

## 机器学习考试范围

4-5 个简答/简述题目，计算题 3-4 个，不会太难，总共 9-10 个大题。(最新)

黄色部分**为必考计算题**

红色字体部分是最后一次课的粗略划范围

蓝色字体部分是最新【出完卷子】之后的划范围

不会考 Python 编程题，只考伪代码编程题。

期末占比：50%

题型参考周志华《机器学习》每个章节的题目，但是不会考类似题目。

### 第1章 引言

- 1、什么是监督学习？什么是非监督学习？什么是训练集？什么是测试集？各个指标的含义？什么是回归任务？什么是分类任务？其他机器学习概念。
- 2、监督、无监督、半监督、强化——**计算或者问答（同时可能考概念，和各种方法怎么做，举例）。**

### 第2章 机器学习概念流程

- 1、模型怎么评估？评估指标的概念（TN,TP 等，ROC 曲线）、
- 2、基本流程（重点）部分：**考损失函数、优化方法、训练、测试等**
- 3、模型评估部分：经验误差与过拟合、性能度量（**考公式**）。

### 第3章 线性模型

- 1、逻辑回归/线性回归怎么计算 是什么？是否涉及阈值 简答/计算/应用。线性回归/逻辑回归的区别？评价模型的好坏，反差/偏差是什么？有什么用？
- 2、**回归和逻辑回归的区别，逻辑回归是什么？**
- 3、**逻辑回归/线性回归的计算题**
- 4、**概念&&计算**

### 第4章 基本分类方法：

- 1、**KNN 计算题必考**
- 2、**KNN 算法步骤（伪代码）**  
贝叶斯：可能结合后面考，或者一个问答题，但是不会单独出贝叶斯计算题，贝叶斯不考很深，考开放性题目。举例，写方法。

### 第5章 神经网络

- 1、**激活函数？链式公式是什么？**
- 2、**权重调整表达式。反向传播必考**
- 3、**常见激活函数（考简答或者问答，概念，定义，常见图像，优缺点）**
- 4、**全连接神经网络的推到计算（会给定数据和激活函数）**

### 第6章 支持向量机

- 1、间隔，对偶，核函数，软/硬间隔 支持向量机距离的计算，支持向量 核心思想，哪些是支持向量，怎么求？考 SVR 的时候结合 SVM 考，不会单独考 SVR。**拉格朗日和对偶问题必考。**
- 2、核函数的**概念**
- 3、**对偶问题的求解**（给定几个点把支持向量机以及目标函数写出，**求出分割平面，支持向量**，给定 7，8 个点或者 3，4 个点）
- 4、**单独考 SVM，或者结合 TSVM，考 SVR。**

### 第7章 集成学习

- 1、Adaboost 算法怎么用，要考计算题，集成学习核心思想
- 2、boosting and adboost 计算题：弱分类器迭代达到强分类器

## 第8章 聚类

- 1、K 均值 必考 考算法/推导/怎么计算 不考概念
- 2、原型、密度、层次聚类 特点、算法流程。
- 3、给定数据按照算法求解（过程求解，10 几分钟的题量）或者问答

### 1、 原型、密度、层次聚类 特点、算法流程。

原型聚类亦称“基于原型的聚类” (prototype-based clustering), 此类算法假设聚类结构能通过一组原型刻画, 在现实聚类任务中极为常用. 通常情形下, 算法先对原型进行初始化, 然后对原型进行迭代更新求解. 采用不同的原型表示、不同的求解方式, 将产生不同的算法. 下面介绍几种著名的原型聚类算法.

“原型”是指样本空间中具有代表性的点.

#### 原型聚类—K 均值聚类

给定样本集  $D = \{x_1, x_2, \dots, x_m\}$ , “k 均值” (k-means) 算法针对聚类所得簇划分  $C = \{C_1, C_2, \dots, C_k\}$  最小化平方误差

$$E = \sum_{i=1}^k \sum_{x \in C_i} \|x - \mu_i\|_2^2, \quad (9.24)$$

#### 算法流程

# k均值算法-伪代码

输入: 样本集  $D = \{x_1, x_2, \dots, x_m\}$ ;  
聚类簇数  $k$ .

