数据挖掘与应用

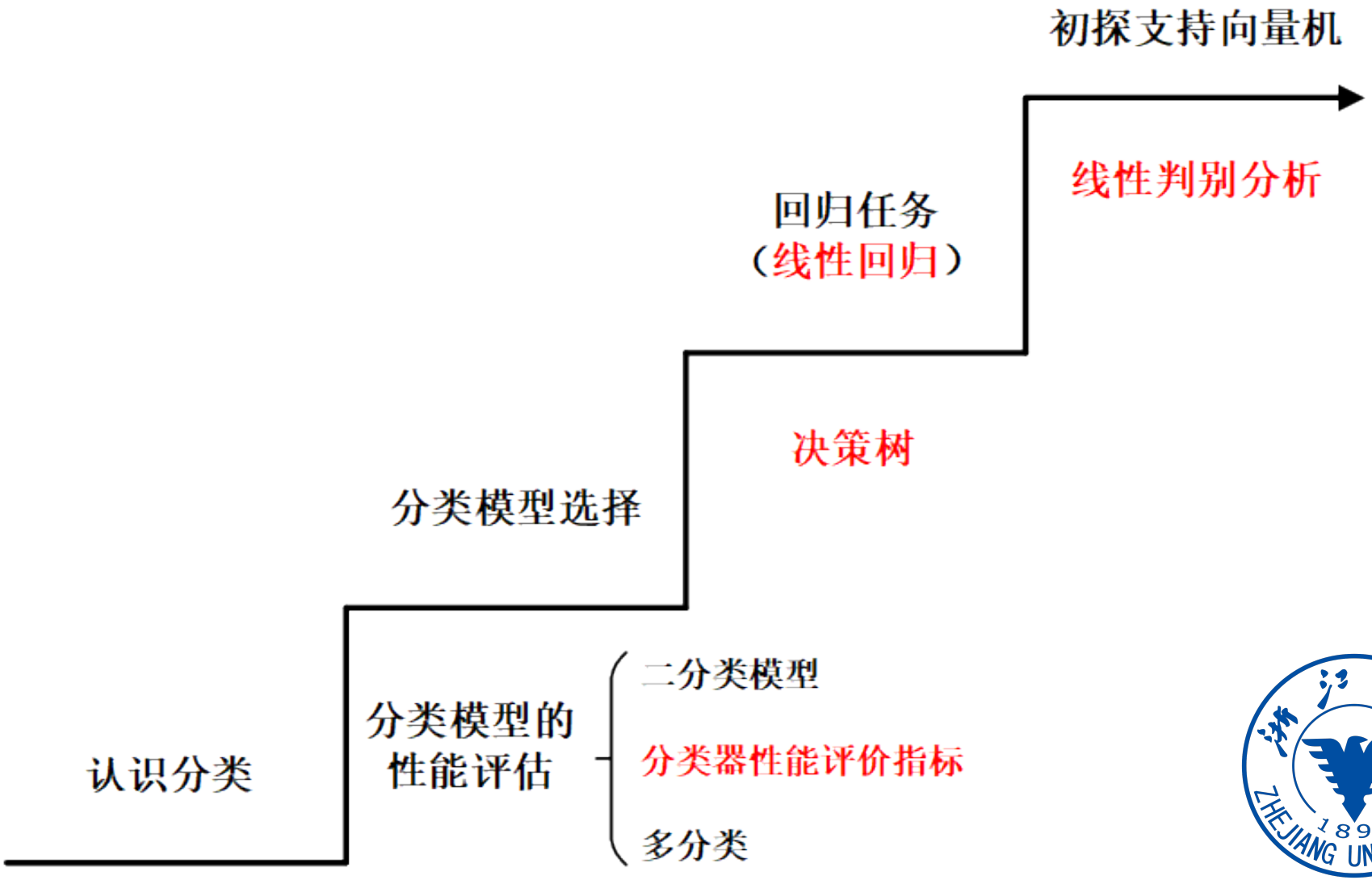
分类

授课教师：周晟

浙江大学　软件学院

2022.09

上节课回顾


初探支持向量机
线性判别分析
回归任务
（线性回归）
决策树
分类模型选择
二分类模型
分类器性能评价指标
多分类
分类模型的
性能评估
认识分类

数据挖掘十大算法 [1]

1. C4.5 √
2. CART √
3. AdaBoost √
4. SVM
5. Naive Bayes
6. EM
7. Apriori（频繁项挖掘）
8. k-Means
9. PageRank
10. kNN

课程内容

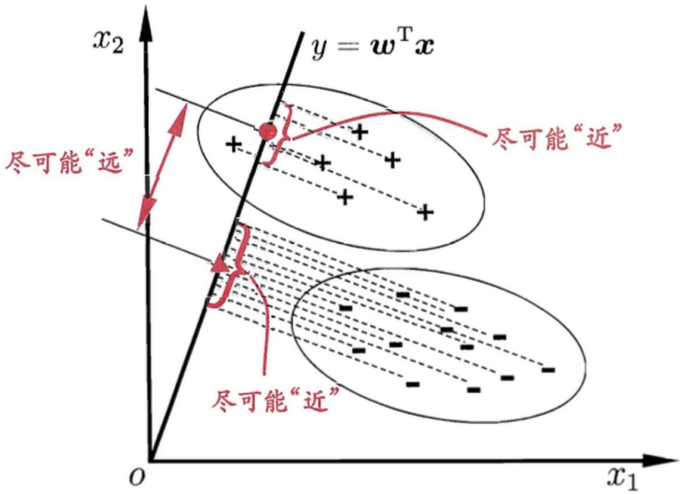
1. 支持向量机 SVM
   - 超平面
   - 函数间隔与几何间隔
   - 支持向量机优化

