机器学习

第 4 章监督学习-分类算法

欧阳毅

浙江工商大学
管理工程与电子商务学院

2023 年 3 月 8 日

目录

感知机模型

- 定义 3.1(感知机) 假设输入空间 (特征空间) 是 $x \in X \subseteq R^n$, 输出空间是 $y \in Y \subseteq \{+1. -1\}$ 表示实例的类别.

$$f(x) = sign(w \cdot x + b)$$

称为感知机。其中 w 叫作权值,b 叫作偏置 (bias), $w \cdot x =< w, x >$ 表示 w 和 x 的内积,sign 符号函数, 即

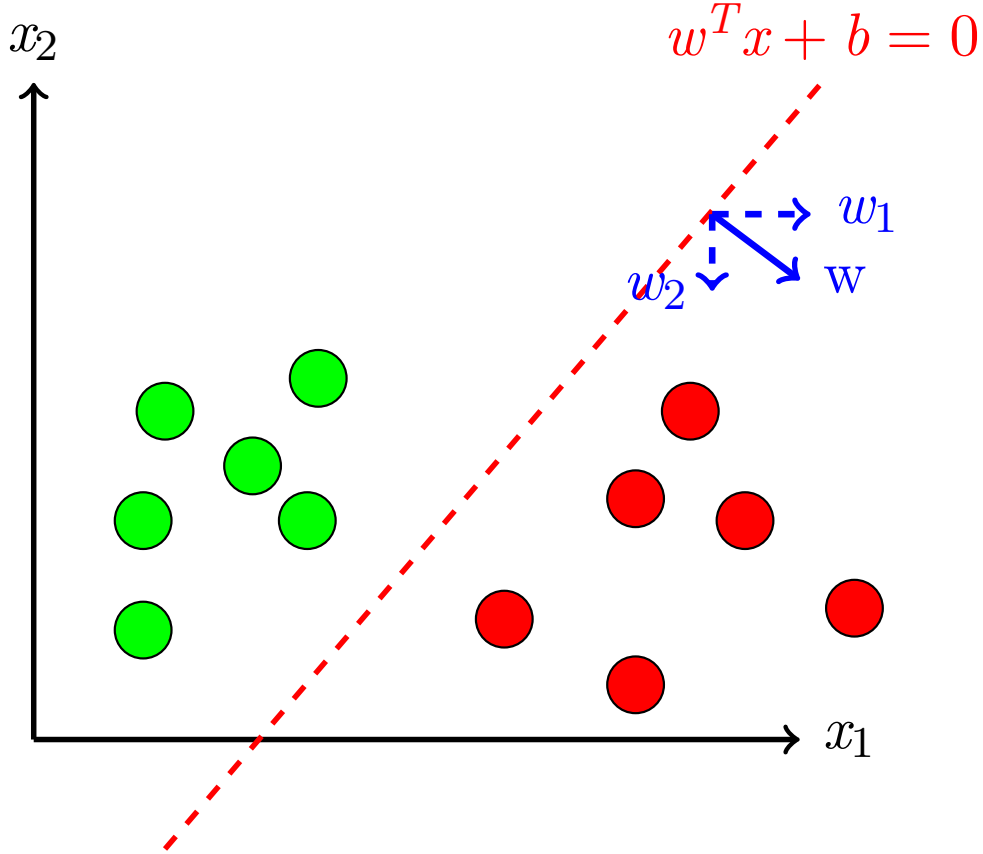
$$sign(x) = \begin{cases} +1 & x \geq 0 \\ -1 & x < 0 \end{cases} \tag{1}$$

感知机模型-几何解释

- 线性方程

$$w \cdot x + b = 0$$

对应于特征空间 R^n 的一个超平面 S，其中 w 是超平面的法向量，b 是超平面的截距



数据集的线性可分性

- 定义 3.2(数据集的线性可分性) 给定一个数据集

$$T = \{(x_1, y_1), (x_2, y_2), ..., (x_N, y_N)\}$$

如果存在某个超平面 S

$$w \cdot x + b = 0$$

能够将数据集的正实例点和负实例点完全正确地划分到超平面的两侧, 则称数据集 T 为线性可分数据集; 否则, 称数据集 T 线性不可分.

即对所有 $y_i = +1$ 的实例, 有 $w \cdot x + b > 0$, 对所有 y=-1 的实例 i, 有 $w \cdot x + b < 0$

感知机学习策略

- 如何选择损失函数?
  - 误分类点的总数（误分类点的总数）
  - 误分类点到超平面 S 的总距离

感知机 $sign(w \cdot x + b)$ 学习的损失函数

给定线性可分训练数据集，感知机 $sign(w \cdot x + b)$ 学习的损失函数定义为:

$$L(w, b) = -\sum_{x \in M} y_i(w \cdot x_i + b)$$

其中 M 为误分类点的集合, 它是感知机学习的经验风险函数. 损失函数 L(w,b) 是非负的, 如果没有误分类点, 损失函数值是 0

感知机学习策略 I

首先写出输入空间 R^n 中任一点 x_0 到超平面 S 的距离：

$$\frac{1}{||w||}|w \cdot x + b|$$

这里，$||w||$ 是 w 的 L2 范数.$||x||_2 = \sqrt{\sum_i x_i^2}$

- 对于误分类的数据 (x_i, y_i) 来说，

$$\begin{cases} (w \cdot x_i + b) > 0, if & y_i = -1 \\ (w \cdot x_i + b) < 0, if & y_i = +1 \end{cases}$$

因此，$-y_i(w \cdot x_i + b) > 0$ 成立. 误分类点 x_i 到超平面 S 的距离是

$$-\frac{1}{||w||} y_i(w \cdot x_i + b)$$

感知机学习策略 II

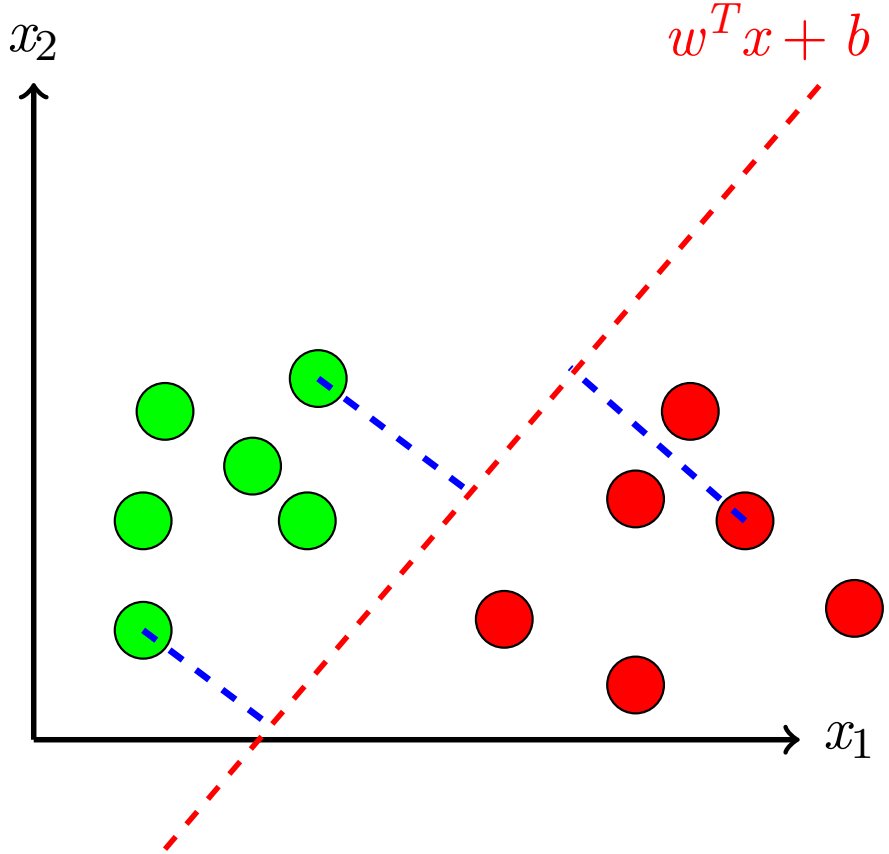
假设超平面 S 的误分类点集合为 M, 那么所有误分类点到超平面 S 总距离为

$$-\frac{1}{||w||}\sum_{x\in M} y_i(w\cdot x_i+b)$$

不考虑 $\frac{1}{||w||}$, 就得到感知机学习的损失函数

感知机模型-几何解释

- 蓝色线段表示数据样本点到超平面的距离



感知机学习算法 I

给定训练数据集 T, 损失函数 L(w,b) 是 w,b 的连续可导函数

- 求参数 w，b，使其为以下损失函数极小化问题的解：

$$\min_{w,b} L(w,b) = -\sum_{x_i \in M} y_i (w \cdot x_i + b)$$

其中 M 为误分类点的集合。

- 首先，任取一个超平面 w0,b0, 然后极小化 L(w,b).
- 假设误分类点集合 M 使固定的，那么损失函数 L 的梯度为：

$$\nabla_w L(w,b) = -\sum_{x \in M} y_i x_i$$

$$\nabla_b L(w,b) = -\sum_{x \in M} y_i$$

感知机学习算法 II

- 随机选取一个误分类点 (x_i, y_i), 对 w,b 进行更新:

$$w \leftarrow w + \eta y_i x_i$$

$$b \leftarrow b + \eta y_i$$

式中 $\eta(0 \leq \eta \leq 1)$ 是步长，也称为学习率。

感知机学习算法的原问题形式：算法 3.1

输入：训练数据集 $T = \{(x_1, y_1), (x_2, y_2), ..., (x_N, y_N)\}$ 其中，$x \in X = R^n$，$y \in Y = \{+1, -1\}$，i=1,2...,N，学习率 η.

