1.

1) 贝叶斯最小风险决策和最小错误决策的决策原则

2) 分类判决面公式

解：

1) 最小错误率决策：应当寻找最大的后验概率，来使错误率最小.

$$\boldsymbol{P(error|x) = 1 - P(wi|x) \quad if\ you\ decide\ wi}$$

- Minimum error decision: Maximum a posteriori (MAP)

Decide ω_i if $P(\omega_i|\mathrm{x}) > P(\omega_j|\mathrm{x})$ for all $j \neq i$

最小风险决策:通过计算不同类别的最小风险决策,选择最小的 R(αi|x)所在的类别.

贝叶斯最小风险决策：

Condition risk

$$R(\alpha_i|\mathrm{x}) = \sum_{j=1}^{c} \lambda(\alpha_i|\omega_j) P(\omega_j|\mathrm{x})$$

Minimum risk decision (Bayes decision)

$$\arg\min_i R(\alpha_i \,|\, x)$$

2)

$$\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{k=1}^{n} \mathrm{x}_k$$ 结论记住

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{k=1}^{n} (\mathrm{x}_k - \hat{\boldsymbol{\mu}})(\mathrm{x}_k - \hat{\boldsymbol{\mu}})^t$$

$$\boldsymbol{u1} = (\mathbf{3}, \mathbf{6})^{\boldsymbol{T}}$$

$$\boldsymbol{\Sigma 1} = \begin{pmatrix} \mathbf{1/2} & \mathbf{0} \\ \mathbf{0} & \mathbf{2} \end{pmatrix}$$

$$\boldsymbol{u2} = (\mathbf{3}, \mathbf{-2})^{\boldsymbol{T}}$$

$$\boldsymbol{\Sigma 2} = \begin{pmatrix} \mathbf{2} & \mathbf{0} \\ \mathbf{0} & \mathbf{2} \end{pmatrix}$$

Case 3: Σ_i= arbitrary

$$g_i(\mathrm{x}) = -\frac{1}{2}(\mathrm{x} - \boldsymbol{\mu}_i)^t \boldsymbol{\Sigma}_i^{-1}(\mathrm{x} - \boldsymbol{\mu}_i) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\boldsymbol{\Sigma}_i| + \ln P(\omega_i)$$

$$g_i(\mathrm{x}) = \mathrm{x}^t \mathbf{W}_i \mathrm{x} + \mathrm{w}_i^t \mathrm{x} + w_{i0}$$

$$\mathbf{W}_i = -\frac{1}{2}\boldsymbol{\Sigma}_i^{-1} \qquad \mathrm{w}_i = \boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i$$

$$w_{i0} = -\frac{1}{2}\boldsymbol{\mu}_i^t \boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i - \frac{1}{2}\ln|\boldsymbol{\Sigma}_i| + \ln P(\omega_i)$$

$g_1(x) = x^t W_1 x + w_1^t x + w_{10}$

$W_1 = -\frac{1}{2}\Sigma_1^{-1} = -\frac{1}{2}\begin{pmatrix} 2 & 0 \\ 0 & \frac{1}{2} \end{pmatrix} = \begin{pmatrix} -1 & 0 \\ 0 & -\frac{1}{4} \end{pmatrix}$

$w_1 = \Sigma_1^{-1}u_1 = \begin{pmatrix} 2 & 0 \\ 0 & \frac{1}{2} \end{pmatrix}\begin{pmatrix} 3 \\ 6 \end{pmatrix} = \begin{pmatrix} 6 \\ 3 \end{pmatrix}$

$w_{10} = \frac{1}{2}u_1^t\Sigma_1^{-1}u_1 - \frac{1}{2}\ln|\Sigma_1| + \ln P(w_1)$

$= -\frac{1}{2}(3\ 6)\begin{pmatrix} 2 & 0 \\ 0 & \frac{1}{2} \end{pmatrix}\begin{pmatrix} 3 \\ 6 \end{pmatrix} - \frac{1}{2}\ln 1 + \ln P(w_1)$

$= -\frac{1}{2}(6\ 3)\begin{pmatrix} 3 \\ 6 \end{pmatrix} + \ln P(w_1)$

$= -18 + \ln P(w_1)$

$x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$

$g_1(x) = (x_1\ x_2)\begin{pmatrix} -1 & 0 \\ 0 & -\frac{1}{4} \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \begin{pmatrix} 6 \\ 3 \end{pmatrix}^T\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} - 18 + \ln P(w_1)$

$= -x_1^2 - \frac{1}{4}x_2^2 + 6x_1 + 3x_2 - 18 + \ln P(w_1)$

同理.