过程:

```

1: 从  $D$  中随机选择  $k$  个样本作为初始均值向量  $\{\mu_1, \mu_2, \dots, \mu_k\}$ 
2: repeat
3:   令  $C_i = \emptyset$  ( $1 \leq i \leq k$ )
4:   for  $j = 1, \dots, m$  do
5:     计算样本  $x_j$  与各均值向量  $\mu_i$  ( $1 \leq i \leq k$ ) 的距离:  $d_{ji} = \|x_j - \mu_i\|_2$ ;
6:     根据距离最近的均值向量确定  $x_j$  的簇标记:  $\lambda_j = \arg \min_{i \in \{1, 2, \dots, k\}} d_{ji}$ ;
7:     将样本  $x_j$  划入相应的簇:  $C_{\lambda_j} = C_{\lambda_j} \cup \{x_j\}$ ;
8:   end for
9:   for  $i = 1, \dots, k$  do
10:    计算新均值向量:  $\mu'_i = \frac{1}{|C_i|} \sum_{x \in C_i} x$ ;
11:    if  $\mu'_i \neq \mu_i$  then
12:      将当前均值向量  $\mu_i$  更新为  $\mu'_i$ 
13:    else
14:      保持当前均值向量不变
15:    end if
16:  end for
17: until 当前均值向量均未更新
18: return 簇划分结果

```

输出: 簇划分  $C = \{C_1, C_2, \dots, C_k\}$

划分样本进簇

更新中心点

## 学习向量量化

与  $k$  均值算法类似, “学习向量量化” (Learning Vector Quantization, 简称 LVQ) 也是试图找到一组原型向量来刻画聚类结构, 但与一般聚类算法不同的是, LVQ 假设数据样本带有类别标记, 学习过程利用样本的这些监督信息来辅助聚类.

---

输入: 样本集  $D = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_m, y_m)\}$ ;  
 原型向量个数  $q$ , 各原型向量预设的类别标记  $\{t_1, t_2, \dots, t_q\}$ ;  
 学习率  $\eta \in (0, 1)$ .

过程:

- 1: 初始化一组原型向量  $\{\mathbf{p}_1, \mathbf{p}_2, \dots, \mathbf{p}_q\}$
- 2: **repeat**
- 3:   从样本集  $D$  随机选取样本  $(\mathbf{x}_j, y_j)$ ;
- 4:   计算样本  $\mathbf{x}_j$  与  $\mathbf{p}_i$  ( $1 \leq i \leq q$ ) 的距离:  $d_{ji} = \|\mathbf{x}_j - \mathbf{p}_i\|_2$ ;
- 5:   找出与  $\mathbf{x}_j$  距离最近的原型向量  $\mathbf{p}_{i^*}$ ,  $i^* = \arg \min_{i \in \{1, 2, \dots, q\}} d_{ji}$ ;
- 6:   **if**  $y_j = t_{i^*}$  **then**
- 7:      $\mathbf{p}' = \mathbf{p}_{i^*} + \eta \cdot (\mathbf{x}_j - \mathbf{p}_{i^*})$
- 8:   **else**
- 9:      $\mathbf{p}' = \mathbf{p}_{i^*} - \eta \cdot (\mathbf{x}_j - \mathbf{p}_{i^*})$
- 10:   **end if**
- 11:   将原型向量  $\mathbf{p}_{i^*}$  更新为  $\mathbf{p}'$
- 12: **until** 满足停止条件

输出: 原型向量  $\{\mathbf{p}_1, \mathbf{p}_2, \dots, \mathbf{p}_q\}$

---

图 9.4 学习向量量化算法

### 原型聚类-高斯混合聚类

与  $k$  均值、LVQ 用原型向量来刻画聚类结构不同, 高斯混合(Mixture-of-Gaussian)聚类采用概率模型来表达聚类原型.

我们先简单回顾一下(多元)高斯分布的定义. 对  $n$  维样本空间  $\mathcal{X}$  中的随机向量  $\mathbf{x}$ , 若  $\mathbf{x}$  服从高斯分布, 其概率密度函数为

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{x}-\mu)^T \Sigma^{-1}(\mathbf{x}-\mu)}, \quad (9.28)$$

---

输入: 样本集  $D = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_m\}$ ;  
 高斯混合成分个数  $k$ .