输出：w,b; 感知机模型 $f(x) = sign(w \cdot x + b)$

Step1: 选取初值 w_0, b_0

Step2: 在训练集中选取数据 (x_i, y_i)

Step3: 如果 $y_i(w \cdot x_i + b) \leq 0$

$$w \leftarrow w + \eta y_i x_i$$

$$b \leftarrow b + \eta y_i$$

Step4: 转至 Step 2，直至训练集中没有误分类点

解释：当一个实例点被误分类，即位于分类超平面的错误一侧时，则调整 w,b 的值，使分离超平面向该误分类点的一侧移动，以减少该误分类点与超平面间的距离，直至超平面越过该误分类点使其被正确分类。

感知机学习算法

例：如图所示的训练数据集，其正实例点是 $x_1 = (3,3)^T, x_2 = (4,3)^T$. 负实例点是 $x_3 = (1,1)^T$，试用感知机学习算法的原始形式求感知机模型

$$f(x) = sign(w \cdot x + b)$$

. 这里，$w = (w^{(1)}, w^{(2)})^T, x = (x^{(1)}, x^{(2)})^T, \eta = 1$

习题

输入： 训练数据集 $T = \{(x_1, y_1), (x_2, y_2), ..., (x_N, y_N)\}$ 其中,$x \in X = R^n$,$y \in Y = \{+1, -1\}$, i=1,2...,N， 学习率 $\eta = 0.1$.

输出： w,b; 感知机模型 $f(x) = sign(w \cdot x + b)$

- 给出下来训练数据的感知机参数学习过程:

x_1	x_2	y
-3.0	3.0	1
-5.0	2.0	1
2.0	4.0	-1
3.0	2.0	-1

感知机学习算法的收敛性 I

定理（Novikoff）

设训练集 $T = \{(x_1, y_1), (x_2, y_2), ..., (x_N, y_N)\}$ 是线性可分的，其中 $x_i \in X = R^n, y_i \in Y = \{-1, +1\}$，则

(1) 存在满足条件 $||w_{opt}|| = 1$ 的超平面 $w_{opt} \cdot x = w_o \cdot \hat{x} + b_o = 0$ 将训练数据集完全正确分开；且存在 $\gamma > 0$, 对于所有 $i = 1, 2.., N$

$$y_i(w_{opt} \cdot x_i) = y_i(w_o \cdot \hat{x}_i + b_o) \geq \gamma \tag{2}$$

(2) 令 $R = \max_{1 \leq i \leq N} ||x_i||$, 则感知机算法在训练数据集上的误分类次数 k 满足不等式

$$k \leq (\frac{R}{\gamma})^2$$

证明

感知机学习算法的收敛性 II

- (1) 由于训练数据集是线性可分的，根据线性可分数据集的定义，存在超平面可将训练数据完全正确分开，取此超平面为

$$w_{opt} \cdot x = w_o \cdot \hat{x} + b_o$$

，使 $\|w_{opt}\| = 1$。由于对有限的 $i = 1, 2, ..., N$ 均有

$$y_i(w_{opt} \cdot x_i) = y_i(w_o \cdot \hat{x}_i + b_o) > 0$$

所以存在

$$\gamma = \min_i \{y_i(w_{opt} \cdot x_i)\}$$

使 $y_i(w_{opt} \cdot x_i) = y_i(w_o \cdot \hat{x}_i + b_o) \geq \gamma$ (2) 感知机算法从 $w_0 = 0$ 开始，如果实例被误分类，则更新权重，令 w_{k-1} 是第 k 个误分类实例之前的扩充权重向量，即

$$w_{k-1} = (\hat{w}_{k-1}, b_{k-1})^T \tag{3}$$

感知机学习算法的收敛性 III

- 则第 k 个误分类实例的条件是

$$y_i(w_{k-1} \cdot x_i) = y_i(\hat{w}_{k-1} \cdot \hat{x}_i + b_{k-1}) \leq 0$$

若 (x_i, y_i) 是被 w_{k-1} 误分类的数据，则 $\hat{w}$ 和 b 的更新是

$$\hat{w}_k \leftarrow \hat{w_{k-1}} + \eta y_i \hat{x}_i$$

$$b_k \leftarrow b_{k-1} + \eta y_i$$

即

$$w_k \leftarrow w_{k-1} + \eta y_i x_i \tag{4}$$

下面推导两个不等式:

$$w_k \cdot w_{opt} \geq k\eta\gamma \tag{5}$$

感知机学习算法的收敛性 IV

(1) 由公式 (2,4) 得:

$$w_k \cdot w_{opt} = w_{k-1} \cdot w_{opt} + \eta y_i w_{opt} \cdot x_i \geq w_{k-1} \cdot w_{opt} + \eta\gamma$$

由此递推即得不等式 (5)

$$w_k \cdot w_{opt} \geq w_{k-1} \cdot w_{opt} + \eta\gamma \geq w_{k-2} \cdot w_{opt} + 2\eta\gamma \geq \ldots \geq 0 + k\eta\gamma$$

(2)

$$||w_k||^2 \leq k\eta^2 R^2 \tag{6}$$

由公式 (4) 得:

$$||w_k||^2 = ||w_{k-1}||^2 + 2\eta y_i w_{k-1} \cdot x_i + \eta^2 ||x_i||^2$$

由误分类条件，公式 (3) 可得:

$$||w_k||^2 \leq ||w_{k-1}||^2 + \eta^2 ||x_i||^2$$

感知机学习算法的收敛性 V

$$||w_k||^2 \leq ||w_{k-1}||^2 + \eta^2 R^2$$

$$||w_k||^2 \leq ||w_{k-2}||^2 + 2\eta^2 R^2 \leq ... \leq k\eta^2 R^2$$

结合公式 (5,6), 和条件 $||w_{opt}|| = 1$ 即得

$$k\eta\gamma \leq w_k \cdot w_{opt} \leq ||w_k||||w_{opt}|| \leq \sqrt{k}\eta R$$

由于是正数不等式两边同时取平方成立

$$k^2\gamma^2 \leq kR^2$$

$$k \leq (\frac{R}{\gamma})^2$$

- 定理表明，误分类的次数 k 是有上界的，经过有限次搜索可以找到将训练数据完全正确分开的分离超平面。也就是说，当训练数据集线性可分时，感知机学习算法原始形式迭代是收敛的。