$g_2(x) = x^t W_2 x + w_2^t x + w_{20}$.

$W_2 = -\frac{1}{2}\Sigma_2^{-1} = -\frac{1}{2}\begin{pmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{2} \end{pmatrix} = \begin{pmatrix} -\frac{1}{4} & 0 \\ 0 & -\frac{1}{4} \end{pmatrix}$

$w_2 = \Sigma_2^{-1}u_2 = \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{2} \end{pmatrix}\begin{pmatrix} 3 \\ 2 \end{pmatrix} = \begin{pmatrix} \frac{3}{2} \\ 1 \end{pmatrix}$

$P(w_1) = P(w_2)$

$w_{20} = -\frac{1}{2}u_2^t\Sigma_2^{-1}u_2 - \frac{1}{2}\ln|\Sigma_2| + \ln P(w_2)$

$= -\frac{1}{2}(3\ 2)\begin{pmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{2} \end{pmatrix}\begin{pmatrix} 3 \\ 2 \end{pmatrix} - \frac{1}{2}\ln 4 + \ln P(w_2)$

$= -\frac{13}{4} - \ln 2 + \ln P(w_2)$

$g_2(x) = -\frac{1}{4}x_1^2 - \frac{1}{4}x_2^2 + \frac{3}{2}x_1 - x_2 - \frac{13}{4} - \ln 2$

令 $g_1(x) = g_2(x)$, 则

$x_2 = 3.514 - 1.125x_1 + 0.1875x_1^2$

2.

1) 最小错误率决策的判别函数,什么条件下为线性判别函数

判别函数 $g_i(\mathbf{x}) = \ln p(\mathbf{x}|\omega_i) + \ln P(\omega_i)$ 取对数，便于计算简化

$$p(\mathbf{x}\,|\,\omega_i) = \frac{1}{(2\pi)^{d/2}\,|\,\Sigma_i\,|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\mu_i)\Sigma_i^{-1}(\mathbf{x}-\mu_i)\right]$$

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\mu_i)^t\Sigma_i^{-1}(\mathbf{x}-\mu_i) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\Sigma_i| + \ln P(\omega_i)$$

Case 3: Σ_i= arbitrary

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\mu_i)^t\Sigma_i^{-1}(\mathbf{x}-\mu_i) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\Sigma_i| + \ln P(\omega_i)$$

$$g_i(\mathbf{x}) = \mathbf{x}^t\mathbf{W}_i\mathbf{x} + \mathbf{w}_i^t\mathbf{x} + w_{i0}$$

$$\mathbf{W}_i = -\frac{1}{2}\Sigma_i^{-1} \qquad \mathbf{w}_i = \Sigma_i^{-1}\mu_i$$

$$w_{i0} = -\frac{1}{2}\mu_i^t\Sigma_i^{-1}\mu_i - \frac{1}{2}\ln|\Sigma_i| + \ln P(\omega_i)$$

下列情况会是线性判别函数,也就是协方差相等的时候

Case 2: $\Sigma_i = \Sigma$

2)

– 二类决策面 $g_i(\mathbf{x}) = g_j(\mathbf{x})$

$$\Rightarrow \quad \mathbf{w}^t(\mathbf{x}-\mathbf{x}_0) = 0 \qquad \mathbf{w} = \Sigma^{-1}(\mu_i - \mu_j)$$ 协方差任意，不一定垂直

$$\mathbf{x}_0 = \frac{1}{2}(\mu_i + \mu_j) - \frac{\ln\left[P(\omega_i)/P(\omega_j)\right]}{(\mu_i-\mu_j)^t\Sigma^{-1}(\mu_i-\mu_j)}(\mu_i - \mu_j)$$

- 注意跟μ_1-μ_2的关系，决策面不一定与之垂直
- 当$P(\omega_1)= P(\omega_2)$，决策面经过$(\mu_1+\mu_2)/2$

P(wi)与 P(wj)的大小决定了其比值是否大于 1 还是小于 1, ln 之后为正值或者负值,间接决定了决策面向先验概率小的方向移动.