过程:

- 1: 初始化高斯混合分布的模型参数  $\{(\alpha_i, \mu_i, \Sigma_i) \mid 1 \leq i \leq k\}$
- 2: repeat
- 3:   for  $j = 1, 2, \dots, m$  do
- 4:     根据式(9.30)计算  $\mathbf{x}_j$  由各混合成分生成的后验概率, 即  
 $\gamma_{ji} = p_{\mathcal{M}}(z_j = i \mid \mathbf{x}_j) \ (1 \leq i \leq k)$
- 5:   end for
- 6:   for  $i = 1, 2, \dots, k$  do
- 7:     计算新均值向量:  $\mu'_i = \frac{\sum_{j=1}^m \gamma_{ji} \mathbf{x}_j}{\sum_{j=1}^m \gamma_{ji}};$
- 8:     计算新协方差矩阵:  $\Sigma'_i = \frac{\sum_{j=1}^m \gamma_{ji} (\mathbf{x}_j - \mu'_i)(\mathbf{x}_j - \mu'_i)^T}{\sum_{j=1}^m \gamma_{ji}};$
- 9:     计算新混合系数:  $\alpha'_i = \frac{\sum_{j=1}^m \gamma_{ji}}{m};$
- 10:   end for
- 11:   将模型参数  $\{(\alpha_i, \mu_i, \Sigma_i) \mid 1 \leq i \leq k\}$  更新为  $\{(\alpha'_i, \mu'_i, \Sigma'_i) \mid 1 \leq i \leq k\}$
- 12: until 满足停止条件
- 13:  $C_i = \emptyset \ (1 \leq i \leq k)$
- 14: for  $j = 1, 2, \dots, m$  do
- 15:   根据式(9.31)确定  $\mathbf{x}_j$  的簇标记  $\lambda_j$ ;
- 16:   将  $\mathbf{x}_j$  划入相应的簇:  $C_{\lambda_j} = C_{\lambda_j} \cup \{\mathbf{x}_j\}$
- 17: end for

输出: 簇划分  $\mathcal{C} = \{C_1, C_2, \dots, C_k\}$

---

图 9.6 高斯混合聚类算法

- 2、 给定数据按照算法求解 (过程求解, 10 几分钟的题量) 或者问答  
 原型聚类例题:  
 k-均值:

# k均值算法-算法实例

- 如下，以西瓜的密度和含糖度数据集为例，来演示k均值算法的学习过程。将编号为  $i$  的样本称为  $x_i$ 。

| 编号 | 密度    | 含糖率   | 编号 | 密度    | 含糖率   | 编号 | 密度    | 含糖率   |
|----|-------|-------|----|-------|-------|----|-------|-------|
| 1  | 0.697 | 0.460 | 11 | 0.245 | 0.057 | 21 | 0.748 | 0.232 |
| 2  | 0.774 | 0.376 | 12 | 0.343 | 0.099 | 22 | 0.714 | 0.346 |
| 3  | 0.634 | 0.264 | 13 | 0.639 | 0.161 | 23 | 0.483 | 0.312 |
| 4  | 0.608 | 0.318 | 14 | 0.657 | 0.198 | 24 | 0.478 | 0.437 |
| 5  | 0.556 | 0.215 | 15 | 0.360 | 0.370 | 25 | 0.525 | 0.369 |
| 6  | 0.403 | 0.237 | 16 | 0.593 | 0.042 | 26 | 0.751 | 0.489 |
| 7  | 0.481 | 0.149 | 17 | 0.719 | 0.103 | 27 | 0.532 | 0.472 |
| 8  | 0.437 | 0.211 | 18 | 0.359 | 0.188 | 28 | 0.473 | 0.376 |
| 9  | 0.666 | 0.091 | 19 | 0.339 | 0.241 | 29 | 0.725 | 0.445 |
| 10 | 0.243 | 0.267 | 20 | 0.282 | 0.257 | 30 | 0.446 | 0.459 |

- 第四步：获得当前簇划分：

$$C_1 = \{x_5, x_6, x_7, x_8, x_9, x_{10}, x_{13}, x_{14}, x_{15}, x_{17}, x_{18}, x_{19}, x_{20}, x_{23}\}$$

$$C_2 = \{x_{11}, x_{12}, x_{16}\}$$

$$C_3 = \{x_1, x_2, x_3, x_4, x_{21}, x_{22}, x_{24}, x_{25}, x_{26}, x_{27}, x_{28}, x_{29}, x_{30}\}$$

