

The background features a complex network of thin grey lines and dots, forming a web-like structure. Scattered throughout are various triangles of different sizes and orientations, some with solid black dots at their vertices. The overall aesthetic is minimalist and technical, suggesting a focus on data or mathematics.

预测类模型

张恒瑞

Outline

1. 时间序列预测
2. 灰色预测模型
3. 微分方程与差分方程
4. 机器学习方法
 1. 线性回归与逻辑回归
 2. 支持向量机
 3. 朴素贝叶斯
 4. 神经网络





01

时间序列预测

时间序列预测


时间序列预测是一种历史资料延伸预测，是以时间序列所能反映的社会经济现象的发展过程和规律性，进行引申外推，预测其发展趋势的方法。

和回归分析模型的预测不同，时间序列模型是依赖于事件发生的先后顺序的，同样大小的值改变顺序后输入模型产生的结果是不同的。

时间序列可以分为平稳序列和非平稳序列

本节介绍自回归模型(AR)、移动平均法(MA)、指数平滑法(ES)...

$$x_1, x_2, \dots, x_{t-1} \rightarrow x_t$$

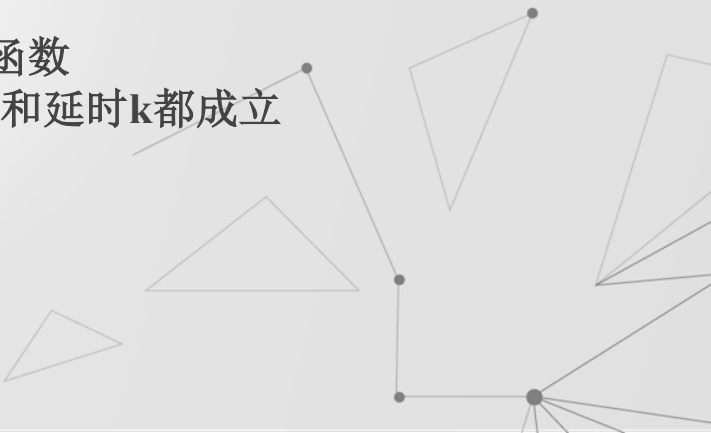


平稳序列

一个时间序列，如果均值没有系统的变化(无趋势)、方差没有系统的变化，且严格消除了周期性变化，就称为平稳序列。

刻画了时间序列的统计性质关于时间平移的不变性(一阶、二阶矩是时间平移不变的)

定义:

1. 随机变量 X_t 的均值 $\mu(t)$ 是常数函数
 2. $\gamma(t, t-k) = \gamma(0, k)$ 对任意的时间 t 和延时 k 都成立
- 

自回归(Auto Regression)

短期预测是时间序列预测的主要目的。时间序列分析的理论基础很简单：设若时间序列(或随机过程)的任一元素 与其前期元素 (x_{t-1} x_{t-2} 等) 之间存在某种关联，则我们可以根据该时间序列的以往观测值来预测其在未来的取值。

上述思路的直接体现便是自回归模型。定义 p 阶自回归过程

$$x_t = \phi_0 + \phi_1 x_{t-1} + \phi_2 x_{t-2} + \cdots + \phi_p x_{t-p} + \varepsilon_t$$

其中 ϕ 为待求参数, p 为滞后期限的数目, ε_p 为白噪声

随机变量 x_t 的取值是前 p 期 $x_{\{t1\}}$, $x_{\{t2\}}$, ..., $x_{\{tp\}}$ 的多元线性回归。





移动平均过程(moving average, MA)

具有下列结构的模型称为q阶自回归模型，简记为MA(q)

$$x_t = \mu + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \cdots - \theta_q \varepsilon_{t-q}$$

即在t时刻的随机变量 x_t 的取值是前 q 期的随机扰动 $\varepsilon_{\{t-1\}}, \varepsilon_{\{t-2\}}, \dots, \varepsilon_{\{t-q\}}$ 的多元线性函数。误差项是当前的随机干扰 ε_t ，为零均值白噪声序列， μ 是序列 $\{x_t\}$ 的均值。认为 x_t 主要受过去 q 期的误差项影响。





自回归移动平均模型(ARMA)


设法将自回归过程AR和移动平均过程MA结合起来

$$x_t = \phi_0 + \phi_1 x_{t-1} + \phi_2 x_{t-2} + \cdots + \phi_p x_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \cdots - \theta_q \varepsilon_{t-q}$$