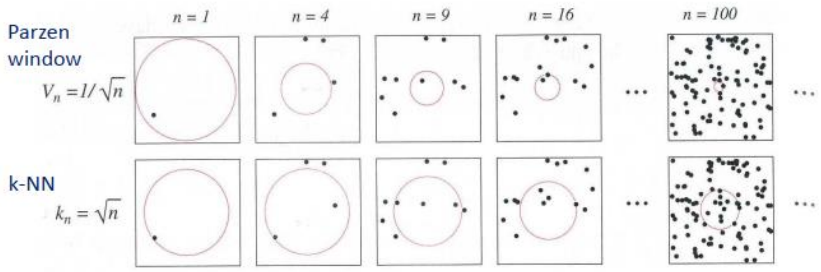
Case 1: $\Sigma_i=\sigma^2 I$ 垂直 Ui-Uj

还有什么情况??????

3.

1) Parzen 窗与 K 紧邻估计的区别

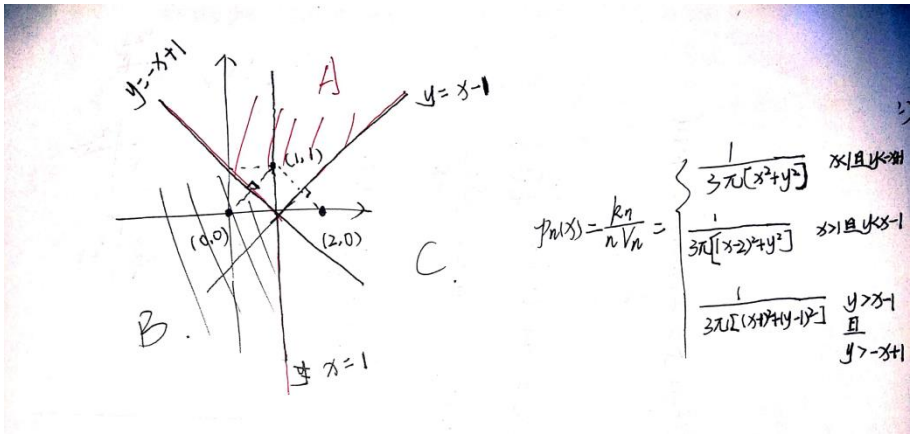
- 非参数概率密度估计
  - Parzen window: 固定局部区域体积V，k变化
  - k-nearest neighbor: 固定局部样本数k，V变化



2) 不知道怎么确定 V

- 一个具体的例子
  - 给定一维空间三个样本点{-2,0,4}，请写出概率密度函数$p(x)$的最近邻（1-NN）估计，并画出概率密度函数曲线图。

$$p_n(x) = \frac{k_n}{nV_n} = \begin{cases} \dfrac{1}{6|x+2|}, & \text{if } x < -1 \\ \dfrac{1}{6|x|}, & \text{if } -1 < x < 2 \\ \dfrac{1}{6|x-4|}, & \text{if } x > 2 \end{cases}$$



3)

$$k = \sum_{i=1}^{c} k_i, \quad p_n(\mathbf{x}, \omega_i) = \frac{k_i}{nV}$$

$$p_n(\omega_i \mid \mathbf{x}) = \frac{p_n(\mathbf{x}, \omega_i)}{\sum_{i=1}^{c} p_n(\mathbf{x}, \omega_i)} = \frac{k_i}{k}, \qquad \omega_m = \arg\max_i \{p_n(\omega_i \mid \mathbf{x})\}$$

这就是K近邻分类器！

4

1)

- **感知准则函数－基本思想**
  - 考虑如下准则函数：

$$J_p(\mathbf{a}) = \sum_{\mathbf{y}\in Y}(-\mathbf{a}^T\mathbf{y}),\ \text{其中，}Y\text{为错分样本集合}$$

  - 当 $\mathbf{y}$ 被错分时， $\mathbf{a}^T\mathbf{y} \le 0$，则 $-\mathbf{a}^T\mathbf{y} \ge 0$。因此 $J_p(\mathbf{a})$ 总是大于等于0。在可分情形下，当且仅当 Y 为空集时 $J_p(\mathbf{a})$ 将等于零，这时将不存在错分样本。
  - 因此，目标是最小化 $J_p(\mathbf{a})$: $\min_{\mathbf{a}} J_p(\mathbf{a})$