- 第五步：计算新的均值向量：

$$\mu'_1 = (0.473; 0.214), \mu'_2 = (0.394; 0.066), \mu'_3 = (0.623; 0.388)$$

- 第六步：重复迭代，直到均向量不发生变化

## k均值算法-算法实例

- 第一步：确定聚类数量：
  - 假定聚类簇数  $k = 3$
- 第二步：初始化中心点：
  - 随机选择3个样本  $x_6$ ,  $x_{12}$ ,  $x_{27}$  作为初始均值向量  
 $\mu_1 = (0.403; 0.237)$ ,  $\mu_2 = (0.343; 0.099)$ ,  $\mu_3 = (0.533; 0.472)$
- 第三步：样本归类：
  - 例如：样本  $x_1 = (0.697; 0.460)$ ，它与当前均值向量  $\mu_1$ ,  $\mu_2$ ,  $\mu_3$  的距离分别为 0.369, 0.506, 0.166，因此  $x_1$  将被划入簇  $C_3$  中。

### 学习向量量化

## 学习向量量化-算法实例

- 如下，以西瓜的密度和含糖度数据集为例，来演示学习向量量化的学习过程。将编号为  $i$  的样本称为  $x_i$ 。

| 编号 | 密度    | 含糖率   | 编号 | 密度    | 含糖率   | 编号 | 密度    | 含糖率   |
|----|-------|-------|----|-------|-------|----|-------|-------|
| 1  | 0.697 | 0.460 | 11 | 0.245 | 0.057 | 21 | 0.748 | 0.232 |
| 2  | 0.774 | 0.376 | 12 | 0.343 | 0.099 | 22 | 0.714 | 0.346 |
| 3  | 0.634 | 0.264 | 13 | 0.639 | 0.161 | 23 | 0.483 | 0.312 |
| 4  | 0.608 | 0.318 | 14 | 0.657 | 0.198 | 24 | 0.478 | 0.437 |
| 5  | 0.556 | 0.215 | 15 | 0.360 | 0.370 | 25 | 0.525 | 0.369 |
| 6  | 0.403 | 0.237 | 16 | 0.593 | 0.042 | 26 | 0.751 | 0.489 |
| 7  | 0.481 | 0.149 | 17 | 0.719 | 0.103 | 27 | 0.532 | 0.472 |
| 8  | 0.437 | 0.211 | 18 | 0.359 | 0.188 | 28 | 0.473 | 0.376 |
| 9  | 0.666 | 0.091 | 19 | 0.339 | 0.241 | 29 | 0.725 | 0.445 |
| 10 | 0.243 | 0.267 | 20 | 0.282 | 0.257 | 30 | 0.446 | 0.459 |



# 学习向量量化-算法实例

- 第一步：确定聚类数量：
  - 假定聚类簇数 $k=5$ 
    - 学习目标是找到5个原型向量 $p_1, p_2, p_3, p_4, p_5$
    - 假设其对应的类标记为 $c_1, c_2, c_3, c_4, c_5$
- 第二步：初始化中心点：
  - 假定初始化样本 $x_5, x_{12}, x_{18}, x_{23}, x_{29}$ 为对应的五个向量原型
- 第三步：样本归类：
  - 例如：样本 $x_1 = (0.697; 0.460)$ ，它与当前原型向量 $p_1, p_2, p_3, p_4, p_5$ 的距离分别为0.283, 0.506, 0.434, 0.260, 0.032，因此 $x_1$ 将被划入簇 $p_5$ 中

---

# 学习向量量化-算法实例

- 第四步：更新原型向量
  - 由于 $p_5$ 与 $x_1$ 距离最近，且两者具有相同的类别标记 $c_1$
  - 假定学习率 $\eta = 0.1$ ，更新 $p_5$ 得到新原型向量：
$$\begin{aligned}\hat{p} &= p_5 + \eta \cdot (x_1 - p_5) \\ &= (0.725; 0.445) + 0.1 \cdot ((0.697; 0.460) - (0.725; 0.445)) \\ &= (0.722; 0.447)\end{aligned}$$
  - 若两者不同， $\hat{p} = p_5 - \eta \cdot (x_1 - p_5)$
- 第五步：
  - 重复迭代，直到达到条件停止循环
    - 超过迭代数量
    - 原型向量的变化值足够小