ARMA最常用的平稳序列拟合模型

随机变量 X_t 的取值 x_t 不仅与以前 p 期的序列值有关，还和前 q 期的随机扰动有关。






使用方法

1. 对观测值序列进行平稳性检测, 如果不平稳, 则对其进行差分运算直到差分后的数据平稳
 2. 进行白噪声检测, 白噪声是指零均值零方差的随机平稳序列
 3. 如果是平稳非白噪声序列就计算其自相关系数和偏自相关系数, 选择合适的模型
 4. 确定模型参数, 应用预测并进行误差分析
- 



02

灰色预测




灰色预测

灰色系统: 系统中的一部分信息是已知的, 另一部分信息是未知的, 系统内各因素之间有不确定的关系

灰色预测: 灰色预测法是一种预测灰色系统的预测方法

通过鉴别系统因素之间发展趋势的相异程度, 即进行关联分析, 并对原始数据进行生成处理来寻找系统变动的规律, 生成有较强规律性的数据序列



灰色生成数列

一切灰色序列都能通过某种生成弱化其随机性，显示其规律性。数据生成的常用方法有：累加生成、累减生成和加权邻值生成。

(1) 累加生成(AGO)

设原始数列为 $x^{(0)} = (x^{(0)}(1), x^{(0)}(2), \dots, x^{(0)}(n))$ 令

$$x^{(1)}(k) = \sum_{i=1}^k x^{(0)}(i) \quad , k = 1, 2, \dots, n$$

$$x^{(1)} = (x^{(1)}(1), x^{(1)}(2), \dots, x^{(1)}(n))$$

所得到的新数列称为为数列 $x^{(0)}$ 的1次累加生成数列，类似的有

$$x^{(r)}(k) = \sum_{i=1}^k x^{(r-1)}(i) \quad , k = 1, 2, \dots, n, r \geq 1$$

灰色生成数列

一切灰色序列都能通过某种生成弱化其随机性，显示其规律性。数据生成的常用方法有：累加生成、累减生成和加权邻值生成。

(2)累减生成(IAGO)

设原始数列为 $x^{(1)} = (x^{(1)}(1), x^{(1)}(2), \dots, x^{(1)}(n))$, 令

$$x^{(0)}(k) = x^{(1)}(k) - x^{(1)}(k-1) \quad , k = 2, 3, \dots, n$$

所得到的新数列 $x^{(0)}$ 称为 $x^{(1)}$ 的1次累减生成数列

可以看出，通过累加数列得到的新数列，可以通过累减生成还原成原始数列

灰色生成数列

一切灰色序列都能通过某种生成弱化其随机性，显示其规律性。数据生成的常用方法有：累加生成、累减生成和加权邻值生成。

(3)加权邻值生成

设原始数列为 $x^{(1)} = (x^1(1), x^1(2), \dots, x^1(n))$,称任意一对相邻元素

$x^{(0)}(k-1), x^{(0)}(k)$ 互为邻值。对于常数 $\alpha \in [0, 1]$,令

$$z^{(0)}(k) = \alpha x^{(0)}(k) + (1 - \alpha)x^{(0)}(k-1) \quad , k = 2, 3, \dots, n$$

所得到的数列称为邻值生成数。

灰色模型GM(1, 1)

G表示grey M表示Model

定义 $x^{(1)}$ 的灰导数为

$$d(k) = x^{(0)}(k) = x^{(1)}(k) - x^{(1)}(k-1)$$

令 $z^{(1)}(k)$ 为数列 $x^{(1)}$ 的邻值生成数列，即

$$z^{(1)}(k) = \alpha x^{(1)}(k) + (1 - \alpha)x^{(1)}(k-1)$$

于是定义GM(1,1)的灰微分方程模型为

$$d(k) + \alpha z^{(1)}(k) = b \quad \text{或} \quad x^{(0)}(k) + \alpha z^{(1)}(k) = b$$

灰色模型GM(1, 1)

将时刻 $k=2,3,\dots$ 代入上式, 有

$$\begin{cases} x^{(0)}(2) + \alpha z^{(1)}(2) = b \\ x^{(0)}(3) + \alpha z^{(1)}(3) = b \\ \dots \\ x^{(0)}(n) + \alpha z^{(1)}(n) = b \end{cases}$$

引入矩阵向量记号

$$u = \begin{bmatrix} a \\ b \end{bmatrix} \quad Y = \begin{bmatrix} x^{(0)}(2) \\ x^{(0)}(3) \\ \dots \\ x^{(0)}(n) \end{bmatrix} \quad B = \begin{bmatrix} -z^{(1)}(2) & 1 \\ -z^{(1)}(3) & 1 \\ \dots & \dots \\ -z^{(1)}(n) & 1 \end{bmatrix}$$