可变增量批处理修正方法（另一种表述形式）

Batch Variable-Increment Perceptron

```
1   begin initialize: a, η0, k=0
2     do k ← k+1 (mod n)
3       Yk={}, j = 0
4       do j ← j + 1
5           if yj is misclassified, then append yj to Yk
6       until j = n
7       a = a + ηk Σ_{y∈Y(k)} y      //发现所有错分，然后再修正
8     until Yk={}                    //直到所有样本均正确分类
9     return a
10  end
```

2)

谱聚类

- 从图切割的角度，聚类就是要找到一种合理的分割图的方法，**分割后能形成若干个子图**。连接不同子图的边的权重尽可能小，子图内部边权重尽可能大。
- 谱聚类算法建立在**图论中的谱图理论**基础之上，其本质是将聚类问题转化为一个**图上的关于顶点划分的最优问题**。

2.2 影响因素

- 局部连接 k-近邻范围
- 点对权值计算方法
- 归一化方法
- 聚类数目
- 聚类方法

3)

Adaboost 基本流程：

- (1) 初始化训练数据的权值分布

$$D_1=\{w_{11}, w_{12}, \ldots, w_{1n}\},\ w_{1i}=1/n,\ i=1,\ldots,n$$

- (2) 对 $m=1, 2, \ldots, M$
- (2a) 使用具有权值分布 D_m 的训练数据，学习基本分类器

$$G_m(\mathbf{x}):\ X \rightarrow \{-1,+1\}$$

- (2b) 计算 $G_m(\mathbf{x})$ 在训练数据集上的分类错误率(加权)：

$$e_m = P(G_m(\mathbf{x}_i) \neq y_i) = \sum_{i=1}^{n} w_{mi}\ I(G_m(\mathbf{x}_i) \neq y_i)$$

真值函数

- (2c) 计算 $G_m(\mathbf{x})$ 的贡献系数：

$$\alpha_m = \frac{1}{2}\ln\frac{1-e_m}{e_m}$$

- (2d) 更新训练数据集的权重分布：

$$D_{m+1}=\{w_{m+1,1}, w_{m+1,2}, \ldots, w_{m+1,n}\}$$

- (3) 构建基本分类器的线性组合：

$$f(\mathbf{x}) = \sum_{m=1}^{M} \alpha_m G_m(\mathbf{x})$$

对于两类分类问题，得到最终的分类器：

$$G(\mathbf{x}) = sign(f(\mathbf{x})) = sign\left(\sum_{m=1}^{M} \alpha_m G_m(\mathbf{x})\right)$$

5.

5. $x_1=(0,3,1,2)^T$, $x_2=(1,3,0,1)^T$, $x_3=(3,3,0,0)^T$, $x_4=(1,1,0,2)^T$
$x_5=(3,2,1,2)^T$, $x_6=(4,1,1,1)^T$

①将每个样本看作一类，计算点对之间的距离，见下表：

	x_1	x_2	x_3	x_4	x_5	x_6
x_1	0					
x_2	$\sqrt{5}$	0				
x_3	$\sqrt{14}$	$\sqrt{5}$	0			
x_4	$\sqrt{6}$	$\sqrt{5}$	$\sqrt{12}$	0		
x_5	$\sqrt{10}$	$\sqrt{7}$	$\sqrt{6}$	$\sqrt{6}$	0	
x_6	$\sqrt{21}$	$\sqrt{14}$	$\sqrt{7}$	$\sqrt{11}$	$\sqrt{3}$	0

x_5和x_6

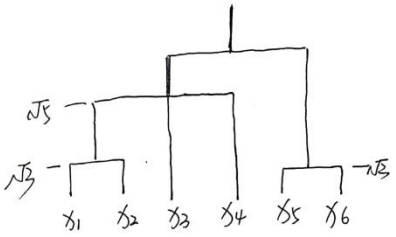
② 查上述矩阵，根据最小距离准则，应把x_5和x_6合并为一类，得$G_1=\{x_1,x_3\}$，
$G_2=\{x_2,x_5,x_6\}$

	G_1	x_3	x_4	G_2
G_1	0			
x_3	$\sqrt{5}$	0		
x_4	$\sqrt{5}$	$\sqrt{12}$	0	
G_2	$\sqrt{7}$	$\sqrt{6}$	$\sqrt{6}$	0

