特征，就无法删除。

后向序贯：每次减掉一个特征，使剩余特征组合最优。一旦删除特征，就无法增加。

区别：前向序贯特征选择是从无到有逐步添加特征，而后向序贯特征选择则是从全到无逐步删除特征。

2. 下面哪个特征评估方法属于距离可分性准则 (A)

(A) $J = \frac{|S_b - S_w|}{|S_w|}$

(B) $J = -\ln\int [p(x\mid\omega_1)\,p(x\mid\omega_2)]^{1/2}\,dx$

(C) $J = \int_x |p(x \mid \omega_1) - p(x \mid \omega_2)|^s \, p(x)^{1-s} dx$

(D) $I(X;Y) = \sum_{x \in S_x} \sum_{y \in S_y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$

3. 下面哪个特征评估方法属于概率可分性准则 (BC)

(A) $I(X;Y \mid Z) = \sum_{x \in S_x} \sum_{y \in S_y} \sum_{z \in S_z} p(x,y,z) \log \frac{p(x,y|z)}{p(x|z)p(y|z)}$

(B) $J = \int_x [p(x \mid \omega_1) - p(x \mid \omega_2)] \ln \frac{p(x|\omega_1)}{p(x|\omega_2)} dx$

(C) $J = -\ln \int p^s(x \mid \omega_1) \, p^{1-s}(x \mid \omega_2) \, dx$

(D) $I(X;Y) = \sum_{x \in S_x} \sum_{y \in S_y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$

4. 请简答特征选择和特征变换的区别
答案： 特征选择是在原始特征集合中选择出最有价值的特征子集，而特征变换则是通过对原始特征进行变换，生成新的特征空间。

5. 下面哪个是特征选择的框架 (ABC)

(A) Filter 方法

(B) Wrapper 方法

(C) Embedded 方法

(D) 主成分分析 (Principal component analysis)

6. 下面哪个是特征变换的方法 (ABCD)

(A) 线性鉴别分析 (LDA)

(B) 主成分分析 (Principal component analysis)

(C) 非负矩阵分解

(D) 局部线性变换

7. 请简答：主成分分析方法中提取的哪个特征方向是最优的?
答案： 标准化样本的协方差矩阵的最大特征值对应的特征方向。

8 信息论

1. 请写出信息熵以及互信息的公式。
答案：
信息熵:
$H(X) = -\sum_{i=1}^{n} p(x_i) \log p(x_i)$
互信息：$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$

2. 信息论中的哪些原则可作为机器学习优化原则？(ABD)

(A) 最大熵模型

(B) 最大互信息模型

(C) 最小熵模型

(D) 最小互信息模型

3. 以下哪些是正确的信息论模型？(ABC)

(A) 最大条件熵模型

(B) 独立成分分析

(C) 最大输入输出互信息

(D) 马尔可夫模型

9 概率图

1. 简述什么是概率图模型，并区分有向和无向概率图模型。
答案： 概率图模型是一种统计模型，用于表示多变量之间的条件依赖关系。有向概率图模型（贝叶斯网络）通过有向边表示变量间的因果关系，而无向概率图模型（马尔可夫随机场）通过无向边表示变量间的关联性。

10 深度学习

1. 请简述梯度消失现象的原因。
答案： 在使用传统的激活函数（如 sigmoid 或 tanh）时，这些激活函数是双线性饱和的。在这些饱和区域，梯度的值非常小。在梯度传递时，随着层数的增加，梯度会逐渐接近于 0，即梯度消失现象。
这可以表示为：
当 n（网络层数）增加时，$[\nabla f(x)]^n \to 0$ ，其中 $\nabla f(x)$ 是激活函数的梯度。

2. 假设你正在设计一个卷积神经网络（CNN），并且正在处理以下层的参数配置：

- 输入图像的尺寸为 $32 \times 32 \times 3$（高度 × 宽度 × 深度）。
- 第一个卷积层使用了尺寸为 5×5 的卷积核，步长为 1，没有填充（padding = 0），并有 10 个这样的卷积核。
- 紧接着第一个卷积层的是一个尺寸为 2×2 的最大池化层，步长为 2。