一元线性回归

$$u = \begin{bmatrix} a \\ b \end{bmatrix} = (B^T B)^{-1} B^T Y$$



03

微分(差分)方程预测

微分方程预测

特点:

1. 描述实际对象某些特性随着时间(空间)而演变的过程
2. 分析它的变化规律
3. 预测它的未来形态
4. 特性会给出关于变化率的一些关系

经典案例:

人口预测模型:

模型1: 马尔萨斯指数增长模型

假设了种群增长率 r 为一个常数

模型2: Logistic模型

假设环境只能供养一定数量的种群, 或者说存在竞争

人口增长模型

指数增长模型——马尔萨斯提出（1798）

基本假设：人口(相对)增长率 r 是常数

$$\frac{x(t + \Delta t) - x(t)}{x(t)} = r\Delta t$$

$$\frac{dx}{dt} = rx, x(0) = x_0$$

$$x(t) = x_0 e^{rt}$$

$$x(t) = x_0 (e^r)^t \approx x_0 (1 + r)$$

阻滞增长模型(Logistic模型)

$$\frac{dx}{dt} = rx$$

$$\frac{dx}{dt} = r(x) \cdot x = rx \left(1 - \frac{x}{x_m}\right)$$

$$x(t) = \frac{x_m}{1 + \left(\frac{x_m}{x_0} - 1\right)e^{-rt}}$$

微分方程预测

也有一些非常典型的微分方程预测模型

- (1) 传染病模型
- (2) 经济增长模型
- (3) 正规战与游击战
- (4) 烟雾的扩散与消失

04

线性回归、逻辑回归

线性回归

线性回归几乎是最简单的机器学习模型，它假设因变量和自变量之间是线性关系的，一条直线简单明了。

Setup 给定一份数据集，由一列观测值(y , 即因变量)和多列特征(X , 即自变量)组成。线性回归的目的就是找到和样本拟合程度最佳的线性模型

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad X = \begin{pmatrix} X_1^T \\ X_2^T \\ \vdots \\ X_n^T \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}$$

线性回归

Dataset

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad X = \begin{pmatrix} X_1^T \\ X_2^T \\ \vdots \\ X_n^T \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}$$

β 是系数向量, ϵ 是干扰项

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

线性回归

最后得到的第*i*个 y (观测值)是这样的

$$y_i = \beta_0 \mathbf{1} + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i (i = 1, \cdots, n, \beta_0 \text{为截距})$$

转换成矩阵形式

$$y = X\beta + \varepsilon$$

线性回归

问题:

1. 如何定义最优?

定义度量标准, 对于回归问题, 最常用的标准是均方差(MSE, Mean Squared Error) 均方差越小, 说明测试值和实际值之间的差距越小, 即模型性能更优。

在线性回归的式子中, y 和 X 是由数据集给定的, 而 β 和 ϵ 是不确定的, 也就是说, 找到了最优的 β 和 ϵ , 就找到了最优的模型。

定义目标函数 $L = \sum_{i=1}^m (f(x_i) - y_i)^2$

则

$$(\beta^*, \epsilon^*) = \operatorname{argmin} \sum_{i=1}^m (f(x_i) - y_i)^2$$

线性回归

问题:

2. 如何找到最优?

最常用的是参数估计方法是最小二乘法(Least Square Method), 最小二乘法试图找到一条直线, 使得样本点和直线的欧氏距离之和最小。

这个寻找的过程的简单描述就是: 根据凸函数的性质, 求得其关于 β 和 ϵ 的二阶导数的零点

容易得到

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

$$\epsilon = y - X\hat{\beta} = y - X(X^T X)^{-1} X^T y.$$

线性回归的推广

考虑下面两个回归模型

容易得到

$$y = \beta X + \varepsilon, \quad \ln y = \beta X + \varepsilon$$

左边是之前得到的线性回归模型，右边是对数线性回归模型

完成了从输入空间到输出空间的非线性映射

$$y = g^{-1}(\beta X + \varepsilon)$$

将以上两个式子综合，可以写成更一般的形式，就是广义线性回归模型 (Generalized Linear Model) 里的 g 是一个单调可微函数，称为联系函数 (Link Function)

逻辑回归

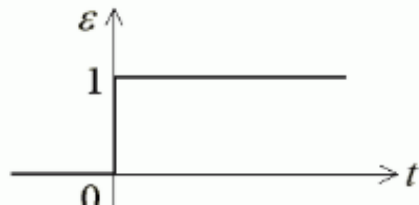
想把线性回归应用到分类问题怎么办？

如：二分类问题，把X对应的y分成类别0和类别1.

