统计机器学习
Statistical Machine Learning

魏莱

上海海事大学信息工程学院

2019 年 3 月 25 日

第二章：线性回归及分类

1. 线性模型的基本形式

假设样本 $\mathbf{x} = (x_1, x_2, \cdots, x_d)^t \in R^{d\times 1}$，$x_j$ 是 $\mathbf{x}$ 第 j 个属性，线性模型试图学得一个通过属性的线性组合来进行预测的函数，即

$$\begin{aligned} f(\mathbf{x}) &= w_0 + w_1 x_1 + \cdots + w_d x_d \\ &= \mathbf{w}^t \hat{\mathbf{x}} \end{aligned} \tag{1}$$

其中 $\mathbf{w} = (w_0, w_1, \cdots, w_d)^t$，
$\hat{\mathbf{x}} = (1, x_1, x_2, \cdots, x_d)^t = (x_0, x_1, x_2, \cdots, x_d)^t$。式 (1) 第二行称为线性回归模型的向量形式。

可见，$\mathbf{w}$ 唯一决定了一个线性模型 $f(\mathbf{x}) = \mathbf{w}^t\hat{\mathbf{x}}$ 的形式。$\mathbf{w}$ 中的每个分量 w_j 直观表达了样本 $\mathbf{x}$ 中第 j 个属性 x_j 的重要性。

例 2：搜集到房屋价格和面积、卧室之间的数据如下表：

Living area (feet^2)	#bedrooms	Price (1000$s)
2104	3	400
1600	3	330
2400	3	369
1416	2	232
3000	4	540
⋮	⋮	⋮

对于新的一个房屋面积为 3402，卧室数量为 3，则其价格因为多少？

给定数据集 $\mathbf{D} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \cdots, (\mathbf{x}_n, y_n)\}$，其中 $\mathbf{x}_i = (x_{i1}, x_{i2}, \cdots, x_{id})^t$，$y_i \in R$。**线性回归 (linear regression) 试图学得一个线性模型以尽可能准确预测新样本 $\mathbf{x}_*$ 的实值输出 y_*。**

由于采用线性模型 $f(\mathbf{x}) = \mathbf{w}^t\mathbf{x}$ 预测，一旦确定 $\mathbf{w}$，那么对于 $\mathbf{x}_*$ 可以计算出 $f(\mathbf{x}_*)$，$f(\mathbf{x}_*)$ 可以视为 y_* 的估计。于是问题转变成，如何在给定数据集下，计算出 $\mathbf{w}$?

定义平方误差函数

$$E(\mathbf{w}) = \frac{1}{2}\sum_{i=1}^{n}(f(\mathbf{x}_i) - y_i)^2 = \frac{1}{2}\sum_{i=1}^{n}(\mathbf{w}^t\mathbf{x}_i - y_i)^2 \tag{2}$$

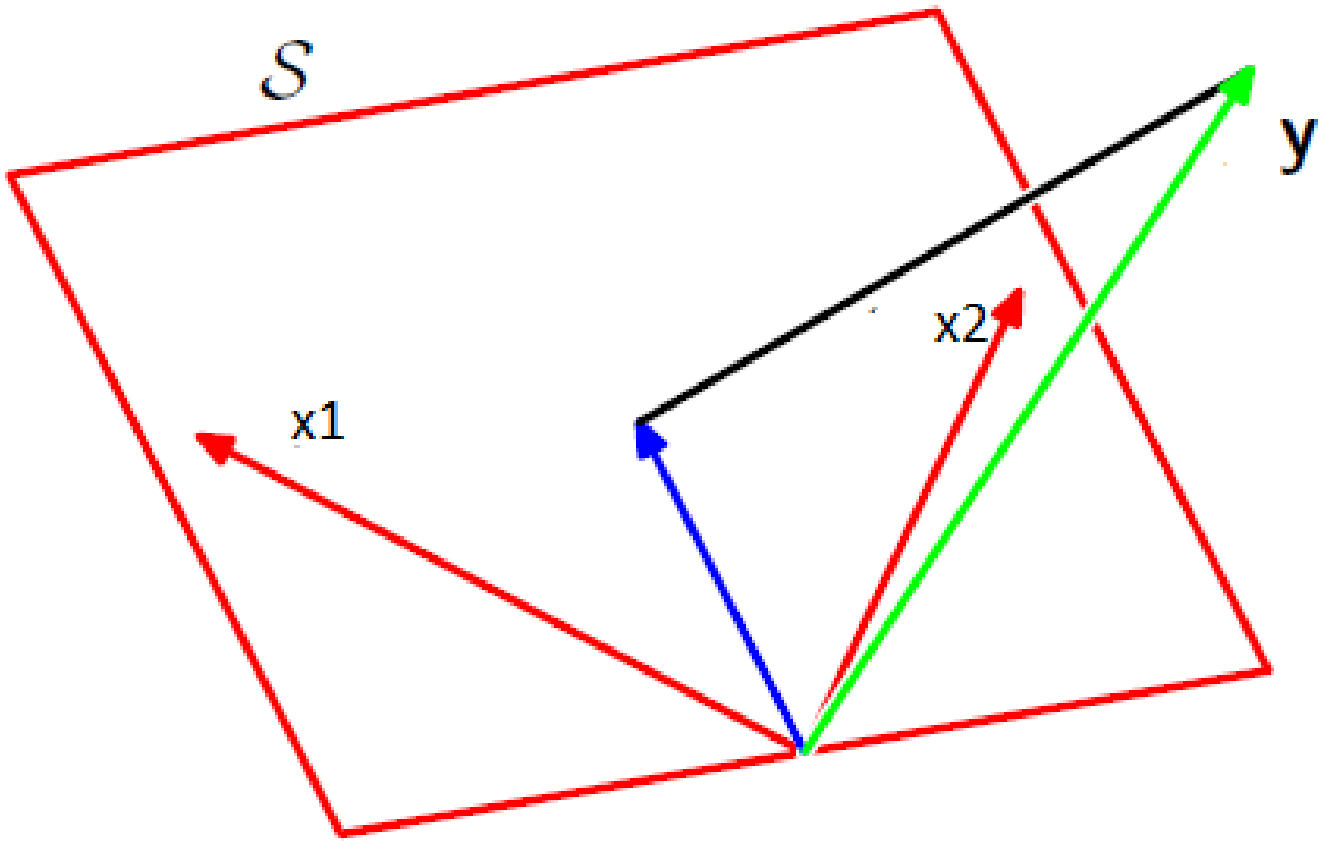
可得 $\mathbf{w}_{opt} = \arg\min_{\mathbf{w}} E(\mathbf{w})$。如何求解 $\mathbf{w}$?