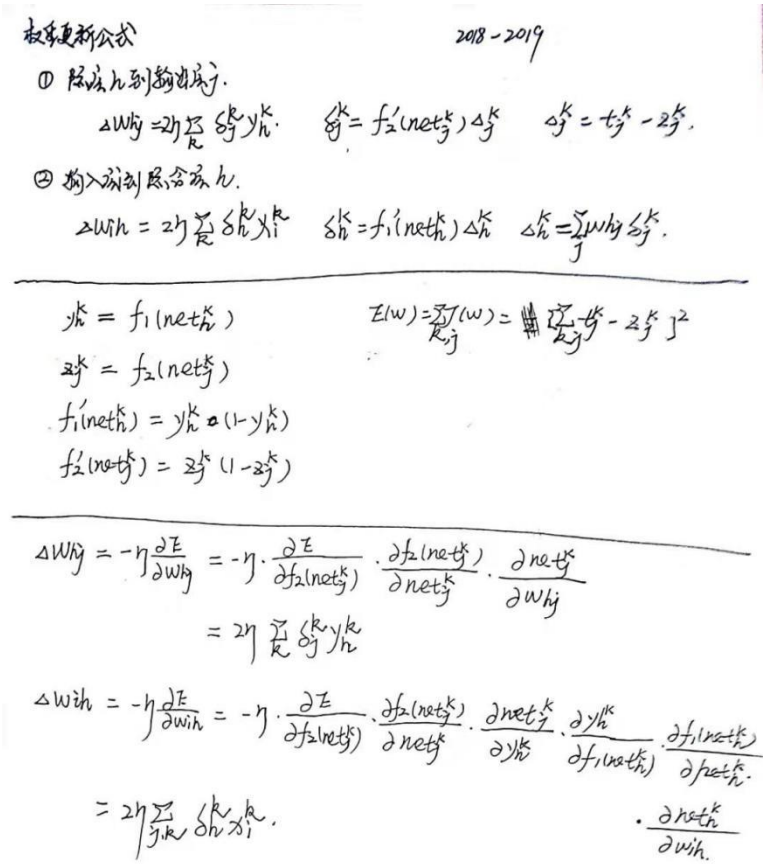
③ $G_3=G_1\cup\{x_3,x_4\}$.

	G_3	G_2
G_3	0	
G_2	$\sqrt{6}$	0

系统树图



6.



7.数据降维

3) 简述并比较 PCA、CCA、LDA、ICA 的区别与适用场景。

4) 详细阐述一种非线性数据降维的方式。

解：

PCA：主成分分析。PCA 的思想是将 n 维特征映射到 m 维上（m<n），这 m 维是全新的正交特征，称为主成分，这 m 维的特征是重新构造出来的，不是简单的从 n 维特征中减去 n-m 维特征。