线性回归本身的输出是连续的，也就是说要将连续的值分为离散的0和1.
所以需要找到一个联系函数，将X映射到 $y = 0/1$

如阶跃函数

$$\varepsilon(t) = \begin{cases} 1 & t > 0 \\ 0 & t < 0 \end{cases}$$

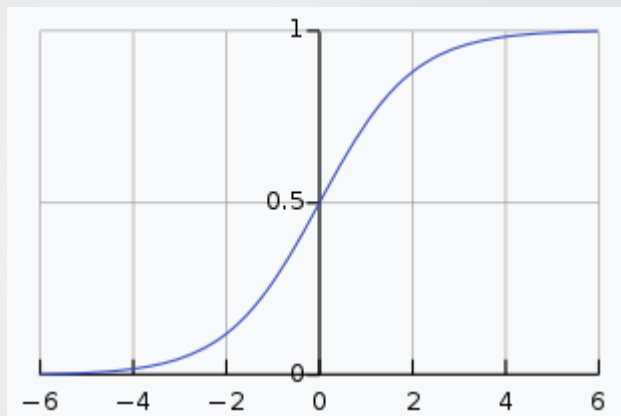


逻辑回归

改用Sigmoid函数(S型函数)作为联系函数，也称为对数几率函数(Logistic Function) 函数图像如下

$$y = \frac{1}{(1 + e^{-z})}$$

$$\ln \frac{y}{1-y} = z$$





逻辑回归

将对数几率函数代入到之前得到的广义线性回归模型中，得到逻辑回归的数学原型

$$y = \frac{1}{1 + e^{-(\beta X + \epsilon)}}$$

逻辑回归是广义线性回归中的一种以对数几率函数为联系函数的特例



逻辑回归

如何求解

定义交叉熵形式的目标函数（代价函数）

$$h_{\theta}(x) = \frac{1}{1 + e^{-\theta^T x}}$$

$$J(\theta) = -\frac{1}{m} \left[\sum_{i=1}^m h_{\theta}(x^{(i)}) \log(h_{\theta}(x^{(i)})) + (1 - h_{\theta}(x^{(i)})) \log(1 - h_{\theta}(x^{(i)})) \right]$$

梯度下降求最优参数

$$\theta_j = \theta_j - \alpha \left(\frac{\partial}{\partial \theta_j} \right) J(\theta) = \theta_j - \alpha \left(\frac{1}{m} \right) \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)}) x_j^{(i)}$$

05

贝叶斯预测模型





先验概率与后验概率

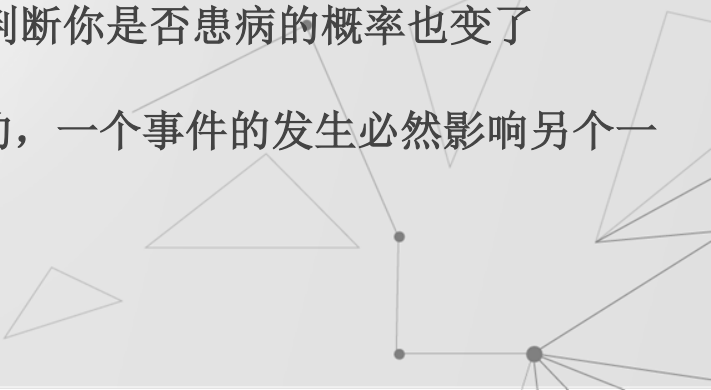
先验概率是不加条件(信息)的情况下判断一件事发生的概率

例：比如你是否聪明的概率、是否患病的概率。

后验概率是在附加了某条件后判断一件事情发生的概率。这个条件可以是已知另一个事件的发生。附加的条件对我们判断签一个事件而言相当于新的信息，因为有了更多信息所以可以做出更可靠的判断。

