1. 机器学习的种类
    1. 三种划分方式：
    2. 按有无监督划分，可以细分成：
        1. 为下面的算法按有无监督分类：KNN、神经网络、朴素贝叶斯、Bagging\Boosting、PCA、层次聚类、Kmeans、DBSCAN、关联检测
        2. 监督学习和强化学习，计算机获得经验的来源分别是什么？
    3. 增量学习又称（），整个过程是（在线/离线）完成的；离线学习又称（）
    4. KNN属于基于（模型/实例）的学习
2. 机器学习的基本步骤
    1. 五个步骤：
    2. 数据预处理阶段，特征缩放的两种方法及其公式
        1. 特征缩放的目的
    3. 模型评估阶段，有哪些评价指标，各自的公式
        1. 在样本（）情况下，准确率会失效
        2. 精确率和召回率可以同时提高吗？
        3. 交叉验证有三种方法，分别简要介绍内容。
        4. k折交叉验证可以防止过拟合吗？
3. 参数调整
    1. 参数和超参数的区别
    2. 超参数会影响到模型的哪些方面？
    3. 调参的两种方法；如果搜索空间维度为5，使用哪种调参方法？

---

4. KNN中的K是什么含义？
5. K值的选择：K值越小，单个样本的影响越（大/小）；近似误差越（大/小）；估计误差越（大/小）；更加（易/不易）受噪声影响；通常选用（）方法来选取最优的k值。
6. 距离度量有哪些，公式是什么？
7. 给定一个二维空间数据集

   T={(2,3),(5,4),(9,6),(4,7),(8,1),(7,2)}

   构造一个平衡kd树。

   在这棵树上搜索点(3,4.5)的k近邻，k=2。
8. 普通寻找K近邻时的时间复杂度：；KD树的时间复杂度：

---

9. 贝叶斯公式
    1. 基础的朴素贝叶斯公式
    2. 先验概率
    3. 后验概率
    4. 边际似然度

5. 似然度

10. 朴素是指什么?

    1. 独立性假设公式

11. 给出朴素贝叶斯进行预测的最终简化公式。

12. 使用朴素贝叶斯算法确定对于样本x(2,S),y的取值。（下面的数据和PPT上例题不一样）

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
x（1）	1	1	1	1	1	2	2	2	2	2	3	3	3	3	3
x（2）	S	M	M	S	S	S	M	M	L	L	L	M	M	L	L
y	-1	-1	1	1	-1	-1	1	1	1	1	1	1	1	1	-1

13. 朴素贝叶斯的估计方法是（极大似然估计/最小二乘法）

14. 为了防止概率值为0，采用（）方法

---

15. 什么是损失函数、代价函数、目标函数

    1. 常用的损失函数

    2. 常用的代价函数

16. 线性回归的模型

17. 线性回归的代价函数，化简后的代价函数

18. 线性回归的算法步骤

19. 线性回归的解析解（闭式解）为?

20. L1/L2正则化，各自的正则化项分别是什么?

21. 如果我需要进行特征选择，那么使用（L1/L2）正则化；数据集比较极端，使用（L1/L2）正则化；（L1/L2）正则化得到的解更稀疏。

---

22. 感知机模型

23. 函数距离、几何距离分别如何定义。

24. 误分类点到超平面的几何距离公式。

25. 感知机的损失函数?

26. 感知机的代价函数?（区分损失函数）

27. 感知机模型学习过程

28. 训练数据集:

正实例点：$x_1 = (-3, 3)^T, x_2 = (-5, 2)^T$

负实例点：$x_3 = (2, 4)^T, x_4 = (3, 2)^T$

学习率设定为0.1

构建感知机模型

29. 数据同26，使用对偶算法，求感知机模型

---

30. 逻辑回归模型

31. 逻辑回归模型通过（极大似然估计/最小二乘法）估计模型参数

32. 信息熵公式
33. 当事件发生的概率为（0/0.5/1）时，信息的含量最大
34. 交叉熵公式
35. 逻辑回归模型的损失函数为
36. 极大似然估计模型参数的全过程（找到似然函数、似然函数最大化）
37. 下面的算法中，可以直接处理多个类别分类的算法有（随机森林、朴素贝叶斯、逻辑回归、感知机、支持向量机）
38. 拆分策略有哪些?