要求：

(a) 计算第一个卷积层后输出特征图的尺寸。

(b) 计算通过最大池化层后输出特征图的尺寸。

答案：

(a) **第一个卷积层的输出特征图尺寸**：

- 使用公式 $\left\lfloor \frac{W-F+2P}{S}+1 \right\rfloor \times \left\lfloor \frac{H-F+2P}{S}+1 \right\rfloor$，其中 W 和 H 是输入的宽度和高度，F 是卷积核的大小，P 是填充，S 是步长。在本题中，$W=H=32$，$F=5$，$P=0$，$S=1$。
- 因此，输出特征图的尺寸为：$\left\lfloor \frac{32-5+0}{1}+1 \right\rfloor \times \left\lfloor \frac{32-5+0}{1}+1 \right\rfloor = 28 \times 28$。
- 所以，第一个卷积层后的输出特征图尺寸为 $28 \times 28 \times 10$，因为有 10 个卷积核。

(b) **最大池化层的输出特征图尺寸**：

- 对于最大池化层，我们同样使用类似的公式。池化核的大小是 2×2 且步长为 2，因此 $F=2$，$S=2$。
- 应用公式：$\left\lfloor \frac{28-2}{2}+1 \right\rfloor \times \left\lfloor \frac{28-2}{2}+1 \right\rfloor = 14 \times 14$。
- 所以，经过最大池化层后的输出特征图尺寸为 $14 \times 14 \times 10$。

3. 循环神经网络（Recurrent Neural Network，RNN）能处理什么样的问题（ABC）

(A) 序列对序列

(B) 序列编码

(C) 序列解码

(D) 长期（Long-Term）依赖关系

4. BP 算法即反向传播算法，通常用于（C）

(A) 加快深度学习模型的推理速度

(B) 提高机器学习模型的解释性

(C) 训练神经网络，通过计算误差梯度并更新网络权重

(D) 优化神经网络的超参数配置

5. 关于 BERT 模型，以下哪些陈述是正确的（ACD）

(A) BERT 使用了 Transformer 的编码器结构。

(B) BERT 只能用于文本分类任务。

(C) BERT 采用了掩码语言模型（Masked Language Model, MLM）预训练方法。

(D) BERT 在预训练阶段使用了“下一个句子预测”（Next Sentence Prediction, NSP）任务。

6. Transformer 模型中的哪一部分显著地改进了对长距离依赖关系的处理能力（A）

(A) 自注意力机制（Self-Attention Mechanism）

(B) 前馈神经网络

(C) 层归一化（Layer Normalization）

(D) 编码器-解码器架构

7. 在生成对抗网络（GAN）中，以下哪种描述是正确的（C）

(A) 生成器的主要目的是提高分类准确性。

(B) 判别器通过生成新的数据样本来欺骗生成器。

(C) 生成器的主要目的是生成逼真的数据样本，以欺骗判别器。

(D) 判别器和生成器是独立训练，没有相互作用。

8. 关于线性鉴别分析的描述最准确的是，找到一个投影方向，使得（B）

(A) 类内距离最大，类间距离最小

(B) 类内距离最小，类间距离最大

(C) 类内距离最大，类间距离最大

(D) 类内距离最小，类间距离最小

9. SVM 的原理的简单描述，可概括为（C）

(A) 最小均方误差分类

(B) 最小距离分类

(C) 最大间隔分类

(D) 最近邻分类

10. 假定你使用阶数为 2 的线性核 SVM，将模型应用到实际数据集上后，其训练准确率和测试准确率均为 100%。现在增加模型复杂度（增加核函数的阶），会发生以下哪种情况（A）

(A) 过拟合

(B) 欠拟合

(C) 什么都不会发生，因为模型准确率已经到达极限

(D) 以上都不对

11. 集成学习中基分类器如何选择，学习效率通常越好（D）

(A) 分类器相似

(B) 都为线性分类器

(C) 都为非线性分类器

(D) 分类器多样，差异大

11 大语言模型相关

1. 在大语言模型中，如何通过“位置编码”（Positional Encoding）处理序列的位置信息？为什么它是必要的?

答案: 由于 Transformer 架构本身缺乏处理序列顺序的机制，位置编码通过为每个词添加一个表示其在序列中位置的向量，使模型能够感知词语的顺序关系。位置编码通常与词嵌入相加，帮助模型理解词语的相对和绝对位置。

2. Transformer 架构中的“自注意力机制”（Self-Attention）在处理长文本时的优势是什么?
   答案：自注意力机制能够捕捉输入序列中各个位置之间的依赖关系，无论距离远近。相比于传统的 RNN 或 LSTM，Transformer 可以并行处理整个序列，显著提高了处理长文本的效率和能力。