2. 贝叶斯分类器
   - 贝叶斯决策论
   - 贝叶斯分类器
   - 拉普拉斯修正

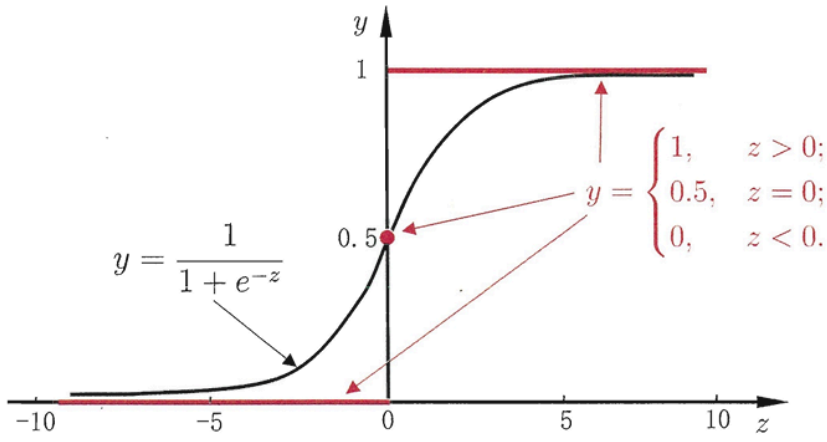
SVM

一直以来学术界和工业界甚至只是学术界里做理论的和做应用的之间，都有一种“鸿沟”。而 SVM 则正好是一个特例，在两边都混得开。

向量空间的分类器



LDA



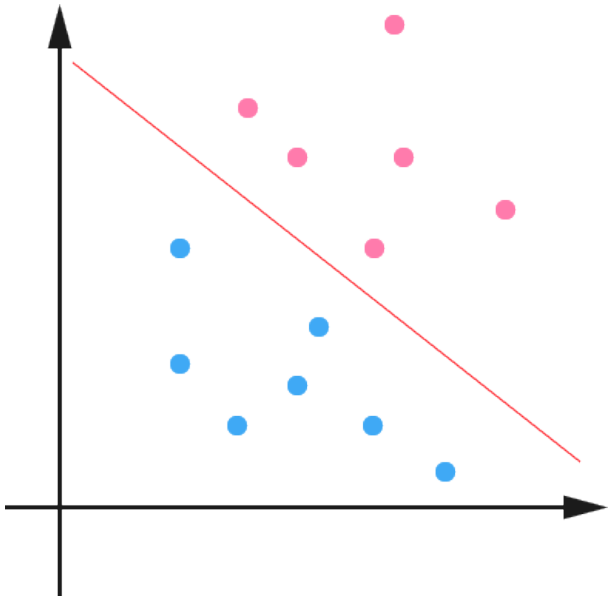
Logistic Regression

分类与超平面

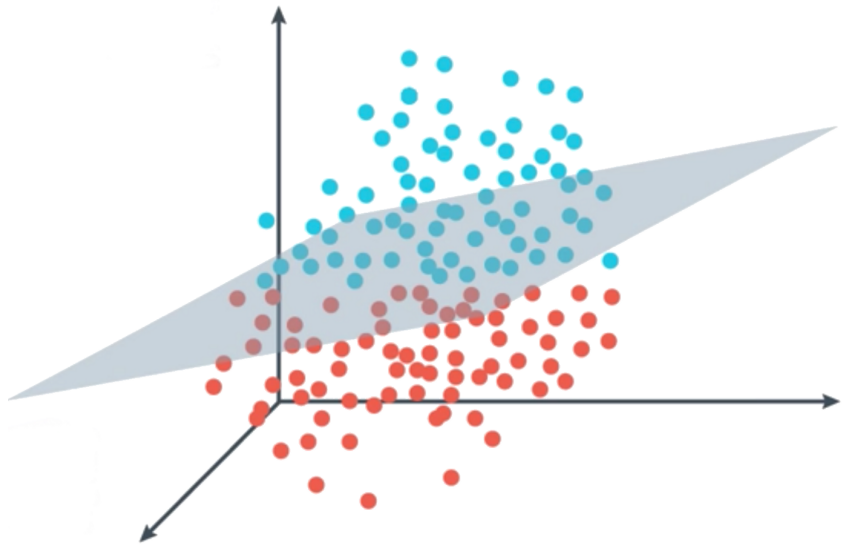
超平面

n 维特征空间中的线性分类器就是要在特征空间中找到一个超平面，使得两类数据尽可能分布在超平面的两侧。超平面可以表示为：

$$w^T x + b = 0$$



二维空间超平面

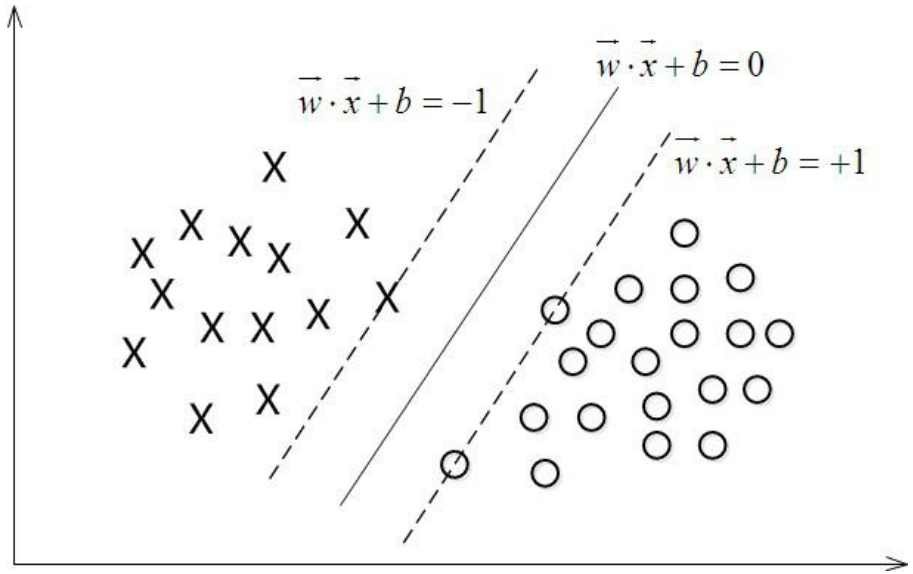


三维空间超平面

分类与超平面

为了计算方便，首先将类别信息数值化。

$$f(x) = w^T x + b \begin{cases} > 0 & y = 1 \\ = 0 & \text{超平面} \\ < 0 & y = -1 \end{cases}$$