    在这些策略中

    1. 分别形成了几个二分类任务
    2. 几个二分类分类器
    3. 选择最终类别的标准是什么
    4. 存储开销、测试时间、训练时间与另外一个策略相比，更（大/小）

---

39. 已知随机变量X的条件下，随机变量Y的条件熵为（），它表示随机变量Y的（不确定性/确定性）
40. 在下面的数据集中，哪一个特征更有价值?

特征1	特征2	分类
A	A	0
B	A	0
A	A	0
B	B	1
A	B	1
B	B	1

41. 以D表示整个数据集，A表示特征名，给出在特征A下的数据集的信息增益、信息增益比、基尼系数的公式
42. 有以下数据集

    （https://blog.csdn.net/wsp_1138886114/article/details/80955528）

ID	性别	车型	衬衣尺码	类
1	男	家用	小	C0
2	男	运动	中	C0
3	男	运动	中	C0
4	男	运动	大	C0
5	男	运动	加大	C0

6	男	运动	加大	C0
7	女	运动	小	C0
8	女	运动	小	C0
9	女	运动	中	C0
10	女	豪华	大	C0
11	男	家用	大	C1
12	男	家用	加大	C1
13	男	家用	中	C1
14	男	豪华	加大	C1
15	女	豪华	小	C1
16	女	豪华	小	C1
17	女	豪华	中	C1
18	女	豪华	中	C1
19	女	豪华	中	C1
20	女	豪华	大	C1

分别使用ID3\C4.5\CART算法构建决策树。

43. 决策树中，叶子节点、分支、非叶子节点分别代表什么实际含义
44. 三种决策树算法：支持分类/回归，树的结构为（多叉/二叉），特征选择基于什么，是否支持连续值处理，是否支持缺失值处理，是否支持剪枝，特征属性能否多次使用

---

45. 支持向量是什么
46. 间隔最大化指什么
47. 使用的间隔为（函数间隔/几何间隔）
48. 写出找到分离超平面的数学问题，约束条件和目标分别是什么实际含义
49. 给出化简后的分离超平面的数学问题
50. 使用SVM算法，针对给出的训练集，写出数学模型，并给出最大分离超平面和分类决策函数

    正例：$x_1 = (3,3)^T, x_2 = (4,3)^T$

    负例：$x_3 = (1,1)^T$
51. 如果两组数据并不完全线性可分（存在较小的训练误差），则应该使用（硬间隔/软间隔）
52. 核函数的作用
53. 列举一些常用的核函数

---

54. 下列哪个关于集成学习的描述是正确的?

    A. 集成学习一定能取得比最好的个体学习器更好的性能

    B. 集成学习的性能可能与个体学习器的平均性能相同

    C. 集成学习的性能一定不差于最差的个体学习器

    D. 集成学习的性能在个体学习器平均性能与个体学习器最佳性能之间
55. 下列关于Boosting算法的说法中错误的是哪个?

    A. Boosting算法适用于分类、回归、排序等机器学习问题

    B. 后一个基学习器更关注前一个基学习器学错的样本

    C. Boosting算法的输出是所有基学习器的加权求和

D. 不同基学习器使用的样本权重是相同的

56. 下列关于Boosting算法中样本权重调整的说法中错误的是哪个?

A. 所有样本的权重和保持不变

B. 前一个基学习器分错的样本会获得更大的权重

C. 只要权重调整的方向正确，Boosting算法的性能就可以获得理论保证

D. 决策树可以直接处理带权重的样本

57. 下列关于Bagging算法中采样的描述哪个是错误的?

