① 英文试题，中文答题，填空题写过程

② 可带计算器

③ 范围：作业＋课件

第一章：基础（强化学习不考）

第二章：公式（条件概率 众数，距离）

第三章：逻辑（通过逻辑连接词计算真假，量词，范围不同导致真值不同）

作用

有监督学习

① KNN

② NB（习题，计算方法）

③ 决策树（记得'-'号，条件熵的权重 取平均离散化，习题x2）

④ PLA（公式，习题）

⑤ LR，NN，SVM（3解）适用范围，特点

无监督学习

① Apriori

② k-means（K值多少，初始中心怎么选）

① 联合概率 $P(A,B) = P(A)P(B|A) = P(B)P(A|B)$

$P(A) = P(A,B) + P(A,\bar{B})$

$P(A) = \sum_{i=1}^{n} P(A,B_i) = \sum_{i=1}^{n} P(A|B_i)P(B_i)$

② 期望：$\begin{cases} E[X] = \sum_i x_i P\{X = x_i\} & \text{(离散型)} \\ E[X] = \int_{-\infty}^{+\infty} x f(x)\,dx & \text{(连续型)} \end{cases}$

③ 性质：$E[X_0 + Y] = E[X] + E[Y]$ （不需要独立）

$E[aX] = aE[X]$

$E[XY] = E[X]\cdot E[Y]$ （独立）

④ 方差：$Var(X) = E[(X - E[X])^2] = E[X^2] - (E[X])^2$

⑤ 协方差：$Cov(X,Y) = E[XY] - E[X]E[Y]$

→ 协方差矩阵（假设只有(X,Y)）

→ $= \begin{bmatrix} Cov(X,X), Cov(X,Y) \\ Cov(X,Y), Cov(Y,Y) \end{bmatrix}$

⑥ 相关系数（Correlation）：$Corr(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$

⑦ 正态分布（Normal random variables）

$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}$

$E[X] = \mu \quad Var(X) = \sigma^2$

⑧ 距离：

Minkowski distance：$d(x,y) = \left(\sum_{k=1}^{n} |x_k - y_k|^r\right)^{1/r}$

- $r=1$ → City block distance（曼哈顿距离）
- $r=2$ → Euclidean distance
- $r=\infty$ → Supremum distance

⑨ Logic：

- Propositional logic（命题逻辑） $\wedge \vee \Rightarrow \Leftrightarrow \neg$（列真值表一定不错）
- First order predicate logic（一阶谓词逻辑） $\forall \exists$, A(…), B(…)

⑩ "⇒": 条件为真，箭头边为真⇒才为真，条件为假，说什么都对。
等等价关系 $(P \Rightarrow Q) \Leftrightarrow (\neg P \vee Q)$

⑪ 一阶谓词逻辑:
object (个体词): a, b, ...
predicate (谓词): A(...) B(...) Z(...) { relations (关系) / functions (性质)
Quantifier (量词) { Universal ~ (全称量词) ∀ / existential ~ (存在量词) ∃
Morgan's Law: { $\forall x L = \neg\exists x \neg L$ / $\neg(\forall x L) = \exists x \neg L$

⑫ Quantifier Scope (量词作用域)
1. 在 $\forall x(F(x) \Rightarrow f(h))$ false → 会飞的都是人，比如鸟会飞，但不是人
2. $\forall x F(x) \Rightarrow f(h)$ true → 是人就会飞，条件错误导致结果说什么都对
{ F: ... can fly
h: ... is human being
1. $\forall x(F(x) \Rightarrow f(h))$ 可以写成 $F(x_1) \Rightarrow f(h) \wedge F(x_2) \Rightarrow f(h) \wedge \cdots F(x_n) \Rightarrow f(h)$ 只要其中一个错误即为错误
2. $\forall x F(x) \Rightarrow f(h)$ 因为当然不是所有 x 都会飞，条件当然错误，即该式为 True，如果为 ⇔，即需要两边等价，两边明显就都错

⑬ KNN, ... 算欧式距离

⑭ NB:
求 max $P(C_i|X)$

$P(C_1|X) = ?$
$P(C_2|X) = ?$
⋮
$P(C_m|X) = ?$
} ⇒ output 中 max

$P(C_i|X) = \frac{P(X, C_i)}{P(X)}$ → 比较大小不用比较方为1)
⇒ $P(X, C_i) = P(X|C_i)P(C_i)$
↓ 条件独立　↓ 统计可得
$= P(X_1|C_i) \cdots P(X_n|C_i)$
↓
考试用 C_i 对应的文本表示，比如出现即为1

⑮ PLA:
$h(X) = \text{sign}(W^T X)$
W 更新 $W_{t+1} \leftarrow W_t + y_{n(t)} X_{n(t)}$

⑯ 决策树
1. ID3: 信息增益 max ← 熵 min
熵: $H(D) = -\sum_{x \in D} P(x) \cdot \log_2 P(x)$
(Entropy)

条件熵(conditional Entropy) $H(Y|D)=\sum_{x\in X} p(x)H(Y|D=x)$

联合熵(joint ~) $H(X,Y)=-\sum_{x\in X}\sum_{y\in Y} P(x,y)\log_2 p(x,y)$

$=H(X)+H(Y|X)$

互信息(Mutual information) $I(X;Y)=H(Y)-H(Y|X)$

等价于 是度量离散变量X和Y的相关度. $=H(X)-H(X|Y)$

$g(D,A)=H(D)-H(D|A)$

2. C4.5 : 信息增益率 max

$GainRatio_A(D)=Gain_A(D)/SplitInfo_A(D)$

其中 $SplitInfo_A(D)=-\sum_{j=1}^{v}\frac{|D_j|}{|D|}\times\log_2\left(\frac{|D_j|}{|D|}\right)$

3. CART (Gini Index) 基尼指数 max

$gini(D)=\sum_{j=1}^{n} p_j(1-p_j)=1-\sum_{j=1}^{n} p_j^2$

$gini_{split}(D)=\sum_{n\in N}\frac{n}{N}gini(D_n)$

4 决策树剪枝: 惩罚

Penalty term: ~~乘法因子~~, 人为规定.

计算错误率: $e_c=\frac{\sum_{i=1}^{L}(r(n_i)+\alpha(n_i))}{\sum_{i=1}^{L} m(n_i)}$

分母: 样本数.

分子: 错误数 + 叶节点数乘惩罚因子

如果 e_c 个则剪枝.

(17) K-means

1. 先给出质心 (prototype):
计算其他点到质心的距离, 划分 cluster
重新计算中心点, 取平均值.
2. 先给出 cluster
计算质心, 然后与 1 相同

→ 直至收敛

(18) DBScan

1. 核心点(core points): 满足 Eps(半径) 和 MinPts(点数)
2. 边界点(border points): 圈中有核心点.
3. 噪声点 (noise points): 除上面之外的点.

算法: 以每个点为圆心画圆. 然后以 1, 2, 3 判断.

(19) 关联规则挖掘 Apriori

1. 支持度(support): $S(C\to A)=S(A\to C)=P(A,C)$ → 联合概率
2. 置信度(confidence): $c(A\to C)=P(A,C)/P(A)=P(C|A)$ → 条件概率

算法:

候选像集 (candidate itemset)

↓ 通过最小支持度

频繁像集 (frequent ~)

最大频繁像集 (maximal ~): 叶子节点, 没有直接 frequent 超集

闭 ~ (closed ~): 其直接超集没有和它支持度相同的.