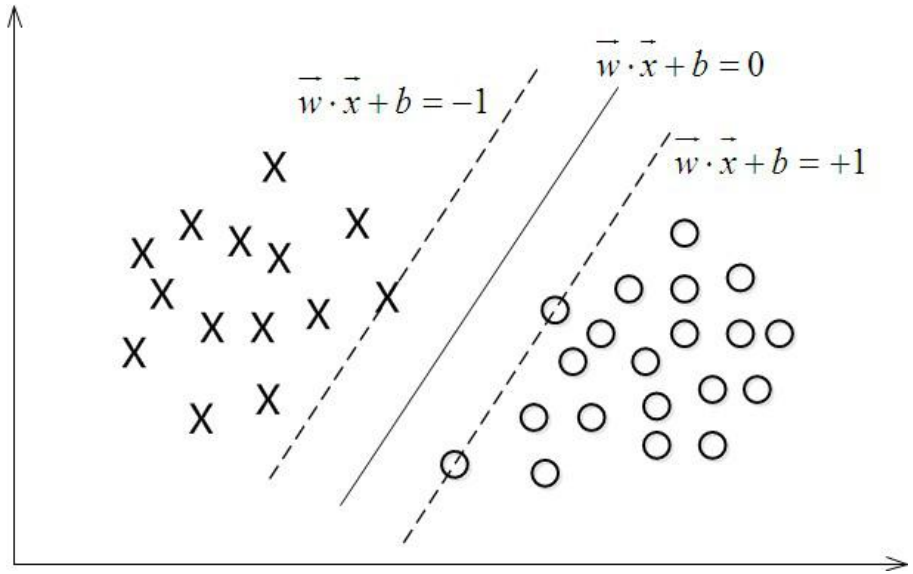


分类的超平面

分类与超平面

为了计算方便，首先将类别信息数值化。

$$f(x) = w^T x + b \begin{cases} > 0 & y = 1 \\ = 0 & \text{超平面} \\ < 0 & y = -1 \end{cases}$$