A. 可以使用Bootstrap采样
B. 每个样本在每个基学习器的数据集中只会出现一次
C. 采样是为了获得不同的基学习器
D. 不同基学习器的数据从相同分布中采样得到

58. 下列关于Bagging算法描述中错误的是哪个?

A. Bagging算法中每个基学习器使用相同的数据集
B. 分类任务中使用投票法获得输出
C. 回归任务中使用平均法获得输出
D. Random Forest是具有代表性的Bagging算法

59. AdaBoost算法是一种常用的Boosting算法，该算法的伪代码如图所示

输入: 训练集 $D = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_m, y_m)\}$;
基学习算法 $\mathfrak{L}$;
训练轮数 T.

过程:

1: $\mathcal{D}_1(\boldsymbol{x}) = 1/m$.
2: **for** $t = 1, 2, \ldots, T$ **do**
3: $\quad h_t = \mathfrak{L}(D, \mathcal{D}_t)$;
4: $\quad \epsilon_t = P_{\boldsymbol{x} \sim \mathcal{D}_t}(h_t(\boldsymbol{x}) \neq f(\boldsymbol{x}))$;
5: $\quad$ **if** $\epsilon_t > 0.5$ **then break**
6: $\quad \alpha_t = \frac{1}{2} \ln \left(\frac{1-\epsilon_t}{\epsilon_t} \right)$;
7: $\quad \mathcal{D}_{t+1}(\boldsymbol{x}) = \frac{\mathcal{D}_t(\boldsymbol{x})}{Z_t} \times \begin{cases} \exp(-\alpha_t), & \text{if } h_t(\boldsymbol{x}) = f(\boldsymbol{x}) \\ \exp(\alpha_t), & \text{if } h_t(\boldsymbol{x}) \neq f(\boldsymbol{x}) \end{cases}$
$= \frac{\mathcal{D}_t(\boldsymbol{x}) \exp(-\alpha_t f(\boldsymbol{x}) h_t(\boldsymbol{x}))}{Z_t}$
8: **end for**

输出: $H(\boldsymbol{x}) = \operatorname{sign}\left(\sum_{t=1}^{T} \alpha_t h_t(\boldsymbol{x}) \right)$

图 8.3 AdaBoost算法

考虑由3个样本组成的训练集，在第1轮中基学习算法将样本1与样本2分类正确，样本3分类错误。则在第2轮中，各个样本的权重为:

60. 随机森林是一种典型的Bagging算法。随机森林使用的决策树的每个结点，先从该结点的属性集合中随机选择包含部分属性的属性子集，再从这个子集中选择一个最优的属性用于划分。这样生成的单棵决策树与单棵传统决策树相比，性能往往 (更高/相同/更低)。

61. 下列关于集成学习的说法中错误的是?

A. 个体学习器准确率很高后，要增加多样性可以不牺牲准确性

B. 当基分类器的错误率相互独立时，随着个体数目的增大，集成错误率将指数级下降

C. 现实任务中，个体学习器很难做到相互独立

D. 集成学习的核心是如何产生并结合好而不同的个体学习器

---

62. Kmeans中，如何找到最优的k值
63. Kmeans得到的聚类结果是（平坦/不平坦）的，（层次化/非层次化）的
64. 对有标签、没有标签的数据集，分别用什么来评估聚类的质量？取值范围是多少？越大越好还是越小越好？
65. 如何利用层次聚类进行初始聚类中心的选择？

---

66. PCA模型的性能会随着特征的增加（上升/下降/先升后降/先降后升）
67. 如何选取坐标轴来降维？
68. 如何得到包含最大差异性的主成分方向？
69. 将下面的数据降到1维

$$X = \begin{pmatrix} -1 & -1 & 0 & 2 & 0 \\ -2 & 0 & 0 & 1 & 1 \end{pmatrix}$$