例： 已知你考上了大学，此时再判断你是否聪明的概率也变了。
已知你检测呈阳性，此时再判断你是否患病的概率也变了

条件概率涉及的两个事件不是独立的，一个事件的发生必然影响另一个事件发生的概率。



贝叶斯定理

后验概率可以用条件概率表示，条件概率公式：

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

$$P(AB) = P(B)P(A|B) = P(A)P(B|A)$$

贝叶斯公式

$$P(A|B) = \frac{P(A)P(B|A)}{P(A)P(B|A) + P(A^{\text{逆}})P(B|A^{\text{逆}})}$$

二分类问题

举例：经典的垃圾邮件过滤问题

一封垃圾邮件 X ，由 d 个单词组成，分别为 x_1, x_2, \dots, x_d (d 个特征)
 $y=1$ 表示是垃圾邮件， $y=0$ 表示不是垃圾邮件，则预测过程为

$$P(y = 1|X) = \frac{P(y = 1)P(X|y = 1)}{P(X)}$$

$$P(y = 0|X) = \frac{P(y = 0)P(X|y = 0)}{P(X)}$$

谁大， y 就取谁

二分类问题

举例：经典的垃圾邮件过滤问题

$P(y = 1)$ 可以直接统计某样本在训练数据集中的分布

$P(X | y=1)$ 比较麻烦，它表示一封邮件是垃圾邮件的情况下，它会包含 X 中的那些词(特征)的概率是多少

$$P(X|y=1)=P(x_1,x_2,...,x_d|y=1)$$

如何求？ 假设各个词之间具备条件独立性

$$P(x_1,x_2,...,x_d|y=1)=P(x_1|y=1)*P(x_2|y=1)*P(x_3|y=1).....$$

拓展到多分类问题

拓展到多分类问题： 比如一共 c_1, c_2, \dots, c_k 类

某一类的后验概率：

$$P(y = c_k | X) = \frac{P(y = c_k)P(X = x | y = c_k)}{P(X = x)}$$

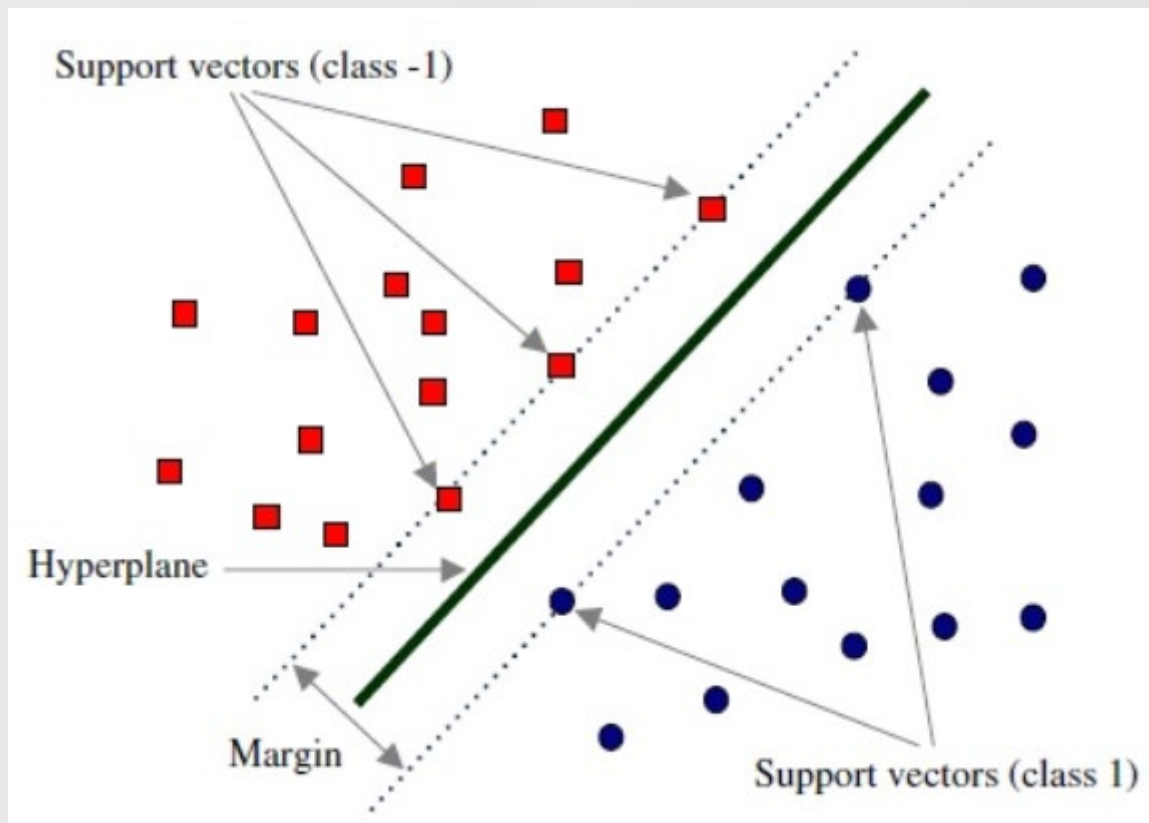
比较每一类的后验概率，取后验概率最大的那一类作为输出的标签值。

The background features a complex network of thin grey lines connecting various-sized dark grey circular nodes. These nodes are scattered across the slide, with a higher concentration on the right side, creating a web-like or molecular structure. The overall aesthetic is clean and technical.