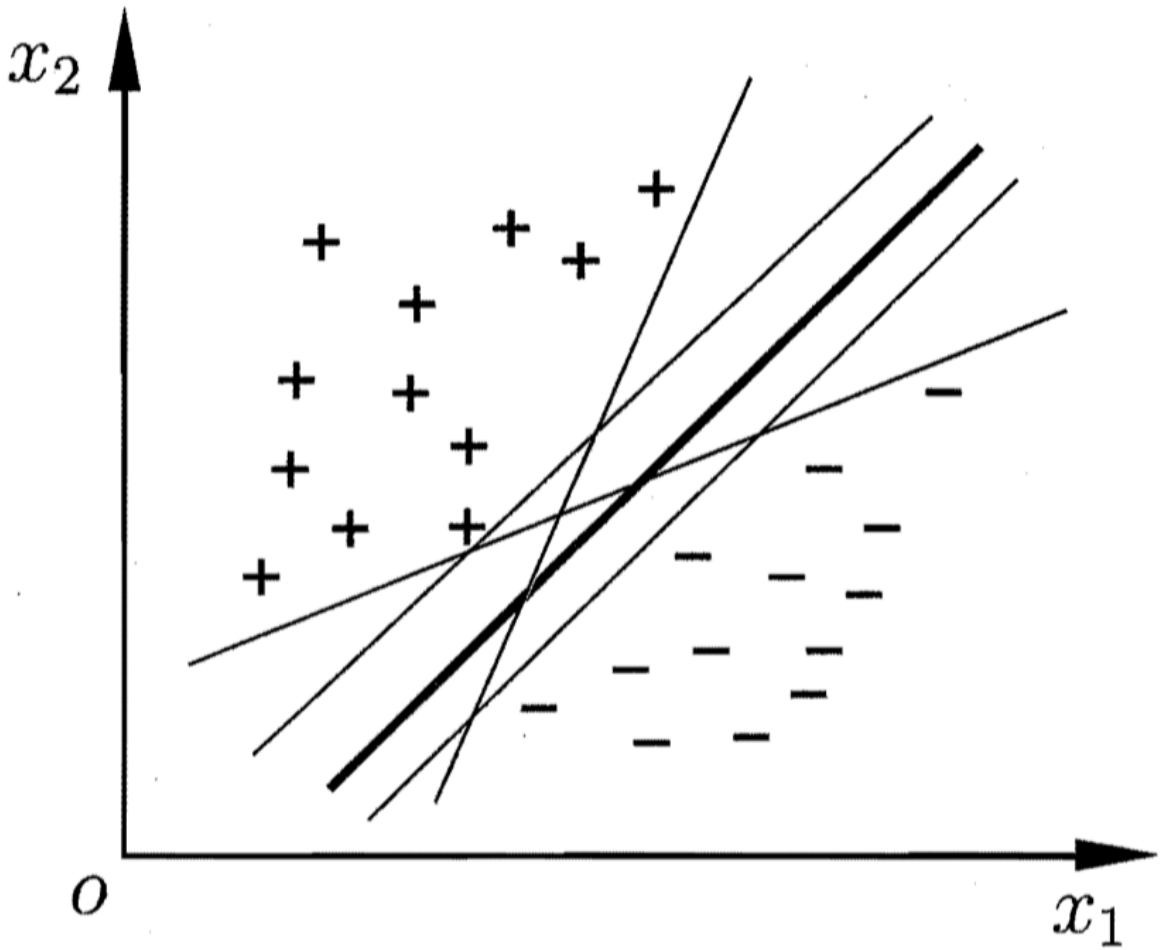


分类的超平面

优点：

1. 通过把 x 带入到 $f(x) = w^T x + b$ 算出结果，即可根据其正负号来进行类别划分。
2. $y \cdot f(x) > 0$

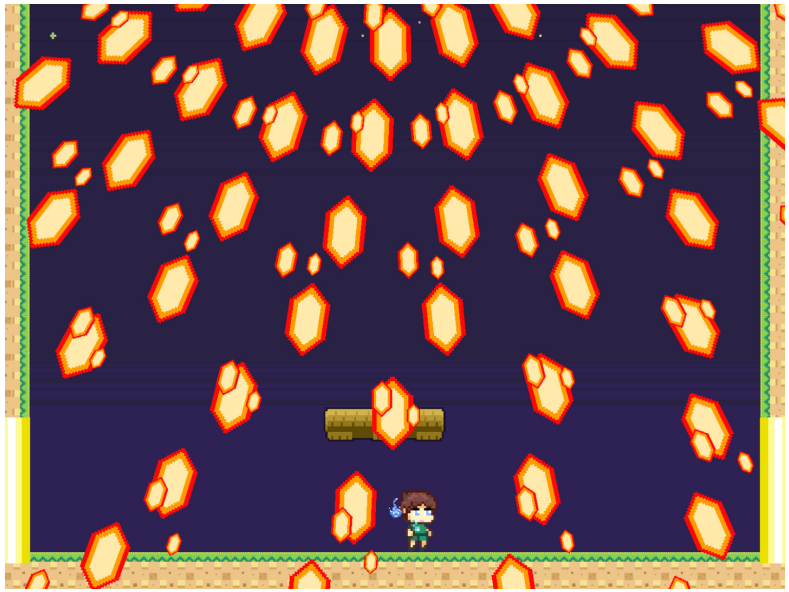
超平面的选择



超平面一定存在嘛？超平面如何选择？

超平面的选择

安全行车



弹幕游戏

SVM 的优化目标

SVM 的优化目标

SVM 的优化目标是寻找空间中的一个超平面，使得两类节点分布在超平面的两侧且距离超平面尽量远。

1. 样本到超平面的距离如何定义?
   1. 函数间隔
   2. 几何间隔
2. 如何定义尽量远?
   1. 距离超平面最近的节点定义
   2. 较远节点可以忽略

函数间隔与几何间隔

函数间隔

函数间隔（Functional margin）定义为：

$$\hat{\gamma} = y\left(w^T x + b\right) = yf(x)$$

几何间隔

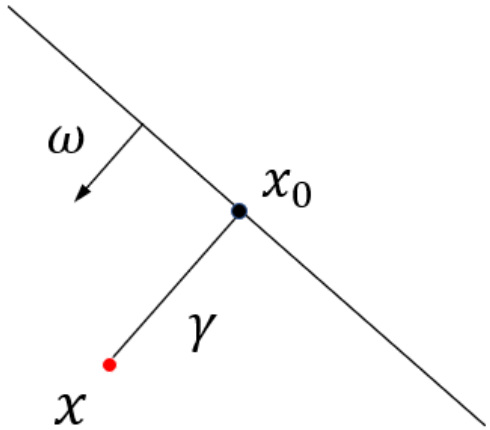
几何间隔（Geometrical margin）定义为点到超平面的距离：

$$\tilde{\gamma} = y\gamma = \frac{\hat{\gamma}}{\|w\|}$$

函数间隔与几何间隔

定义样本 x 在超平面上的投影为 x_0，w 是垂直于超平面的向量，易得：

$$x = x_0 + \gamma \frac{w}{\|w\|}$$



由于 x_0 是超平面上的点，满足 $f(x_0) = 0$，易得：

$$\gamma = \frac{w^T x + b}{\|w\|} = \frac{f(x)}{\|w\|}$$

因此，函数间隔与几何间隔满足关系：

$$\tilde{\gamma} = y\gamma = \frac{\hat{\gamma}}{\|w\|}$$

SVM 的优化目标

SVM 的优化目标 2

SVM 的优化目标是寻找空间中的一个超平面，使得节点到超平面的函数/几何间隔足够大。

函数间隔的缺陷

通过等比例缩放 $||w||$，函数间隔可以在超平面不变的情况下被取得任意大，因此在同一超平面下，函数间隔的取值可以无限大，不适合作为优化目标。

SVM 的优化目标 3

$$\max \tilde{\gamma} = \max \frac{\hat{\gamma}}{\|w\|} = \max \frac{y\left(w^T x + b\right)}{\|w\|}$$

$$s.t. \quad y_i\left(w^T x_i + b\right) = \hat{\gamma}_i \geq \hat{\gamma}, \quad i = 1, \ldots, n$$

最优超平面

为了计算方便，令函数间隔为 1(距离超平面最近的点)：

$$\hat{\gamma} = y\left(w^T x + b\right) = yf(x) = 1$$

目标函数转变为：

$$\max \frac{1}{\|w\|} \quad \text{s.t.} \quad y_i(w^T x_i + b) \geq 1, i = 1, \ldots, n$$

支持向量

距离超平面最近的点，称之为支持向量：

$$y(w^T x + b) = 1$$

特征空间中的其他点 $(y(w^T x + b) > 1)$ 点对最有超平面的学习没有实质贡献，因此可以最大程度地提升存储和计算的效率。

支持向量机的优化

支持向量机的优化 4

$$\max \frac{1}{\|w\|} \quad \text{s.t. } , y_i \left(w^T x_i + b\right) \geq 1, i = 1, \ldots, n$$

为了计算方便，将优化目标等价变换为带约束的二次线性优化：

支持向量机的优化 5

$$\min \frac{1}{2}\|w\|^2 \quad \text{s.t. } , y_i \left(w^T x_i + b\right) \geq 1, i = 1, \ldots, n$$

优化方法：

1. 二次优化直接求解
2. 拉格朗日乘子

使用拉格朗日乘子优化 SVM

利用拉格朗日乘子法，将带约束的优化问题转化为不带约束的优化问题：

$$\mathcal{L}(\boldsymbol{w}, b, \alpha) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{n} \alpha_i \left(y_i \left(\boldsymbol{w}^T x_i + b\right) - 1\right)$$

其中 α_i 是每个约束条件对应的拉格朗日乘子。

令

$$\theta(w) = \max_{\alpha_i \geq 0} \mathcal{L}(w, b, \alpha)$$

易证带约束的最小化 $\frac{1}{2}\|\boldsymbol{w}\|^2$ 等价于最小化 $\theta(w)$

使用拉格朗日乘子优化 SVM

SVM 优化目标

$$\theta(w) = \max_{\alpha_i \geq 0} \mathcal{L}(w, b, \alpha) = \max \frac{1}{2}\|w\|^2 - \sum_{i=1}^{n} \alpha_i \left(y_i \left(\boldsymbol{w}^T x_i + b \right) - 1 \right)$$

为何等价性成立?

1. 若样本不满足约束条件:$y_i(\boldsymbol{w}^T x_i + b) < 1$，则 $\theta(w) = +\infty$
2. 若样本满足约束条件，则 $\alpha_i = 0$，因此 $\theta(w) = \frac{1}{2}\|w\|^2$

Support Vector!