3. 在大语言模型中，为什么通常使用“因果语言建模”（Causal Language Modeling）而不是传统的序列预测任务?
   答案：因果语言建模通过仅依赖前文来预测下一个词，符合语言生成的因果性，避免了信息泄露问题。传统的序列预测任务可能需要同时考虑前后文，而因果建模更适合生成任务。

4. 大语言模型中如何处理超大规模的参数（如数百亿或更多参数)?
   答案：处理超大规模参数的方法包括：
   - 模型并行：将模型的不同部分分布到多个设备上。
   - 数据并行：将数据分割，分别在不同设备上训练相同的模型副本。
   - 混合精度训练：使用较低精度的数据类型（如 FP16）来减少内存占用。
   - 梯度累积：在多个小批次上累积梯度，然后再进行一次更新，以减少内存需求。

5. “稀疏注意力”（Sparse Attention）在大语言模型中的应用有哪些？它如何提升效率?
   答案：稀疏注意力通过限制每个词只关注输入序列中的部分词，减少了计算复杂度和内存消耗。常见的稀疏模式包括局部窗口注意力和固定模式注意力，适用于处理长序列。

6. 大语言模型使用的“知识蒸馏”（Knowledge Distillation）技术在训练时有什么特别的优势?
   答案：知识蒸馏通过让小模型模仿大模型的行为，传递其知识，减少小模型的训练数据需求，同时保留大模型的性能。对于资源受限的环境，蒸馏可以显著提高小模型的效果。

7. 为什么大语言模型通常使用“解码器”（Decoder）部分进行文本生成，而不是完整的编码器-解码器架构?
   答案：在文本生成任务中，解码器通过自回归生成的方式可以逐步生成下一个词，而编码器-解码器架构通常用于序列到序列的任务（如机器翻译)。因此，单独使用解码器能够更专注于生成任务，提高效率。

8. 在大语言模型的训练过程中，如何通过“梯度累积”（Gradient Accumulation）来处理内存限制?
   答案：梯度累积通过在多个小批次上累积梯度，然后再进行一次更新，减少了每次更新所需的内存。这样可以在不增加内存占用的情况下，使用更大的批次进行训练，提升训练稳定性和效率。

9. 大语言模型如何通过“混合精度训练”（Mixed Precision Training）来提高训练效率?
   答案：混合精度训练通过在训练过程中使用较低精度的数据类型（如 FP16），减少内存占用和计

算量，同时保持数值稳定性。现代硬件（如 NVIDIA 的 Tensor Cores）对低精度计算有优化，能够进一步提升训练速度。

10. 监督微调阶段如何帮助 ChatGPT 更好地理解和生成符合人类意图的文本?
答案：监督微调阶段通过使用高质量的人工标注数据，模型学习了如何根据人类的指令生成更符合预期的文本，从而提高了对人类意图的理解和响应能力。

11. ChatGPT 训练的奖励建模阶段的主要目标是什么?
答案：奖励建模阶段的主要目标是训练模型预测其生成的文本可能获得的人类评分，以便在后续的强化学习阶段，模型能够生成更符合人类偏好的内容。

12. 在强化学习阶段，ChatGPT 是如何根据人类反馈调整其生成策略的?
答案：在强化学习阶段，模型根据奖励建模阶段获得的反馈，调整其生成策略，以最大化预期的奖励，从而生成更符合人类偏好的文本。

13. 监督微调阶段与预训练阶段的主要区别是什么?
答案：预训练阶段主要使用大量的未标注数据进行自监督学习，关注语言的基本结构和模式；而监督微调阶段使用高质量的人工标注数据，关注特定任务或指令的响应能力。

14. ChatGPT 的训练过程如何确保模型生成的内容符合伦理和安全标准?
答案：ChatGPT 的训练过程通过精心设计的数据集和训练策略，结合人工审核和反馈机制，旨在减少模型生成有害或不安全内容的风险，确保生成的文本符合伦理和安全标准。

15. 大语言模型如何在多模态任务中整合视觉和语言信息?
答案：通过引入视觉编码器，将图像信息转换为特征向量，与语言模型的文本嵌入进行融合，利用多模态预训练策略，使模型能够同时理解和生成视觉和语言信息。

16. 大语言模型如何实现对特定领域知识的精确掌握和应用?
答案：通过在特定领域的大规模高质量数据上进行微调，使模型能够学习该领域的专业知识和术语，从而在相关任务中表现出色。

17. 在大语言模型的训练中，如何确保模型对时间敏感信息的更新和适应?
答案：定期在最新的数据上进行增量训练或微调，使模型能够及时更新其知识库，保持对时间敏感信息的准确性和时效性。