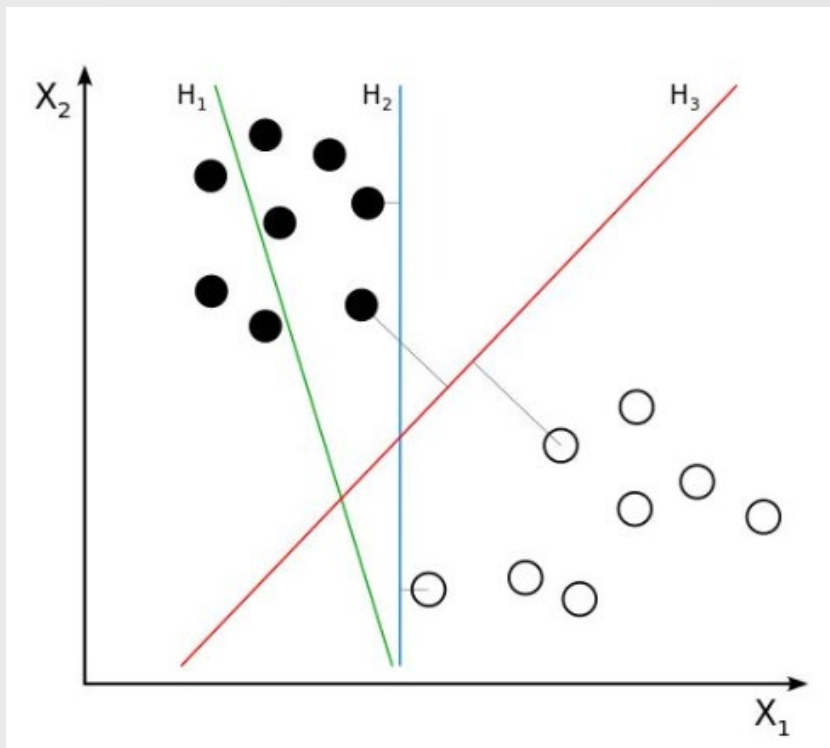
06

支持向量机

支持向量机



直观理解



图中有分别属于两类的一些二维数据点和三条直线。如果三条直线分别代表三个分类器的话，请问哪一个分类器比较好？

线性可分SVM

考虑以下形式的线性可分的训练数据集：

$$(X_1, y_1), (X_2, y_2), \dots, (X_n, y_n)$$

其中 X_i 是一个含有 d 个元素的列向量，即 $X_i \in \mathbf{R}^d$ ， y_i 是标量

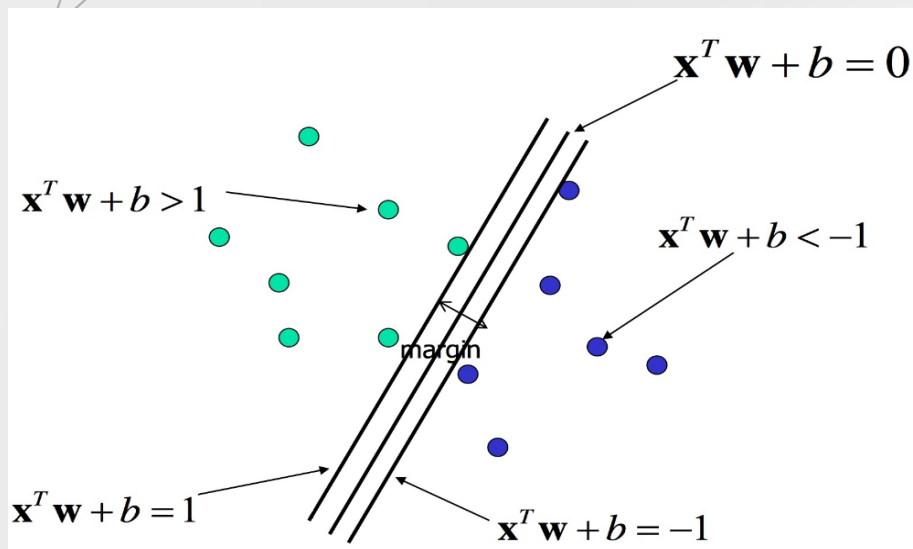
$y_i \in +1, -1, y_i = +1$ ， $+1$ 表示 X_i 属于正类别， -1 表示属于负类别