使用拉格朗日乘子优化 SVM

利用拉格朗日乘子，SVM 的优化目标转化为：

SVM 的优化目标

$$\min_{w,b} \theta(w) = \min_{w,b} \max_{\alpha_i \geq 0} \mathcal{L}(w, b, \alpha) = p^*$$

由于 SVM 优化目标满足KKT 条件，因此将 SVM 的优化进一步转化为拉格朗日乘子的对偶问题:

$$\max_{\alpha_i \geq 0} \min_{w,b} \mathcal{L}(w, b, \alpha) = d^*$$

SVM 的拉格朗日对偶问题求解

$$\frac{\partial \mathcal{L}}{\partial w} = 0 \Rightarrow w = \sum_{i=1}^{n} \alpha_i y_i x_i$$

$$\frac{\partial \mathcal{L}}{\partial b} = 0 \Rightarrow \sum_{i=1}^{n} \alpha_i y_i = 0$$

代入可得

$$\begin{aligned}\mathcal{L}(w, b, \alpha) &= \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^T x_j - \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^T x_j - b \sum_{i=1}^{n} \alpha_i y_i + \sum_{i=1}^{n} \alpha_i \\ &= \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^T x_j\end{aligned}$$

SVM 的拉格朗日对偶问题求解

$$\begin{aligned} &\max_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^T x_j \\ &\text{s.t. } , \alpha_i \geq 0, i = 1, \ldots, n \\ &\sum_{i=1}^{n} \alpha_i y_i = 0 \end{aligned}$$

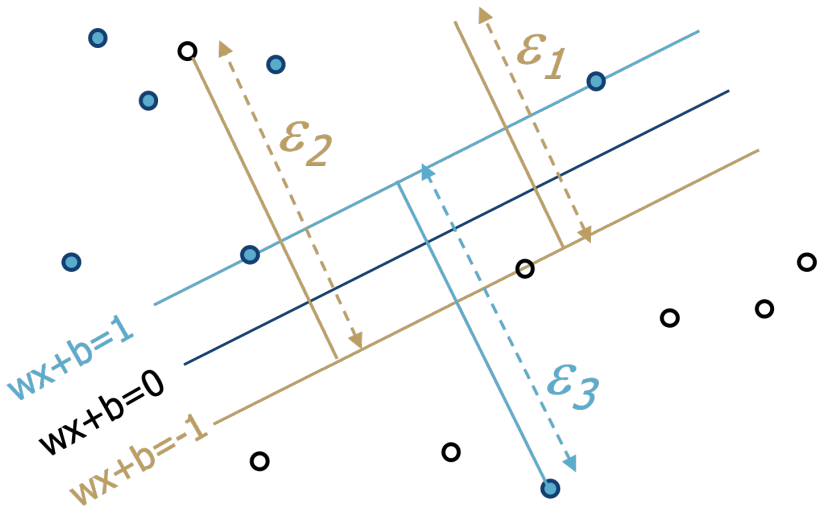
由于在梯度为 0 时，需满足

$$w = \sum_{i=1}^{n} \alpha_i y_i x_i$$

因此对于未知标签的节点，可得：

$$\begin{aligned} f(x) &= \left(\sum_{i=1}^{n} \alpha_i y_i x_i \right)^T x + b \\ &= \sum_{i=1}^{n} \alpha_i y_i \langle x_i, x \rangle + b \end{aligned}$$

非线性可分下的 SVM



非线性可分下的 SVM

$$\min_{w,b,\xi} \frac{1}{2} w^T w + C \left(\sum_{i=1}^{l} \xi_i \right)$$

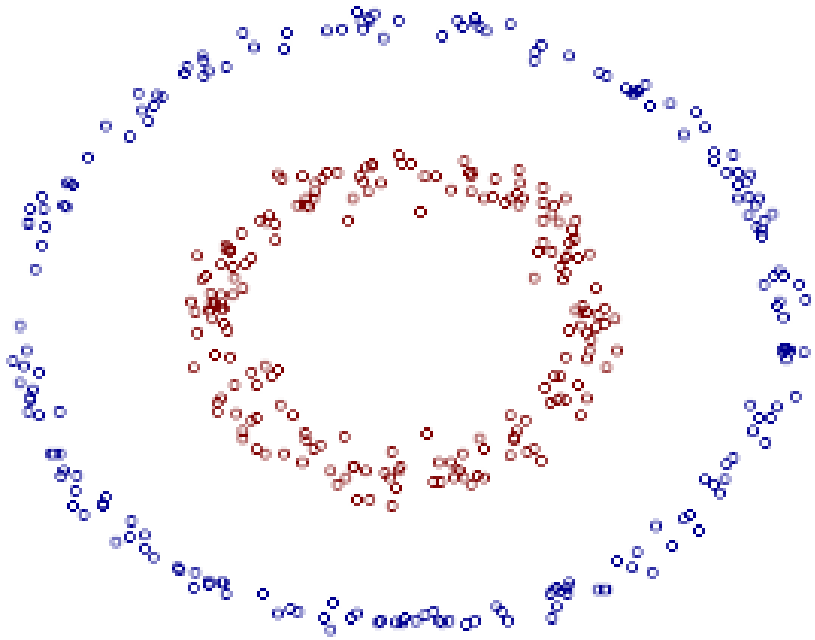
$$y_i \left(\left(w^T x_i \right) + b \right) \geq 1 - \xi_i$$

$$\xi_i \geq 0, i = 1, \ldots, l$$

$$L_P \equiv \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{l} \xi_i - \sum_{i=1}^{l} \alpha_i \left[y_i \left(w \cdot x_i + b \right) - 1 + \xi_i \right] - \sum_{i=1}^{l} \mu_i \xi_i$$

$$L_D \equiv \sum_i \alpha_i - \frac{1}{2} \alpha^T H \alpha \quad \text{s.t. } 0 \leq \alpha_i \leq C \quad \text{and} \quad \sum_i \alpha_i y_i = 0$$

非线性可分下的 SVM



$$a_1X_1 + a_2X_1^2 + a_3X_2 + a_4X_2^2 + a_5X_1X_2 + a_6 = 0$$

$$\sum_{i=1}^{5} a_iZ_i + a_6 = 0$$

非线性可分下的 SVM

在新空间（假设线性可分）中的对偶优化问题：

$$\max_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \left\langle \phi\left(x_i\right), \phi\left(x_j\right) \right\rangle$$

$$\text{s.t.} \alpha_i \geq 0, i = 1, \ldots, n$$

$$\sum_{i=1}^{n} \alpha_i y_i = 0$$

在新空间（假设线性可分）中的 SVM：

$$f(x) = \sum_{i=1}^{n} \alpha_i y_i \left\langle \phi\left(x_i\right), \phi(x) \right\rangle + b$$