目录

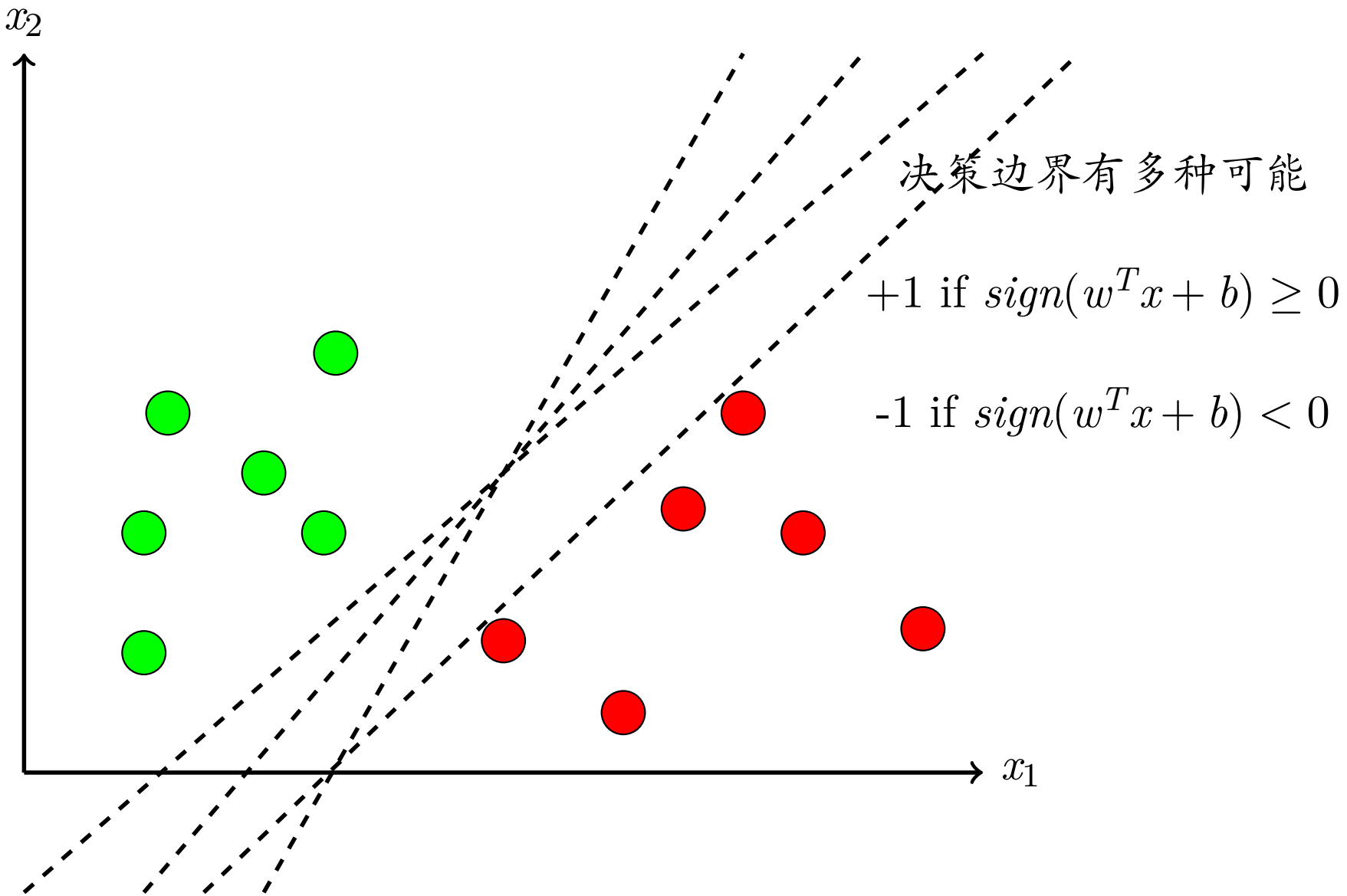
感知机回顾 I

- 属于判别式方法，直接估计决策边界
- 回想一下感知机分类
- 给定一个训练数据集，求参数 w，b，使其为以下损失函数极小化问题的解：

$$\min_{w,b} L(w,b) = -\sum_{x_i \in M} y_i(w \cdot x_i + b)$$

- 感知机是分类误差驱动的

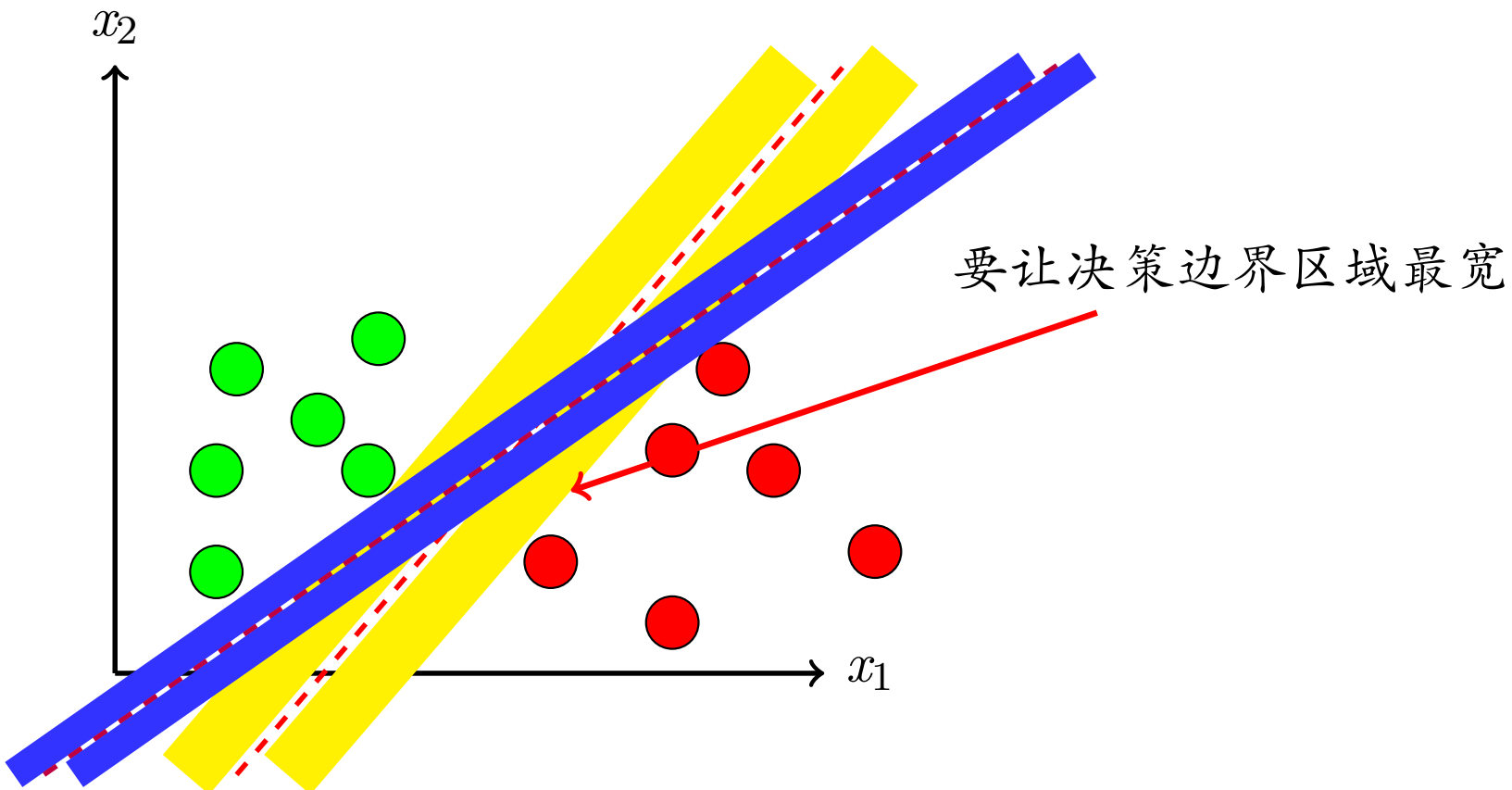
感知机回顾 II



支持向量机（SVM） I

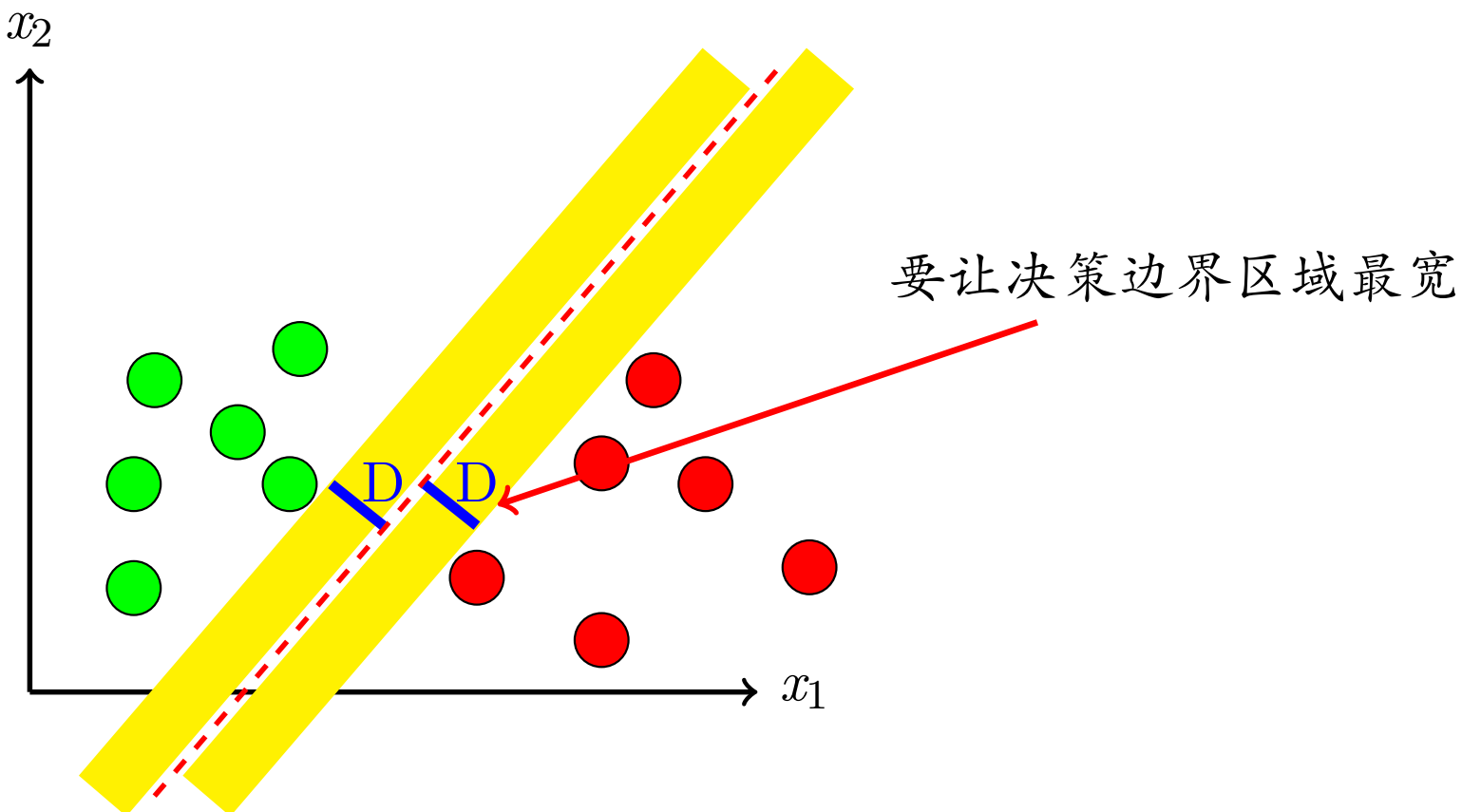
SVM 不是拟合所有点，而是集中在处于边界的点集上