目标：找到一个超平面使其能够正确地将每个样本正确分类

超平面与间隔

一个超平面由其法向量 \mathbf{W} 和截距 b 决定，其方程为 $\mathbf{x}^T \mathbf{W} + b = 0$ ，可以规定法向量指向的一侧为正类，另一侧为反类。

左图画出了三个平行的超平面，法方向取左上方向



为了找到最大间隔超平面，我们可以先选择分离两类数据的两个平行超平面，使得它们之间的距离尽可能大。在这两个超平面范围内的区域称为“间隔(margin)”，最大间隔超平面是位于它们正中间的超平面。

间隔最大化

间隔

$$\text{margin} = \rho = \frac{2}{\|W\|}$$

目标是使得 ρ 最大，等价于使得 ρ^2 最大

$$\max_{W,b} \rho \iff \max_{W,b} \rho^2 \iff \min_{W,b} \frac{1}{2} \|W\|^2$$

上式的 $1/2$ 上是为了后续求导时刚好可以消去，没有其他的特殊意义

间隔最大化

再加一些约束条件

$$\begin{aligned} X_i^T W + b &\geq +1, y_i = +1 \\ X_i^T W + b &\leq -1, y_i = -1 \end{aligned}$$

总结一下，间隔最大化问题的数学表达就是

$$\begin{aligned} \min_{W, b} J(W) &= \min_{W, b} \frac{1}{2} \|W\|^2 \\ \text{s.t.} \quad &y_i (X_i^T W + b) \geq 1, i = 1, 2, \dots, n. \end{aligned}$$