核函数

核函数

计算两个向量在映射过后的空间中的内积的函数叫做核函数 (Kernel Function)，它能简化映射空间中的内积运算

常见的核函数包括：

1. 线性核 $\kappa(x_1, x_2) = \langle x_1, x_2 \rangle$
2. 多项式核 $\kappa(x_1, x_2) = (\langle x_1, x_2 \rangle + R)^d$
3. 高斯核 $\kappa(x_1, x_2) = \exp\left(-\frac{\|x_1 - x_2\|^2}{2\sigma^2}\right)$

SVM 总结

1. Supprot Vector 和 Margin
2. 使用拉格朗日乘子优化 SVM
3. 转化为拉格朗日乘子的对偶问题
4. 非线性可分下的 SVM
5. Kernel 函数

1. 支持向量机 SVM
   - 超平面
   - 函数间隔与几何间隔
   - 支持向量机优化

2. 贝叶斯分类器
   - 贝叶斯决策论
   - 贝叶斯分类器
   - 拉普拉斯修正

贝叶斯决策论

贝叶斯决策论

在所有相关概率都已知的理想情形下，贝叶斯决策论考虑如何基于这些概率和误判所造成的损失来选择最优的类别标签。

基于后验概率 $P(c_i|x)$，可以获得将样本 x 分类为 c_i 所产生的期望损失，即在样本 x 上的“条件风险”:

$$R(c_i|x) = \sum_{j=1}^{K} \lambda_{ij} P(c_j|x)$$

$\lambda_{i,j}$ 是将一个真实标签为 c_j 的样本误分类为 c_i 所产生的损失。

贝叶斯决策论

优化目标是最小化分类错误率，则误判损失 λ_{ij} 可以写做：

$$\lambda_{ij} = \begin{cases} 0, & if \quad i = j; \\ 1, & otherwise, \end{cases}$$

此时条件风险为：

$$R(c|x) = 1 - P(c|x) \qquad (P(c_j|x)\text{为 0 或 1})$$

因此，最小化分类错误率的贝叶斯最优分类器为：

$$h^*(x) = \underset{c \in Y}{\arg\max}\, P(c|x)$$

即对于每个样本 x，选择能使其后验概率 $P(c|x)$ 最大的类别标签

贝叶斯分类器

贝叶斯分类器

贝叶斯分类器的目标是基于有限的训练样本集来尽可能准确地估计出后验概率 $P(c|x)$

大体来说，主要有两种策略：

- 给定 x，可通过直接建模 $P(c|x)$ 来预测 c，这样得到的是“判别式模型”(决策树、逻辑回归、SVM)
- 先对联合概率分布 $P(x, c)$ 建模，然后再由此获得 $P(c|x)$，这样得到的是“生成式模型”

贝叶斯公式

$$P(c|x) = \frac{P(x,c)}{P(x)} = \frac{P(c)P(x|c)}{P(x)}$$

类先验概率 似然 证据因子

$P(c)$ 是类“先验”概率；
$P(x|c)$ 是样本 x 相对于类标记 c 的类条件概率，或称为“似然”；
$P(x)$ 是用于归一化的“证据”因子。

Thomas Bayes
1702 - 1761

贝叶斯分类器

$$P(c|x) = \frac{P(c)P(x|c)}{P(x)}$$

证据因子$P(x)$ 与类标签无关 $\Longrightarrow$ 如何基于训练数据 D 来估计先验$P(c)$ 和似然$P(x|c)$

根据大数定律，当训练集包含充足的独立同分布样本时，类先验概率$P(c)$ 可通过各类样本出现的概率进行估计。

对于类条件概率（似然）$P(x|c)$ 来说，直接用频率来估计是不可行的，因为“未被观测到”$\neq$“出现概率为零”

极大似然估计

统计学界的两个学派分别提供了不同的解决方案：

- 频率主义学派认为参数虽然未知，但却是客观存在的固定值。因此，可通过优化似然函数等准则来确定参数值
- 贝叶斯学派则认为参数是未观察到的随机变量，其本身也可有分布。因此，可假定参数服从一个先验分布，然后基于观测到的数据来计算参数的后验分布。

极大似然估计源自频率主义学派，是根据数据采样来估计概率分布参数的经典方法。

极大似然估计

极大似然估计

假设 $P(x|c)$ 具有确定的形式并被参数向量 θ_c 唯一确定，令 D_c 表示训练集 D 中第 c 类样本组成的集合，且假设这些样本是独立同分布的。参数 θ_c 对于数据集 D_c 的似然是：

$$P(D_c|\theta_c) = \prod_{x \in D_c} P(x|\theta_c)$$

对 θ_c 进行极大似然估计，就是去寻找能最大化似然 $P(D_c|\theta_c)$ 的参数值 $\hat{\theta}_c$。

极大似然估计

上式中的连乘操作易造成下溢，通常使用对数似然（log-likehood）：

$$\begin{aligned} LL(\theta_c) &= \log P(D_c|\theta_c) \\ &= \sum_{x \in D_c} \log P(x|\theta_c) \end{aligned}$$

此时参数 θ_c 的极大似然估计 $\hat{\theta}_c$ 为：

$$\hat{\theta}_c = \arg\max_{\theta_c} LL(\theta_c)$$

极大似然估计

假设似然函数 $p(x|c) \sim \mathcal{N}(\mu_c, \sigma_c^2)$，则参数 μ_c 和 σ_c^2 的极大似然估计为：

$$\hat{\mu}_c = \frac{1}{|D_c|} \sum_{x \in D_c} x$$

$$\hat{\sigma}_c^2 = \frac{1}{|D_c|} \sum_{x \in D_c} (x - \hat{\mu}_c)(x - \hat{\mu}_c)^T$$

通过极大似然法得到正态分布均值就是样本均值，方差就是 $(x - \hat{\mu}_c)(x - \hat{\mu}_c)^T$ 的均值，这显然是一个符合直觉的结果。