- 属于判别式方法，直接估计决策边界，使得两组点集的分类边界最大化（同决策边界最近点与决策边界的距离最大化）



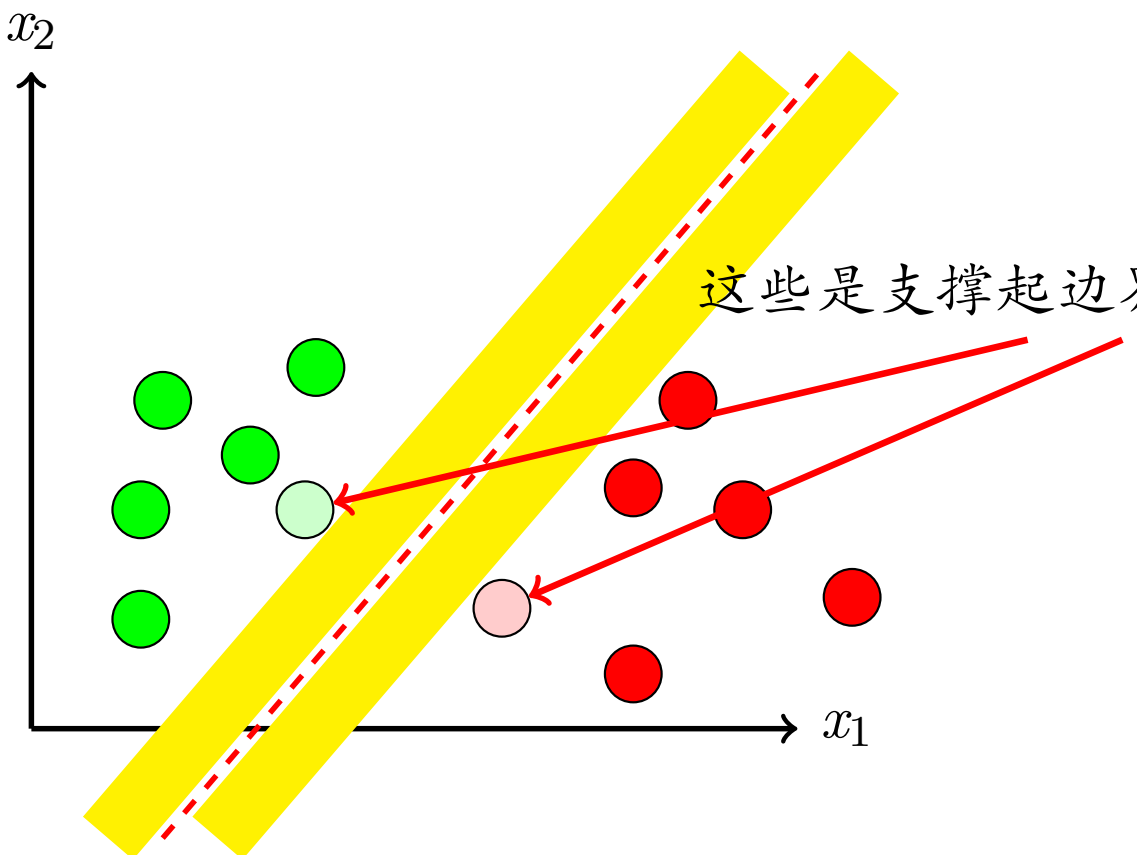
支持向量机（SVM） I

- 属于判别式方法，直接估计决策边界，使得两组点集的分类边界最大化（同决策边界最近点与决策边界的距离最大化）



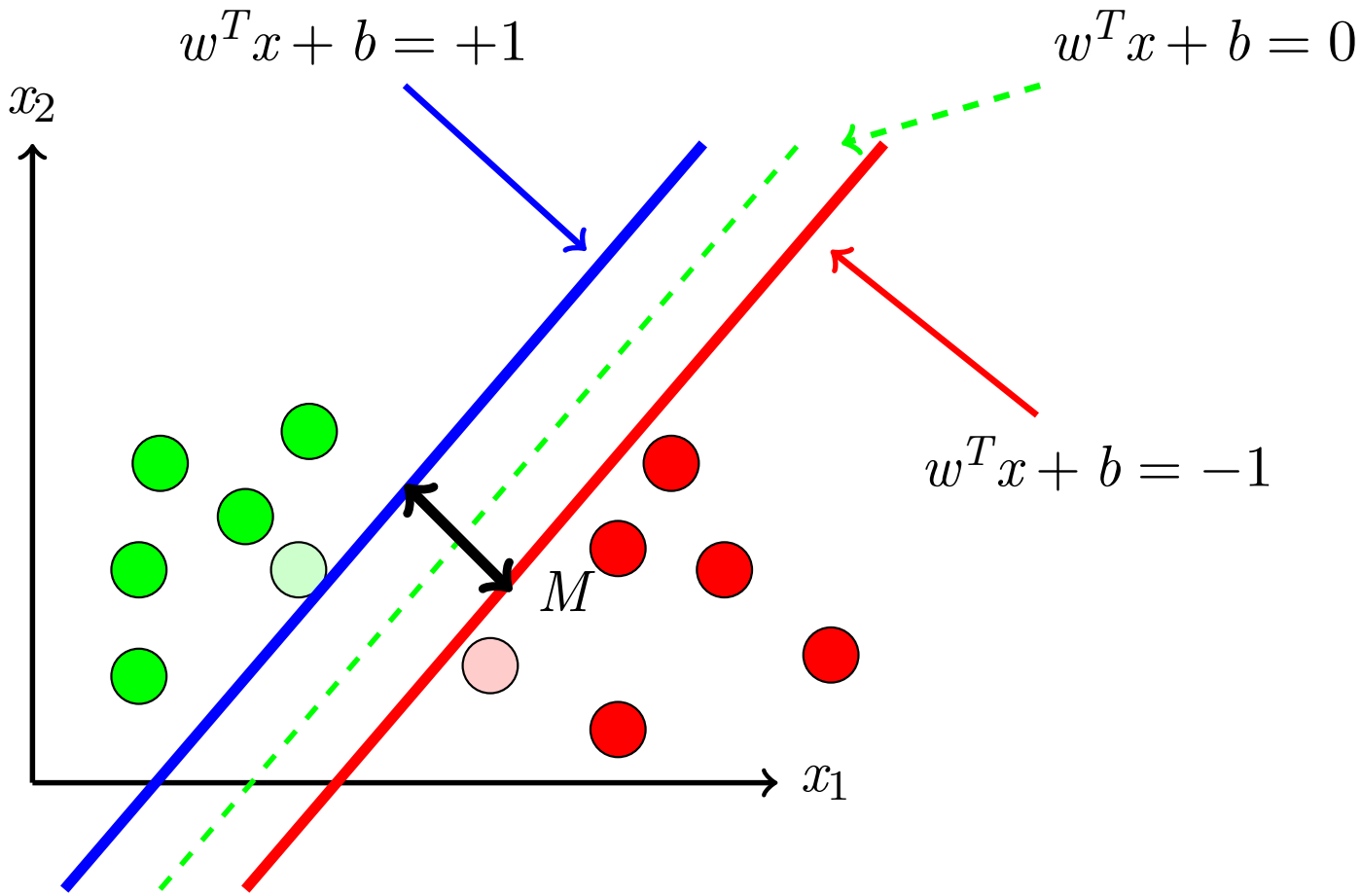
支持向量机（SVM） I

- 属于判别式方法，直接估计决策边界，使得两组点集的分类边界最大化（同决策边界最近点与决策边界的距离最大化）



支持向量机（SVM） I

- 定义边界宽度为 M，如何最大化 M?