PCA 属于无监督（训练样本无标签）的降维方法，是一种正交投影，侧重选择样本点投影方差最大的方向，减少特征相关性。

适用场景：

(1) 原始数据特征多而且特征冗余。

(2) 需要对样本进行可视化的时候，三维以上的特征无法进行可视化。

CCA：典型相关分析。它选择的投影标准是降维到 1 维后，两组数据的相关系数最大。

适用场景：侧重于两组数据有相关关系的时候。

LDA：线性判别分析，是从更利于分类的角度的有监督（训练样本有标签）的降维方法。希望数据投影后类内方差最小，类间方差最大。

适用场景：侧重于分类。

ICA：独立成分分析。ICA 信号需要是非高斯的，寻找的是最能使数据的相互独立的方向。

应用场景：盲信号分离。

非线性数据降维：

核 PCA：

1) 先通过核方法把低维线性不可分的数据升维到高维空间，得到线性可分的核矩阵
2) 对核矩阵进行 PCA 降维

流形及其流形学习

流形学习的本质：当样本空间为一个高维光滑流形时，要从样本数据中学习这个高维流形的内在几何结构或内在规律，得到对应的低维数据集，实际也就是非线性降维。

流形学习的主要算法

（1）基于全局的方法，如等距映射（ISOMAP）

（2）基于局部的方法，如局部线性嵌入算法（LLE）

LLE 算法主要分为三步：

（1）求 k 个近邻的过程，这个过程使用了和 KKN 算法一样的求最近邻的方法

（2）对每个样本求它在邻域里的 k 个近邻的线性关系，得到线性关系的权重系数 w

（3）利用权重系数在低维里重构样本数据

ISOMAP：引入图论框架，将数据作为图中的点，点与其邻近点之间使用边来连接，逼近的测地线使用最短路径代替。

步骤 1：构建邻接图 G（

步骤 2：计算所有点对之间的最短路径

步骤 3：把最短路径输入 MDS 算法得到输出。

8.决策树

1) 描述并比较 ID3、C4.5、CART 三种决策树方法的区别。

2) 阐述随机森林的核心思想。

解：

ID3：核心是在决策树各个结点上应用**信息增益准则**选择特征，递归的构建决策树。相当于用极大似然法进行概率模型的选择。

（1）不能对连续数据进行处理，只能通过连续数据离散化进行处理；

（2）采用信息增益进行数据分裂容易偏向取值较多的特征，准确性不如信息增益率；

（3）缺失值不好处理。

（4）没有采用剪枝，决策树的结构可能过于复杂，出现过拟合。

C4.5：继承了 ID3 的优点，并从以下四个方面进行改进。

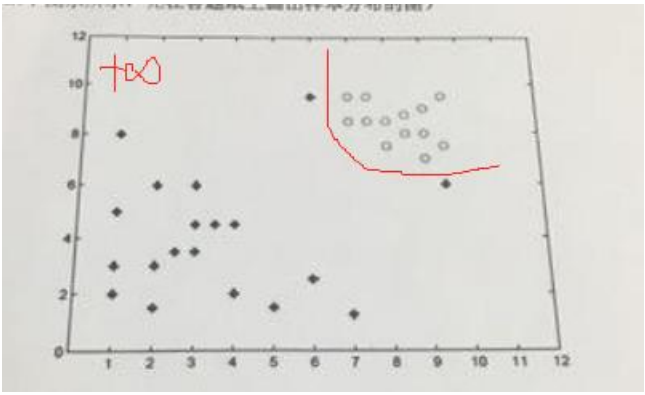
1) 用**信息增益率**来选择属性，克服了用信息增益选择属性时偏向选择取值多的属性的不足
2) 在树构造过程中进行剪枝
3) 能够完成对连续属性的离散化处理
4) 能够对不完整数据进行处理

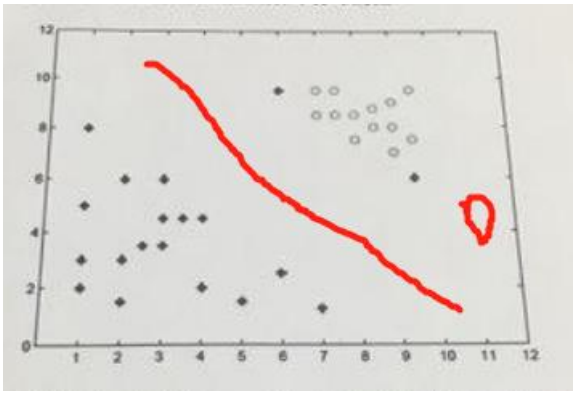
CART：相比 ID3 和 C4.5，CART 既可以用于分类也可以用于回归。 CART 树的生成就是递归地构建二叉决策树的过程。对回归树用平方误差最小化准则，对分类树用**基尼指数**最小化准则，进行特征选择，生成二叉树。

随机森林：该算法用随机的方式建立起一棵棵决策树，然后由这些决策树组成一个森林，其中每棵决策树之间没有关联，当有一个新的样本输入时，就让每棵树独立的做出判断，按照多数原则决定该样本的分类结果。