朴素贝叶斯分类器

极大似然估计的问题

类条件概率 $P(x|c)$ 是所有属性上的联合概率，难以从有限的训练样本直接估计而得到。

朴素贝叶斯分类器 (naive Bayes classifier) 采用了属性条件独立性假设：对已知类别，假设所有属性相互独立。即假设每个属性独立地对分类结果发生影响。

$$P(c|x) = \frac{P(c)P(x|c)}{P(x)} = \frac{P(c)}{P(x)} \prod_{i=1}^{d} P(x_i|c)$$

其中 d 为属性数目，x_i 为 x 在第 i 个属性上的取值。

朴素贝叶斯分类器

由于对于所有类别来说 $P(x)$ 相同，因此对应的贝叶斯判定准则为：

$$h_{nb}(x) = \arg\max_{c \in Y} P(c) \prod_{i=1}^{d} P(x_i|c)$$

这也就是朴素贝叶斯分类器的表达式。

由上式可知，朴素贝叶斯分类器的训练过程就是基于训练集 D 来估计类先验概率 $P(c)$，并为每个属性估计条件概率 $P(x_i|c)$。

朴素贝叶斯分类器

令 D_c 表示训练集 D 中第 c 类样本组成的集合，若有充足的独立同分布样本，则可容易地估计出类先验概率：

$$P(c) = \frac{|D_c|}{|D|}$$

对离散属性而言，令 D_{c,x_i} 表示 D_c 中在第 i 个属性上取值为 x_i 的样本组成的集合，则条件概率 $P(x_i|c)$ 可估计为：

$$P(x_i|c) = \frac{|D_{c,x_i}|}{|D_c|}$$

朴素贝叶斯分类器

而对于连续属性，可以考虑概率密度函数。假定 $p(x_i|c) \sim \mathcal{N}(\mu_{c,i}, \sigma_{c,i}^2)$，其中 $\mu_{c,i}$ 和 $\sigma_{c,i}^2$ 分别为第 c 类样本在第 i 个属性上取值的均值和方差，则有：

$$p(x_i|c) = \frac{1}{\sqrt{2\pi}\sigma_{c,i}} \exp\left(-\frac{(x_i - \mu_{c,i})^2}{2\sigma_{c,i}^2}\right)$$

朴素贝叶斯分类器——以西瓜数据集为例

编号	色泽	根蒂	敲声	纹理	脐部	触感	密度	含糖率	好瓜
1	青绿	蜷缩	浊响	清晰	凹陷	硬滑	0.697	0.460	是
2	乌黑	蜷缩	沉闷	清晰	凹陷	硬滑	0.774	0.376	是
3	乌黑	蜷缩	浊响	清晰	凹陷	硬滑	0.634	0.264	是
4	青绿	蜷缩	沉闷	清晰	凹陷	硬滑	0.608	0.318	是
5	浅白	蜷缩	浊响	清晰	凹陷	硬滑	0.556	0.215	是
6	青绿	稍蜷	浊响	清晰	稍凹	软粘	0.403	0.237	是
7	乌黑	稍蜷	浊响	稍糊	稍凹	软粘	0.481	0.149	是
8	乌黑	稍蜷	浊响	清晰	稍凹	硬滑	0.437	0.211	是
9	乌黑	稍蜷	沉闷	稍糊	稍凹	硬滑	0.666	0.091	否
10	青绿	硬挺	清脆	清晰	平坦	软粘	0.243	0.267	否
11	浅白	硬挺	清脆	模糊	平坦	硬滑	0.245	0.057	否
12	浅白	蜷缩	浊响	模糊	平坦	软粘	0.343	0.099	否
13	青绿	稍蜷	浊响	稍糊	凹陷	硬滑	0.639	0.161	否
14	浅白	稍蜷	沉闷	稍糊	凹陷	硬滑	0.657	0.198	否
15	乌黑	稍蜷	浊响	清晰	稍凹	软粘	0.360	0.370	否
16	浅白	蜷缩	浊响	模糊	平坦	硬滑	0.593	0.042	否
17	青绿	蜷缩	沉闷	稍糊	稍凹	硬滑	0.719	0.103	否

训练数据

朴素贝叶斯分类器——以西瓜数据集为例

编号	色泽	根蒂	敲声	纹理	脐部	触感	密度	含糖率	好瓜
测 1	青绿	蜷缩	浊响	清晰	凹陷	硬滑	0.697	0.460	?

测试数据

首先，我们来估计类先验概率 $P(c)$，显然有：

$$P(\text{好瓜} = \text{是}) = \frac{8}{17} \approx 0.471$$

$$P(\text{好瓜} = \text{否}) = \frac{9}{17} \approx 0.529$$

朴素贝叶斯分类器——以西瓜数据集为例

接着，我们为每个属性估计条件概率 $P(x_i|c)$：

$$P_{\text{青绿}\,|\,\text{是}} = P(\text{色泽} = \text{青绿} \mid \text{好瓜} = \text{是}) = \frac{3}{8} = 0.375$$

$$P_{\text{青绿}\,|\,\text{否}} = P(\text{色泽} = \text{青绿} \mid \text{好瓜} = \text{否}) = \frac{3}{9} \approx 0.333$$

$$P_{\text{蜷缩}\,|\,\text{是}} = P(\text{根蒂} = \text{蜷缩} \mid \text{好瓜} = \text{是}) = \frac{5}{8} = 0.625$$

$$P_{\text{蜷缩}\,|\,\text{否}} = P(\text{根蒂} = \text{蜷缩} \mid \text{好瓜} = \text{否}) = \frac{3}{9} \approx 0.333$$

$$P_{\text{浊响}\,|\,\text{是}} = P(\text{敲声} = \text{浊响} \mid \text{好瓜} = \text{是}) = \frac{6}{8} = 0.750$$

$$P_{\text{浊响}\,|\,\text{否}} = P(\text{敲声} = \text{浊响} \mid \text{好瓜} = \text{否}) = \frac{4}{8} \approx 0.444$$