支持向量机（Linear-SVM） I

- 向量 w 是同 +1 超平面正交的

证明.

令 u, v 为 +1 超平面的任意两个向量，则有 $w^T(u-v)=0$ □

- 同样，向量 w 是同-1 超平面正交的
- 若 x^+ 是 +1 超平面上的一个点，x^{-1} 是-1 超平面上与 x^+ 最近的一个点，则有

$$x^+ = \lambda w + x^-$$

支持向量机（Linear-SVM） II

- 因为 w 正交于它们所在的平面，过 x^+ 点的垂足在 x^- 上。也就是说向量 $V_{x^+ \to x^-}$ 是和 w 方向是相同的

$$w^T x^+ + b = +1, \quad w^T x^- + b = -1, x^+ = \lambda w + x^-$$

$$w^T(\lambda w + x^-) + b = +1$$

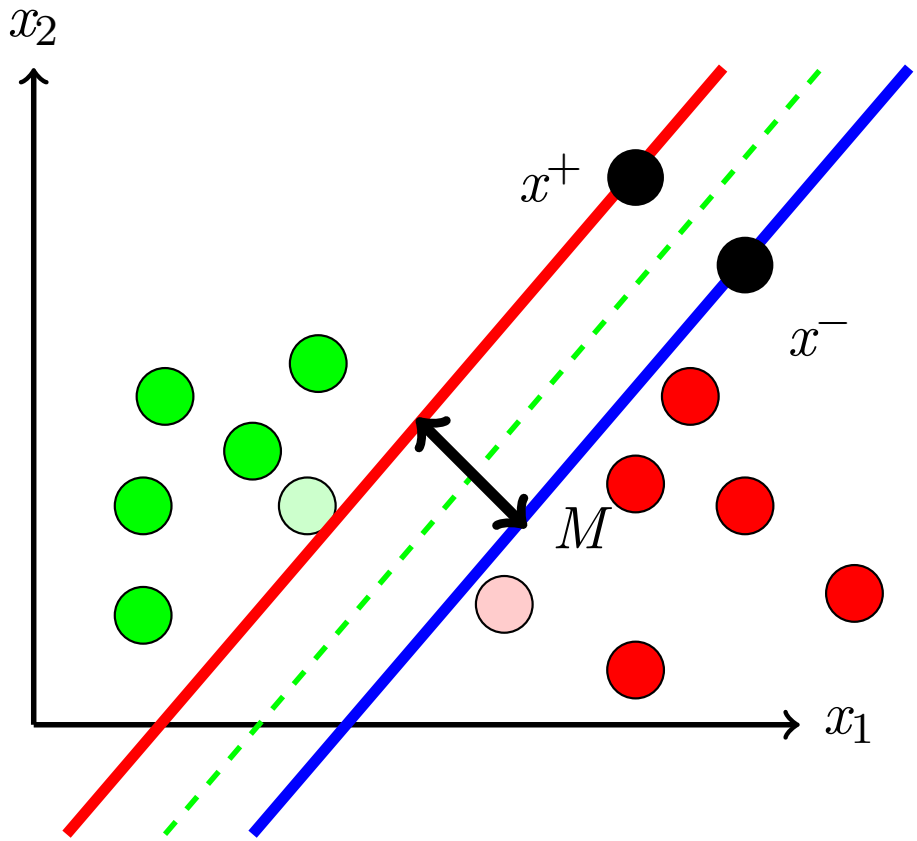
$$\lambda w^T w + w^T x^- + b = +1 \to \lambda w^T w + (-1) = +1$$

$$\lambda w^T w = 2 \to \lambda = \frac{2}{w^T w}$$

$$|x^+ - x^-| = M$$

$$M = |\lambda w| = \lambda \sqrt{w^T w} = \frac{2}{\sqrt{w^T w}}$$

支持向量机（Linear-SVM） III



支持向量机（Linear-SVM） IV

- 最大化 M 等于，最小化分母, 即求解二次规划问题:

$$min_w \frac{w^T w}{2}$$

，服从于如下不等式:

$$w^T x + b \geq 1 \quad if \quad x \in C_+ \tag{7}$$

$$w^T x + b \leq -1 \quad if \quad x \in C_- \tag{8}$$

支持向量机（Linear-SVM） I

$$\min_{w,b} \frac{1}{2}||w||^2$$

- 服从于如下不等式:

$$g_i(w) = -y^{(i)}(w^T x^{(i)} + b) + 1 \leq 0, i = 1.., m$$

- 构造拉格朗日函数

$$L(w, b, a) = \frac{1}{2}||w||^2 - \sum_{i=1}^{m} a_i[y^{(i)}(w^T x^{(i)} + b) - 1]$$

支持向量机（Linear-SVM） II

- 最小化问题，则求其极值点，满足 L 的 w 偏导为 0

$$\nabla_w L(w, b, a) = w - \sum_{i=1}^{m} a_i y^{(i)} x^{(i)} = 0$$

$$w = \sum_{i=1}^{m} a_i y^{(i)} x^{(i)} \quad EQ(1)$$

- 对 L 关于 b 求偏导：

$$\frac{\partial}{\partial b} L(w, b, a) = \sum_{i=1}^{m} a_i y^{(i)} = 0 \quad EQ(2)$$

支持向量机（Linear-SVM） III

- 将 EQ(1) 代入拉格朗日函数

$$L(w,b,a)=\sum_{i=1}^{m}a_i-\frac{1}{2}\sum_{i,j=1}^{m}a_j y^{(j)}a_i y^{(i)}(x^{(i)})^T x^{(j)}-b\sum_{i=1}^{m}a_i y^{(i)}$$

- 利用 EQ(2) 得到对偶形式

$$\max_a L(w,b,a)=\sum_{i=1}^{m}a_i-\frac{1}{2}\sum_{i,j=1}^{m}y^{(i)}y^{(j)}a_i a_j(x^{(i)})^T x^{(j)}$$

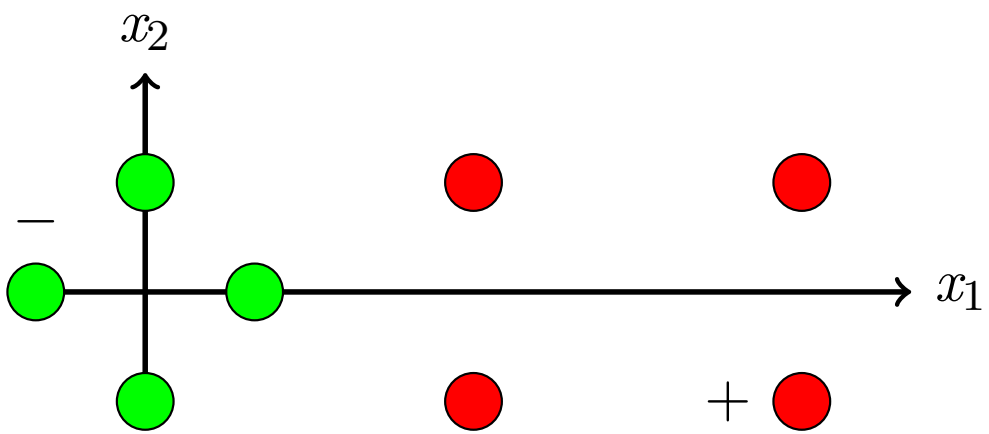
$$s.t.\quad a_i\geq 0,\sum_{i=1}^{m}a_i y^{(i)}=0$$