通过求解上式即可得到最优超平面参数




支持向量

在线性可分的情况下，训练数据集的样本点中与分离超平面举例最近的数据点称为支持向量，满足

$$y_i(X_i^T W + b) = 1$$

也就是所有在直线 $X^T W + b = 1$ 或直线 $X^T W + b = -1$ 的点

在决定最佳超平面时，只有支持向量起作用，而其他数据点并不起作用

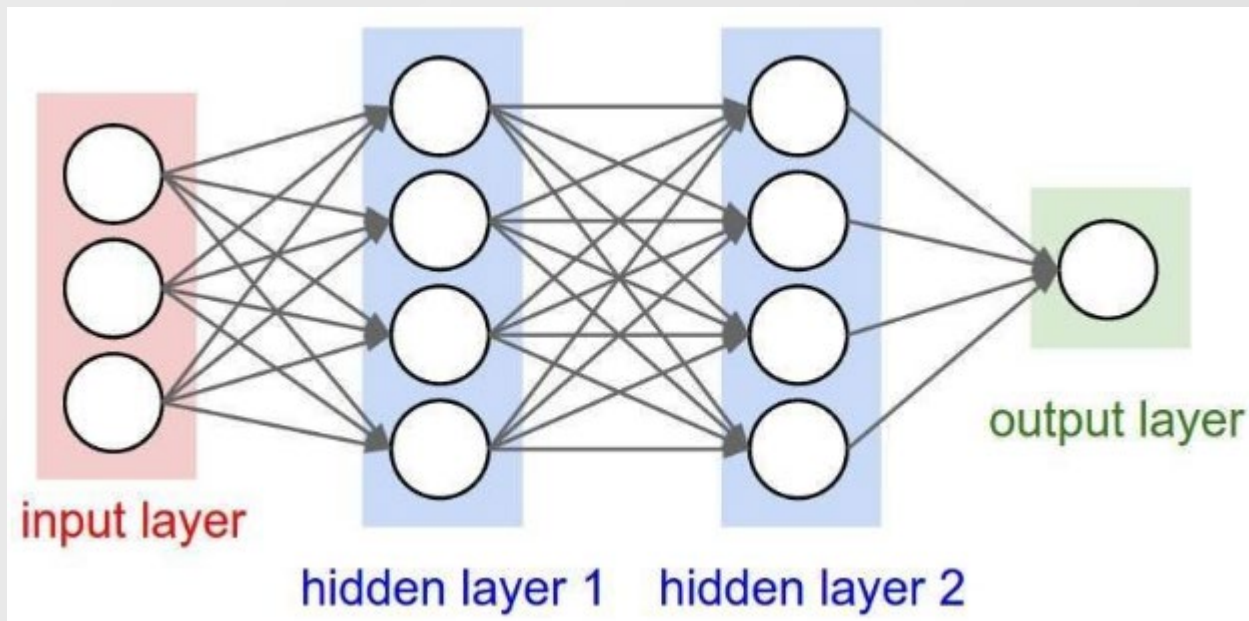




07

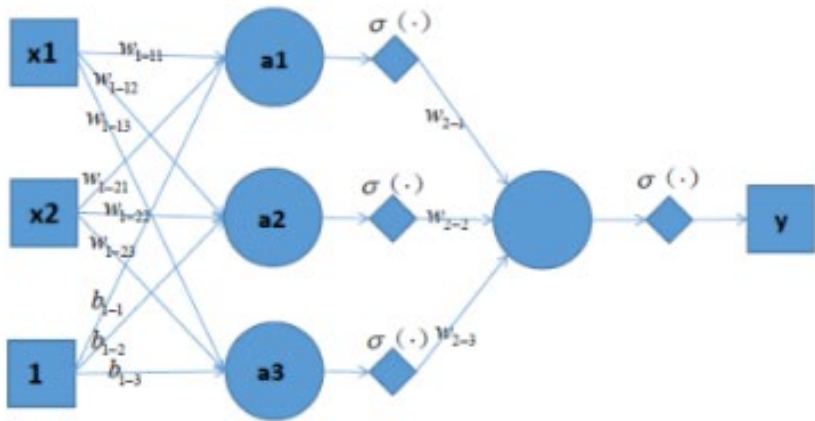
神经网络预测

神经网络预测



神经网络预测

Setup: 数据特征 $X = (X_1, X_2, X_n)$, 标签 $Y = (y_1, y_2, y_n)$



$$a1 = w_{1-11}x_1 + w_{1-21}x_2 + b_{1-1}$$

$$a2 = w_{1-12}x_1 + w_{1-22}x_2 + b_{1-2}$$

$$a3 = w_{1-13}x_1 + w_{1-23}x_2 + b_{1-3}$$

$$y = \sigma(w_{2-1}\sigma(a1) + w_{2-2}\sigma(a2) + w_{2-3}\sigma(a3))$$

神经网络预测

多层神经网络

$$H_0 = X = \{x_1, x_2, x_3, x_4\}$$

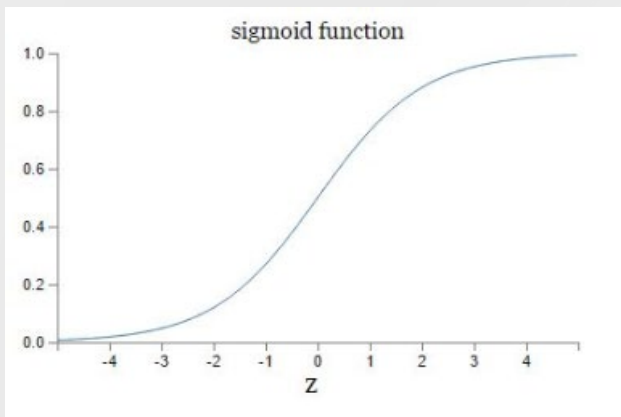
$$H_i = \sigma(W_i H_{i-1} + b_i)$$

$$O = \sigma(W_l H_{l-1} + b_l)$$

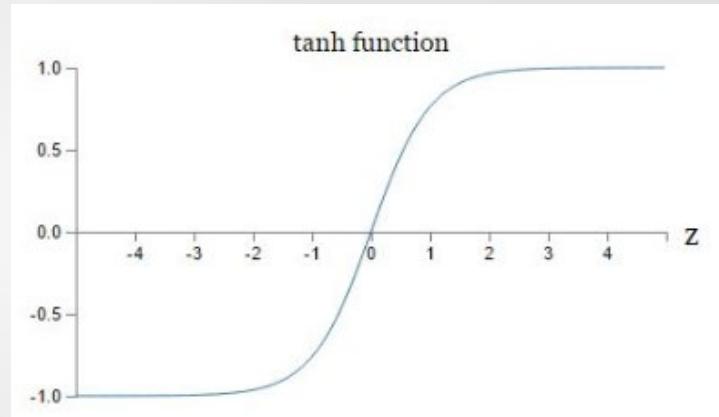
W, b 均为待求参数， σ 为激活函数

激活函数

Sigmoid



tanh



Outline

1. 时间序列预测
2. 微分方程与差分方程
3. 灰色预测模型
4. 机器学习方法
 1. 线性回归与逻辑回归
 2. 支持向量机
 3. 朴素贝叶斯
 4. 神经网络





谢谢

CREDITS: This presentation template was created by [Slidesgo](#) including icons by [Flaticon](#) and infographics & images by [Freepik](#)

Please keep this slide for attribution.