朴素贝叶斯分类器——以西瓜数据集为例

$$P_{\text{清晰}\mid\text{是}} = P(\text{纹理} = \text{清晰} \mid \text{好瓜} = \text{是}) = \frac{7}{8} = 0.875$$

$$P_{\text{清晰}\mid\text{否}} = P(\text{纹理} = \text{清晰} \mid \text{好瓜} = \text{否}) = \frac{2}{9} \approx 0.222$$

$$P_{\text{凹陷}\mid\text{是}} = P(\text{脐部} = \text{凹陷} \mid \text{好瓜} = \text{是}) = \frac{6}{8} = 0.750$$

$$P_{\text{凹陷}\mid\text{否}} = P(\text{脐部} = \text{凹陷} \mid \text{好瓜} = \text{否}) = \frac{2}{9} \approx 0.222$$

$$P_{\text{硬滑}\mid\text{是}} = P(\text{触感} = \text{硬滑} \mid \text{好瓜} = \text{是}) = \frac{6}{8} = 0.750$$

$$P_{\text{硬滑}\mid\text{否}} = P(\text{触感} = \text{硬滑} \mid \text{好瓜} = \text{否}) = \frac{6}{8} \approx 0.667$$

朴素贝叶斯分类器——以西瓜数据集为例

$$
\begin{aligned}
p_{\text{密度:0.697}|\text{ 是}} &= p(\text{密度 } = 0.697 |\ \text{好瓜} = \text{是}) \\
&= \frac{1}{\sqrt{2\pi}\cdot 0.129} exp\left(-\frac{(0.697-0.574)^2}{2\cdot 0.129^2}\right) \approx 1.959 \\
p_{\text{密度:0.697}|\text{ 否}} &= p(\text{密度 } = 0.697 |\ \text{好瓜} = \text{否}) \\
&= \frac{1}{\sqrt{2\pi}\cdot 0.195} exp\left(-\frac{(0.697-0.496)^2}{2\cdot 0.195^2}\right) \approx 1.203 \\
p_{\text{含糖:0.460}|\text{ 是}} &= p(\text{含糖 } = 0.460 |\ \text{好瓜} = \text{是}) \\
&= \frac{1}{\sqrt{2\pi}\cdot 0.101} exp\left(-\frac{(0.460-0.279)^2}{2\cdot 0.101^2}\right) \approx 0.788 \\
p_{\text{含糖:0.460}|\text{ 否}} &= p(\text{含糖 } = 0.460 |\ \text{好瓜} = \text{否}) \\
&= \frac{1}{\sqrt{2\pi}\cdot 0.108} exp\left(-\frac{(0.460-0.154)^2}{2\cdot 0.108^2}\right) \approx 0.066
\end{aligned}
$$

朴素贝叶斯分类器——以西瓜数据集为例

于是，有

$$P(\text{好瓜} = \text{是}) \times P_{\text{青绿}\,|\,\text{是}} \times P_{\text{蜷缩}\,|\,\text{是}} \times P_{\text{浊响}\,|\,\text{是}} \times P_{\text{清晰}\,|\,\text{是}} \times P_{\text{凹陷}\,|\,\text{是}} \times P_{\text{硬滑}\,|\,\text{是}} \times p_{\text{密度:0.697}|\,\text{是}} \times p_{\text{含糖:0.460}|\,\text{是}} \approx 0.063$$

$$P(\text{好瓜} = \text{否}) \times P_{\text{青绿}\,|\,\text{否}} \times P_{\text{蜷缩}\,|\,\text{否}} \times P_{\text{浊响}\,|\,\text{否}} \times P_{\text{清晰}\,|\,\text{否}} \times P_{\text{凹陷}\,|\,\text{否}} \times P_{\text{硬滑}\,|\,\text{否}} \times p_{\text{密度:0.697}|\,\text{否}} \times p_{\text{含糖:0.460}|\,\text{否}} \approx 6.80 \times 10^{-5}$$

由于 $0.063 > 6.80 \times 10^{-5}$，因此，朴素贝叶斯分类器将测试样本“测 1”判别为“好瓜”。

朴素贝叶斯分类器——拉普拉斯修正

需注意，若某个属性值在训练集中没有与某个类同时出现过，则基于以上公式进行概率估计和判别将会出现问题。

例如，在使用西瓜数据集训练朴素贝叶斯分类器时，对一个“敲声 = 清脆”的测试用例，有

$$P_{\text{清脆}\mid\text{是}} = P(\text{敲声} = \text{清脆} \mid \text{好瓜} = \text{是}) = \frac{0}{8} = 0$$

因为朴素贝叶斯表达式的连乘结果为零，因此，无论该样本的其他属性是什么，哪怕在其他属性上明显像好瓜，分类的结果都将是“好瓜 = 否”，这显然不太合理。

朴素贝叶斯分类器——拉普拉斯修正

为了避免其他属性携带的信息被训练集中未出现的属性值“抹去”，在估计概率值时通常要进行“平滑”，常用“拉普拉斯修正”。

具体来说，令 K 表示训练集 D 中可能的类别数，N_i 表示第 i 个属性可能的取值数，则可将估计概率的公式修正为：

$$\hat{P}(c) = \frac{|D_c| + 1}{|D| + K}$$

$$\hat{P}(x_i|c) = \frac{|D_{c,x_i} + 1|}{|D_c| + N_i}$$

朴素贝叶斯分类器——拉普拉斯修正

例如，在上述例子中，类先验概率可估计为：

$$\hat{P}(\text{好瓜} = \text{是}) = \frac{8+1}{17+2} \approx 0.474, \quad \hat{P}(\text{好瓜} = \text{否}) = \frac{9+1}{17+2} \approx 0.526$$

类似的，$P_{\text{青绿} \mid \text{是}}$ 可估计为：

$$\hat{P}_{\text{青绿} \mid \text{是}} = \hat{P}(\text{色泽} = \text{青绿} \mid \text{好瓜} = \text{是}) = \frac{3+1}{8+3} \approx 0.364$$

同时，上文提到的概率 $P_{\text{清脆} \mid \text{是}}$ 可估计为：

$$\hat{P}_{\text{清脆} \mid \text{是}} = \hat{P}(\text{敲声} = \text{清脆} \mid \text{好瓜} = \text{是}) = \frac{0+1}{8+3} \approx 0.091$$

总结

1. 支持向量机 SVM
   - 超平面
   - 函数间隔与几何间隔
   - 支持向量机优化

2. 贝叶斯分类器
   - 贝叶斯决策论
   - 贝叶斯分类器
   - 拉普拉斯修正

参考文献

Wu, X., Kumar, V., Ross Quinlan, J., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Ng, A., Liu, B., Yu, P. S., et al.

Top 10 algorithms in data mining.

Knowledge and information systems 14, 1 (2008), 1–37.