# 高斯混合聚类-算法实例

- 如下，以西瓜的密度和含糖度数据集为例，来演示高斯混合聚类算法的学习过程。将编号为  $i$  的样本称为  $x_i$ 。

| 编号 | 密度    | 含糖率   | 编号 | 密度    | 含糖率   | 编号 | 密度    | 含糖率   |
|----|-------|-------|----|-------|-------|----|-------|-------|
| 1  | 0.697 | 0.460 | 11 | 0.245 | 0.057 | 21 | 0.748 | 0.232 |
| 2  | 0.774 | 0.376 | 12 | 0.343 | 0.099 | 22 | 0.714 | 0.346 |
| 3  | 0.634 | 0.264 | 13 | 0.639 | 0.161 | 23 | 0.483 | 0.312 |
| 4  | 0.608 | 0.318 | 14 | 0.657 | 0.198 | 24 | 0.478 | 0.437 |
| 5  | 0.556 | 0.215 | 15 | 0.360 | 0.370 | 25 | 0.525 | 0.369 |
| 6  | 0.403 | 0.237 | 16 | 0.593 | 0.042 | 26 | 0.751 | 0.489 |
| 7  | 0.481 | 0.149 | 17 | 0.719 | 0.103 | 27 | 0.532 | 0.472 |
| 8  | 0.437 | 0.211 | 18 | 0.359 | 0.188 | 28 | 0.473 | 0.376 |
| 9  | 0.666 | 0.091 | 19 | 0.339 | 0.241 | 29 | 0.725 | 0.445 |
| 10 | 0.243 | 0.267 | 20 | 0.282 | 0.257 | 30 | 0.446 | 0.459 |

## 高斯混合聚类-算法实例

- 第一步：确定聚类数量：

假定高斯混合成分数  $k=3$

- 第二步：初始化模型参数：

假定初始参数为： $a_1 = a_2 = a_3 = \frac{1}{3}$ ； $\mu_1 = x_6, \mu_2 = x_{22},$

$\mu_3 = x_{27}; \Sigma_1 = \Sigma_2 = \Sigma_3 = \begin{pmatrix} 0.1 & 0.0 \\ 0.0 & 0.1 \end{pmatrix}$

- 第三步：计算后验概率：

例如：样本  $x_1 = (0.697; 0.460)$ ，由后验概率公式可得到对应的三个后验概率分别为： $\gamma_{11} = 0.219, \gamma_{12} = 0.404, \gamma_{13} = 0.377$

# 高斯混合聚类-算法实例

❑ 第四步：更新参数 $\hat{a}_i$ 、 $\hat{\mu}_i$ 、 $\hat{\Sigma}_i$ ：

$$\hat{a}_1 = 0.361, \hat{a}_2 = 0.323, \hat{a}_3 = 0.316$$

$$\hat{\mu}_1 = (0.491; 0.251), \hat{\mu}_2 = (0.571; 0.281), \hat{\mu}_3 = (0.534; 0.295)$$

$$\hat{\Sigma}_1 = \begin{pmatrix} 0.025 & 0.004 \\ 0.004 & 0.016 \end{pmatrix}, \hat{\Sigma}_2 = \begin{pmatrix} 0.023 & 0.004 \\ 0.004 & 0.017 \end{pmatrix}, \hat{\Sigma}_3 = \begin{pmatrix} 0.024 & 0.005 \\ 0.005 & 0.016 \end{pmatrix}$$

❑ 第五步：重复迭代，直到达到最大迭代次数或者参数变化足够小

密度聚类例题：

层次聚类例题：

## 第9章 半监督学习

- 1、TSVM 必考 计算题/问答题/公式推导 考怎么做？计算方式？核心/基本思想 算法流程（重点） 伪代码（重点） track 在哪里？SVM 与 TSVM 的区别在哪里，SVM 相比 TSVM 少了什么步骤，差的点在哪里？半监督学习的流程。

## 第10章 降维和度量

- 1、考动机 距离和度量 什么是度量？考 PCA，但是分值不高。
- 2、计算&&简答

## 第11章 专题部分：

- 1、只考一个问答题或者不考，知识工程？但是由于安全部分没发 PPT，因此不考安全部分，不考其他