9.SVM

1) C 取无穷大和无穷小时分类边界



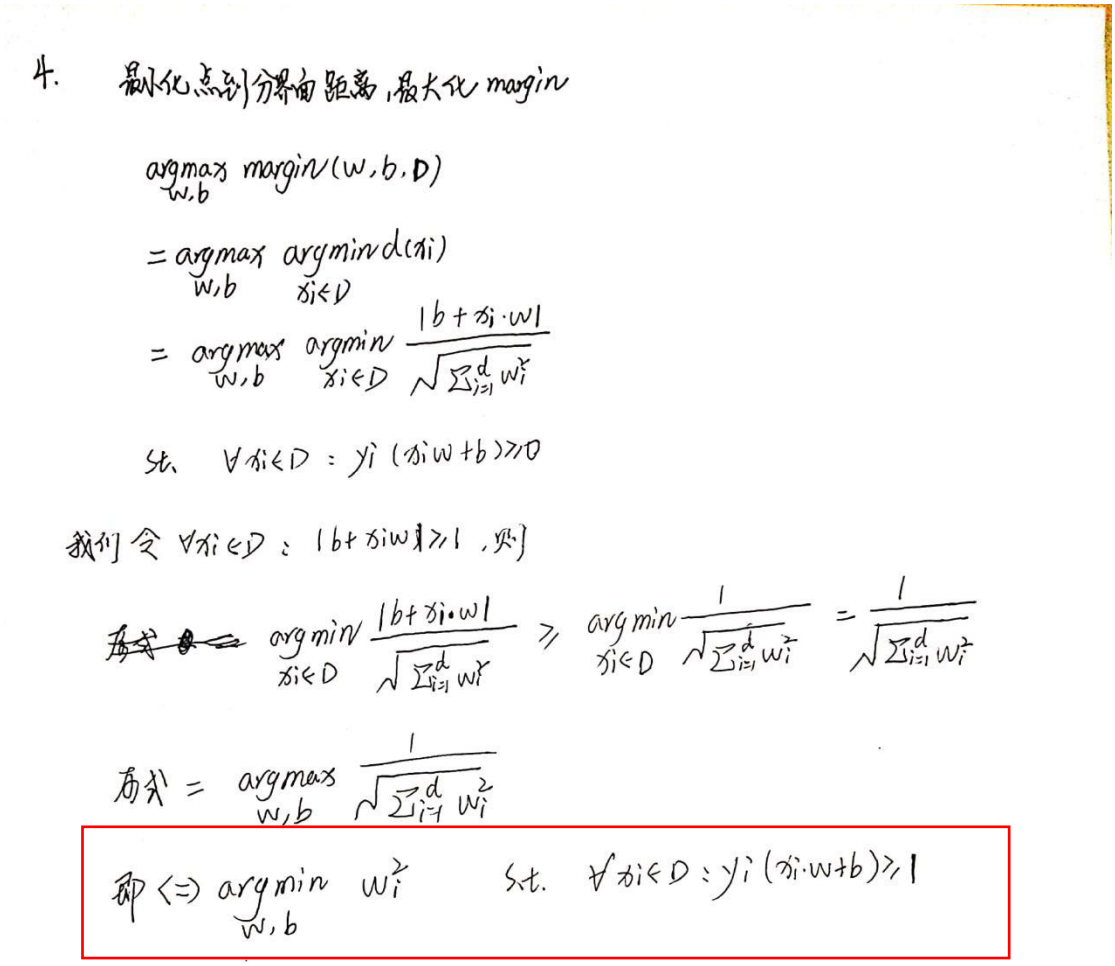


C 在等于合适的值的时候，不是特别大也不是特别小的时候。既保证了分类，而且又最大化 magin，避免过拟合现象。

惩罚因子 C 越大，则 SVM 会更倾向把所有数据分对，往往出现较小的 margin，最终导致过拟合现象，泛化性能不好。

C 过于小，则惩罚力度不够，SVM 会更倾向实现最大化的 margin，而对样本分对分错不关心，不利于分类。

2) 二分类线性可分的数据，SVM 的优化目标最小化是什么？如何从最大化 margin 的角度推过来。



3) 阐述核方法的基本思想是如何将线性模型转换非线性模型的

将在原始低维空间线性不可分的分类问题通过非线性变换成高维空间线性可分的分类问题，在高维空间学习线性支持向量机。在线性支持向量机学习的对偶问题中，把低维到高维的非线性变换的内积形式用核函数表示。

补充

1. LDA 的缺点与改进:

缺点：

最多可以降到 k-1 维;

可能会过拟合数据。

类分离问题。

(1) LDA 至多可生成 C-1 维子空间

LDA 降维后的维度区间在[1,C-1]，与原始特征数 n 无关，对于二值分类，最多投影到 1 维。

1) LDA 不适合对非高斯分布样本进行降维

2) LDA 在样本分类信息依赖方差而不是均值时，效果不好。

改进：

改进目标函数，对容易混淆的类别加入较大的惩罚因子，最好主要与 KNN 结合。