Linear-SVM 例题 I

例

有 8 个数据样本点，按照 SVM 算法计算超平面方程

+1正实例:[3 1],[3,-1],[6,1],[6,-1]
-1负实例:[1,0],[0,1],[0,-1],[-1,0]



Linear-SVM 习题 I

例

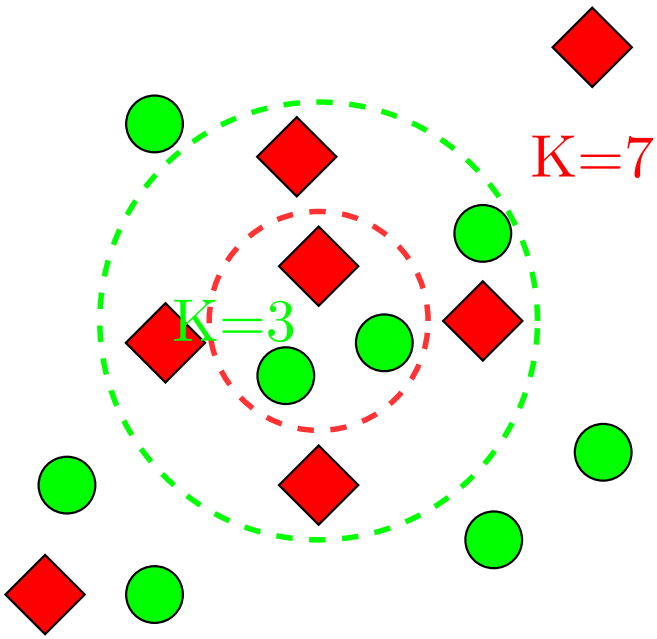
根据数据样本点，按照 SVM 算法计算分类超平面方程

x_1	x_2	y
-3.0	3.0	1
-5.0	2.0	1
2.0	4.0	-1
3.0	2.0	-1

目录

K 近邻法 I

- 算法思路：如果一个样本在特征空间中的 k 个最相似 (即特征空间中最邻近) 的样本中的大多数属于某一个类别，则该样本也属于这个类别



KNN 应用实例

威廉康星乳腺癌诊断数据集

- 这个数据集包含 500 例细胞活检案例，每个案例有 32 个肿块活检图像显示的细胞核的特征。
- 第一个特征是 ID，第二个是这个案例的癌症诊断结果，癌症诊断结果用编码”M” 表示恶性，B 表示良性。其他 30 个特征是数值型的其他指标，包括细胞核的半径 (Radius)、质地 (Texture)、周长 (Perimeter)、面积 (Area) 和光滑度 (Smoothness) 等的‘均值、标准差和最大值。部分数据如下：

A	B	C	D	E	F	G	H	I	J	K	L
ID	Diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave_mean	symmetry_mean	fractal_mean
842302	M	17.99	10.38	122.8	1001	0.1184	0.2776	0.3001	0.1471	0.2419	0.0787
842517	M	20.57	17.77	132.9	1326	0.08474	0.07864	0.0869	0.07017	0.1812	0.0566
84300903	M	19.69	21.25	130	1203	0.1096	0.1599	0.1974	0.1279	0.2069	0.0599
84348301	M	11.42	20.38	77.58	386.1	0.1425	0.2839	0.2414	0.1052	0.2597	0.0974
84358402	M	20.29	14.34	135.1	1297	0.1003	0.1328	0.198	0.1043	0.1809	0.0588
843786	M	12.45	15.7	82.57	477.1	0.1278	0.17	0.1578	0.08089	0.2087	0.0761
844359	M	18.25	19.98	119.6	1040	0.09463	0.109	0.1127	0.074	0.1794	0.0574
84458202	M	13.71	20.83	90.2	577.9	0.1189	0.1645	0.09366	0.05985	0.2196	0.0745
844981	M	13	21.82	87.5	519.8	0.1273	0.1932	0.1859	0.09353	0.235	0.0738
84501001	M	12.46	24.04	83.97	475.9	0.1186	0.2396	0.2273	0.08543	0.203	0.0824
845636	M	16.02	23.24	102.7	797.8	0.08206	0.06669	0.03299	0.03323	0.1528	0.0569
84610002	M	15.78	17.89	103.6	781	0.0971	0.1292	0.09954	0.06606	0.1842	0.0608
846226	M	19.17	24.8	132.4	1123	0.0974	0.2458	0.2065	0.1118	0.2397	0.07
846381	M	15.85	23.95	103.7	782.7	0.08401	0.1002	0.09938	0.05364	0.1847	0.0533
84667401	M	13.73	22.61	93.6	578.3	0.1131	0.2293	0.2128	0.08025	0.2069	0.0768
84799002	M	14.54	27.54	96.73	658.8	0.1139	0.1595	0.1639	0.07364	0.2303	0.0707
848406	M	14.68	20.13	94.74	684.5	0.09867	0.072	0.07395	0.05259	0.1586	0.0592
84862001	M	16.13	20.68	108.1	798.8	0.117	0.2022	0.1722	0.1028	0.2164	0.0735
849014	M	19.81	22.15	130	1260	0.09831	0.1027	0.1479	0.09498	0.1582	0.0539
8510426	B	13.54	14.36	87.46	566.3	0.09779	0.08129	0.06664	0.04781	0.1885	0.0576
8510653	B	13.08	15.71	85.63	520	0.1075	0.127	0.04568	0.0311	0.1967	0.0681
8510824	B	9.504	12.44	60.34	273.9	0.1024	0.06492	0.02956	0.02076	0.1815	0.0690
8511133	M	15.34	14.26	102.5	704.4	0.1073	0.2135	0.2077	0.09756	0.2521	0.0703

KNN 与 Kmeans 的算法的区别

- 1. KNN 算法是分类算法，分类算法肯定是需要有训练数据，然后通过学习训练数据之后的模型来匹配测试数据集，将测试数据集合进行按照预先学习的模型来分类。
- 2. Kmeans 算法是聚类算法，聚类算法与分类算法最大的区别是聚类算法没有训练数据集合。

K 近邻法

输入：训练数据集 $T = \{(x_1, y_1), (x_2, y_2), ..., (x_N, y_N)\}$ 其中，$x \in X = R^n$，$y \in Y = \{c_1, c_2, ..., c_K\}$，i=1,2...,N，实例特征向量 x;.

输出：实例 x 所属于的类 y

- 根据给定的距离度量，在训练集 T 中找出与 x 最近邻的 k 个点，涵盖这 k 个点的领域记为 $N_k(x)$
- 在 $N_k(x)$ 中根据分类决策规则（如多数表决）决定 x 的类别 y：

$$y = \arg\max_{C_j} \sum_{x_i \in N_K(x)} I(y_i = c_i), i = 1, 2, ..., N; j = 1, 2..., K$$

度量空间 I

特征空间中两个实例点的距离是两个实例点相似度的反映，KNN 的也可采用其它距离公式：

$$L_p(x_i, x_j) = (\sum_{l=1}^{n} |x_i^{(l)} - x_j^{(l)}|^p)^{\frac{1}{p}}$$

这里 $p \geq 1$. 当 p=2 时，称为欧氏距离

$$L_2(x_i, x_j) = (\sum_{l=1}^{n} |x_i^{(l)} - x_j^{(l)}|^2)^{\frac{1}{2}}$$

当 p=1 时，称为曼哈顿距离

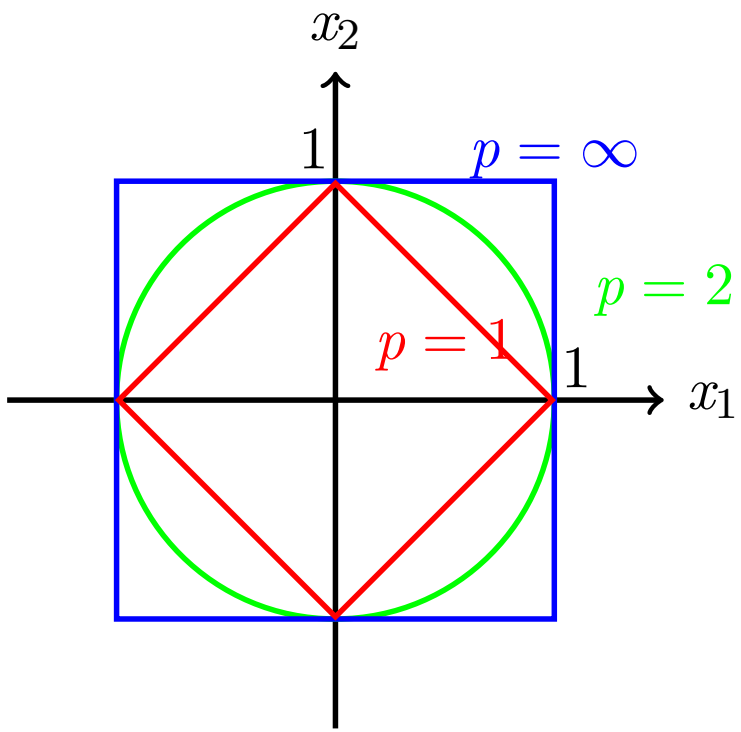
$$L_1(x_i, x_j) = \sum_{l=1}^{n} |x_i^{(l)} - x_j^{(l)}|$$

度量空间 II

当 $p = \infty$. 它是各个坐标距离的最大值（切比雪夫距离）

$$L_\infty(x_i, x_j) = \max_l |x_i^{(l)} - x_j^{(l)}|$$

二维空间 p 取不同值时，与原点的 L_p 距离为 1 的点的图形



KNN-实例分析

例 1: 已知二维空间 3 个点 $x_1 = (1,1)^T, x_2 = (5,1)^T, x_3 = (4,4)^T$，试求在 p 取不同值时，$L_p$ 距离下 x_1 的最近邻?

解: 因为 x_1, x_2 只有第二维上值不同，所以 p 为任何值，$L_p(x_1, x_2) = 4$.

$$L_1(x_1, x_3) = 3 + 3 = 6$$

$$L_2(x_1, x_3) = \sqrt{9+9} = 4.24$$

$$L_3(x_1, x_3) = \sqrt[3]{27+27} = \sqrt[3]{54} = 3.78$$

$$L_4(x_1, x_3) = 3.57$$

于是得到:

p 等于 1 或 2 时，x_2, x_1 是最近邻点，

p 大于 3 时，x_3, x_1 是最近邻点

由此可见近邻与所选取的度量相关

K 值的选择 I

K 值的选择会对 k 近邻法的结果产生重大影响

较小的 K k 值的减小就意味着整体模型变得复杂，容易发生过拟合。

较大的 K k 值的增大就意味着整体的模型变得简单。

K==N k 值一般取一个比较小的数值。通常采用交叉验法来选取 k 值。

KNN 实现数据挖掘 I

习题 1：给出 KNN 学习方法的 python 实现

习题 2：k 值的选择会对 k 近邻法的结果有何影响?

目录

朴素贝叶斯学习与分类 I

- 贝叶斯公式：根据条件概率的定义，在事件 B 发生的条件下，A 发生的概率为：

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

- 重点在于，在碰到问题时，如何能够将实际问题与之联系起来，也就是从实际问题到数学模型的转换过程。

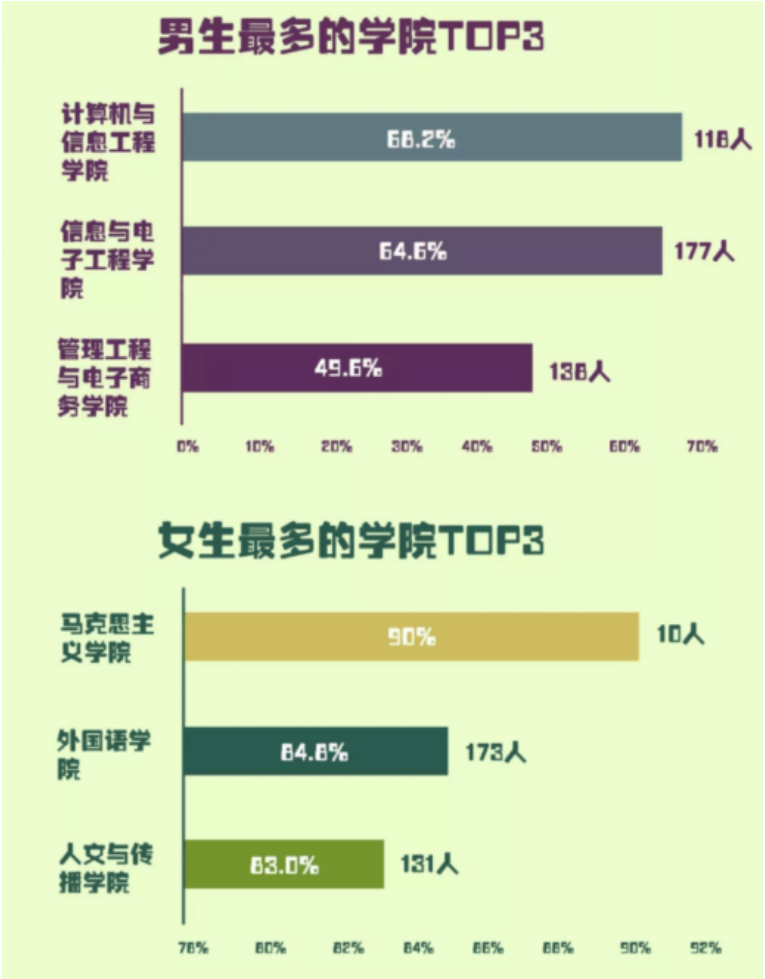
朴素贝叶斯学习与分类-根据身高体重预测性别 I

- 贝叶斯公式: 根据条件概率的定义，在事件 B 发生的条件下，A 发生的概率为:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

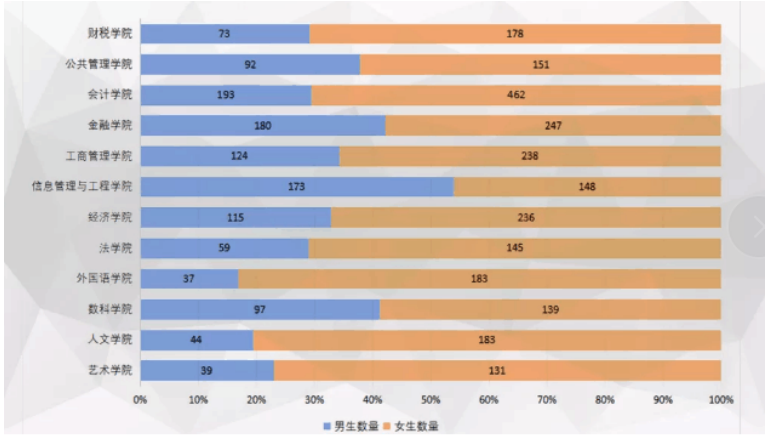
- 先验概率 (或边缘概率)
- 条件概率
- 后验概率

朴素贝叶斯学习与分类-根据身高体重预测性别 II

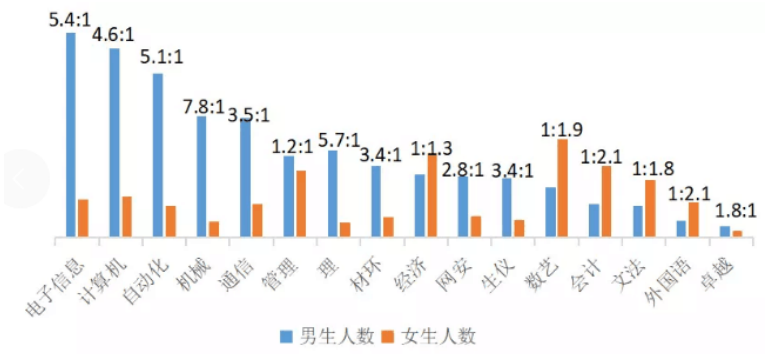


图：浙江工商大学

朴素贝叶斯学习与分类-根据身高体重预测性别 III



图：浙江财经大学



图：杭电

朴素贝叶斯法的学习与分类 I

设输入空间 $x \in X \subseteq R^n$ 为 n 维向量的集合, 类标记集合 $y \in Y = \{c_1, c_2, ..., c_K\}$. P(X,Y) 是 X 和 Y 的联合概率分布. 训练数据集

$$T = \{(x_1, y_1), (x_2, y_2), ..., (x_N, y_N)\}$$

由 P(X,Y) 独立同分布产生.
朴素贝叶斯法通过训练数据集学习联合概率分布 P(X,Y).

- 先验概率分布

$$P(Y = c_k), k = 1, 2..., K$$

- 条件概率分布

$$P(X = x | Y = c_k)$$

但是条件概率分布 $P(X = x | Y = c_k)$ 有指数级数量的参数, 其估计实际是不可行的

朴素贝叶斯法的学习与分类 II

- 假设 $x^{(i)}$ 可取值有 S_j 个，$j = 1, 2, \ldots, n, Y$ 可取值有 K 个，那么参数个数为 $K\prod_{j=1}^{n} S_j$

朴素贝叶斯对条件概率分布做了条件独立的假设:

$$\begin{aligned} P(X = x|\, Y = c_k) &= P(X^{(1)} = x^{(1)}, \ldots, X^{(n)} = x^{(n)}|\, Y = c_k) \\ &= \prod_{j=1}^{n} P(X^{(j)} = x^{(j)}|\, Y = c_k) \end{aligned} \tag{9}$$

- 朴素贝叶斯法分类时, 对给定的输入 x, 通过学习到的模型计算后验概率分布 $P(Y = c_k|X = x)$, 将后验概率最大的类作为 x 的类输出. 后验概率计算根据贝叶斯定理进行:

$$P(Y = c_k|X = x) = \frac{P(X = x|\, Y = c_k)P(Y = c_k)}{\sum_k P(X = x|\, Y = c_k)P(Y = c_k)} \tag{10}$$

朴素贝叶斯法的学习与分类 III

将式 (9) 代入式 (10) 有

$$P(Y=c_k|X=x)=\frac{\prod_{j=1}^{n}P(X^{(j)}=x^{(j)}|Y=c_k)P(Y=c_k)}{\sum_k(\prod_{j=1}^{n}P(X^{(j)}=x^{(j)}|Y=c_k))P(Y=c_k)} \tag{11}$$

朴素贝叶斯分类器可表示为

$$y=f(x)=\arg\max_{c_k}\frac{\prod_{j=1}^{n}P(X^{(j)}=x^{(j)}|Y=c_k)P(Y=c_k)}{\sum_k(\prod_{j=1}^{n}P(X^{(j)}=x^{(j)}|Y=c_k))P(Y=c_k)} \tag{12}$$

注意到, 在式 (12) 中分母对所有 c 都是相同的, 所以,

$$y=f(x)=\arg\max_{c_k}P(Y=c_k)\prod_{j=1}^{n}P(X^{(j)}=x^{(j)}|Y=c_k) \tag{13}$$

期望风险最小化等价于后验概率最大化 I

假设选择 0-1 损失函数:

$$L(Y, f(X)) = \begin{cases} 1 & , \quad Y \neq f(X) \\ 0 & , \quad Y = f(X) \end{cases} \tag{14}$$

式中 f(x) 是分类决策函数. 这时, 期望风险函数为

$$R_{exp}(f) = E[L(Y, f(X))]$$

期望是对联合分布 P(X,Y) 取的. 由此取条件期望

$$R_{exp}(f) = E_X \sum_{k=1}^{K} [L(c_k, f(X))] P(c_k|X)$$

期望风险最小化等价于后验概率最大化 II

为了使期望风险最小化, 只需对 X=x 逐个极小化, 由此得到:

$$\begin{aligned} f(x) &= \arg\min_{y\in Y}\sum_{k=1}^{K} L(c_k, y)P(c_k|X=x) \\ &= \arg\min_{y\in Y}\sum_{k=1}^{K} P(y\neq c_k|X=x) \\ &= \arg\min_{y\in Y}(1-P(y=c_k|X=x)) \\ &= \arg\max_{y\in Y} P(y=c_k|X=x)) \end{aligned} \tag{15}$$

这样一来, 根据期望风险最小化准则就得到了后验概率最大化准则:

$$f(x) = \arg\max_{c_k} P(y=c_k|X=x)$$

即朴素贝叶斯法所采用的原理

朴素贝叶斯法的参数估计

在朴素贝叶斯法中，学习意味着估计 $P(Y=c_k)$ 和 $P(X^{(j)}=x^{(j)}|Y=c_k)$. 可以用极大似然估计法估计相应的概率.

- 先验概率 $P(Y=c_k)$ 的极大似然估计是

$$P(Y=c_k)=\frac{\sum_{i=1}^{N} I(y_i=c_k)}{N}, k=1,2...,K$$

设第 j 个特征 $x^{(j)}$ 可能取值的集合为 $\{a_{j1},a_{j2},...,a_{jS_j}\}$

- 条件概率 $P(X^{(j)}=a_{js}|Y=c_K)$ 的极大似然估计是

$$P(X^{(j)}=a_{js}|Y=c_k)=\frac{\sum_{i=1}^{N} I(x_i^{(j)}=a_{js},y_i=c_k)}{\sum_{i=1}^{N} I(y_i=c_k)}$$

式中，$x_i^{(j)}$ 是第 i 个样本的第 j 个特征；a_{js} 是第 j 个特征可能取的第 s 个值；I 为指示函数

学习与分类算法 I

朴素贝叶斯算法 (naive Bayes algorithm)

输入 训练数据 $T = \{(x_1, y_1), (x_2, y_2), ..., (x_N, y_N)\}$, 其中
$x = (x_i^{(1)}, x_i^{(2)}, ..., x_i^{(n)})^T$,$x_i^{(j)}$ 是第 i 个样本的第 j 个特征,
$x_i^{(j)} \in \{a_{j1}, a_{j2}, ..., a_{jS_j}\}$,$a_{js}$ 是第 j 个特征可能取的第 s 个值,
j={1,2,...,n}, s={1,2,...,S_j}, $y_i \in \{c_1, c_2, ..., c_K\}$;
实例 x

输出 实例 x 的分类

- (1) 计算先验概率及条件概率

$$P(Y = c_k) = \frac{\sum_{i=1}^{N} I(y_i = c_k)}{N}, k = 1, 2, ..., K$$

$$P(X^{(j)} = a_{js} | Y = c_k) = \frac{\sum_{i=1}^{N} I(x_i^{(j)}, y_i = c_k)}{\sum_{i=1}^{N} I(y_i = c_k)}$$

,

$$j = 1, 2, ..., n; s = 1, 2, ..., S_j; k = 1, 2, ..., K$$

学习与分类算法 II

- (2) 对于给定的实例 $x = (x^{(1)}, x^{(2)}, ..., x^{(n)})^T$, 计算

$$P(Y = c_k) \prod_{j=1}^{n} P(X^{(j)} = x^{(j)} | Y = c_k), k = 1, 2, ..., K$$

- (3) 确定实例 x 的类

$$y = \arg\max_{c_k} P(Y = c_k) \prod_{j=1}^{n} P(X^{(j)} = x^{(j)} | Y = c_k)$$

目录

贝叶斯估计 I

- 设想一下：工商大学新生来了个
  身高：190cm，
  体重：150kg
  性别：?
  已知：190cm 的男生有 50 人，体重均小于 100kg，女生 0 人
- 确定实例 x 的类

$$y = \arg\max_{c_k} P(Y = c_k) \prod_{j=1}^{n} P(X^{(j)} = x^{(j)} | Y = c_k)$$

贝叶斯估计 I

用极大似然估计可能会出现所要估计的概率值为 0 的情况. 这时会影响到后验概率的计算结果, 使分类产生偏差解决这一问题的方法是采用贝叶斯估计. 具体地, 条件概率的贝叶斯估计是

$$P_\lambda(X^{(j)} = a_{js}|\, Y = c_k) = \frac{\sum_{i=1}^{N} I(x_i^{(j)} = a_{js}, y_i = c_k) + \lambda}{\sum_{i=1}^{N} I(y_i = c_k) + S_j\lambda} \tag{16}$$

式中 $\lambda \geq 0$. 等价于在随机变量各个取值的频数上赋予一个正数 $\lambda > 0$. 当 $\lambda = 0$ 时就是极大似然估计. 常取 $\lambda = 1$, 这时称为拉普拉斯平滑 (Laplace smoothing). 显然, 对任何 $s = 1, 2, ..., S_j, k = 1, 2, ..., K$, 有

$$P_\lambda(X^{(j)} = a_{js}|\, Y = c_k) > 0$$

$$\sum_{s=1}^{S_j} P_\lambda(X^{(j)} = a_{js}|\, Y = c_k) = 1$$

贝叶斯估计 II

表明式 (16) 确为一种概率分布. 同样, 先验概率的贝叶斯估计是

$$P_\lambda(Y = c_k) = \frac{\sum_{i=1}^{N} I(y_i = c_k) + \lambda}{N + K\lambda}, k = 1, 2, ..., K$$

作业

- 有 1000 个水果样例. 它们可能是香蕉，橙子或其它水果，已知每个水果的 3 种特性:
  - 是否偏长
  - 是否甜
  - 颜色是否是黄色

类型	长	不长	甜	不甜	黄色	非黄	Total
香蕉	400	100	350	150	450	50	500
橙子	0	300	150	150	300	0	300
其它	100	100	150	50	50	150	200
Total	500	500	650	350	800	200	1000

- 根据上表数据，分别利用朴素贝叶斯分类和贝叶斯估计方法，对一个（长，甜，黄色）水果进行识别，判断该水果属于：香蕉，橙子或其它水果哪一类?