- 方法一：最小二乘法 (least squares)。将 $E(\mathbf{w})$ 对 $\mathbf{w}$ 求导，然后迫使导数等于 0，通过等式求得 $\mathbf{w}$。为此，令 $\mathbf{y} = (y_1, y_2, \cdots, y_n)^t$，$\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n)$ 则

$$\begin{aligned} E(\mathbf{w}) &= \frac{1}{2}\sum_{i=1}^{n}(f(\mathbf{x}_i) - y_i)^2 \\ &= \frac{1}{2}(\mathbf{w}^t\mathbf{X} - \mathbf{y}^t)(\mathbf{w}^t\mathbf{X} - \mathbf{y}^t)^t \\ &= \frac{1}{2}(\mathbf{w}^t\mathbf{X}\mathbf{X}^t\mathbf{w} - \mathbf{w}^t\mathbf{X}\mathbf{y} - \mathbf{y}^t\mathbf{X}^t\mathbf{w} + \mathbf{y}^t\mathbf{y}) \end{aligned} \tag{3}$$

因此，$\partial E(\mathbf{w})/\partial \mathbf{w} = \mathbf{X}\mathbf{X}^t\mathbf{w} - \mathbf{X}\mathbf{y} = 0$，故 $\mathbf{w}_{opt} = (\mathbf{X}\mathbf{X})^{-1}\mathbf{X}\mathbf{y}$。

- 最小二乘法的几何解释。



- 方法二：梯度下降法 (gradient descent)。将 $E(\mathbf{w})$ 对每个 w_j 求导，然后令

$$w_j \doteq w_j - \alpha \partial E(\mathbf{w})/\partial w_j$$

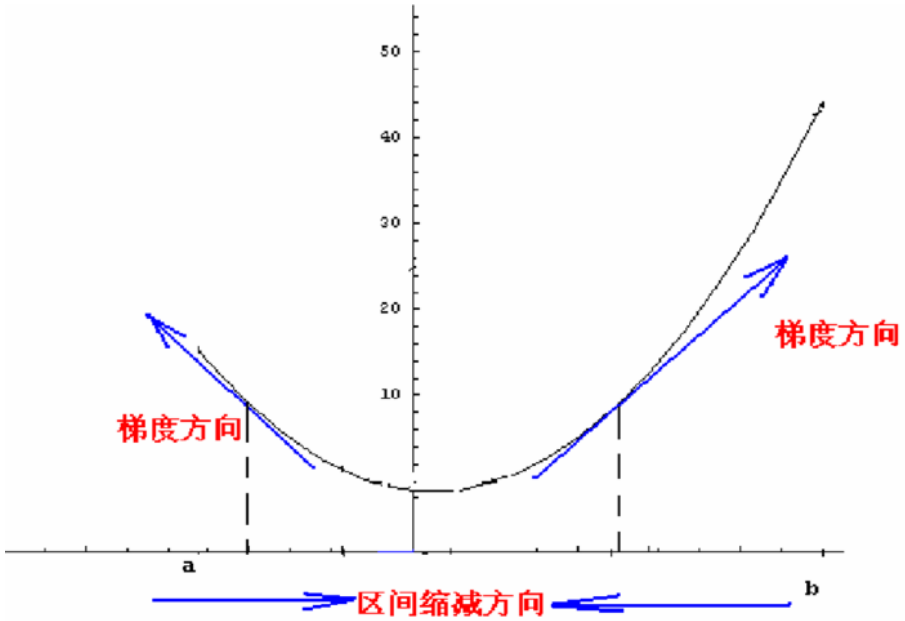
其中 α 为**学习率 (learning rate)**。$\mathbf{y} = (y_1, y_2, \cdots, y_n)^t$，$\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n)$ 则

$$\begin{aligned} \partial E(\mathbf{w})/\partial w_j &= \tfrac{1}{2}\textstyle\sum_{i=1}^{n}(\mathbf{w}^t\mathbf{x}_i - y_i)^2/\partial w_j \\ &= \textstyle\sum_{i=1}^{n}(\mathbf{w}^t\mathbf{x}_i - y_i)\partial(\mathbf{w}^t\mathbf{x}_i - y_i)/\partial w_j \\ &= \textstyle\sum_{i=1}^{n}(\mathbf{w}^t\mathbf{x}_i - y_i)\partial(\sum_{j=1}^{d} w_j x_{ij} - y_i)/\partial w_j \\ &= \textstyle\sum_{i=1}^{n}(\mathbf{w}^t\mathbf{x}_i - y_i)x_{ij} \end{aligned} \tag{4}$$

因此，梯度下降法更新规则可以表示成

$$w_j \doteq w_j - \alpha \sum_{i=1}^{n} (\mathbf{w}^t \mathbf{x}_i - y_i) x_{ij} \tag{5}$$

- 梯度下降法的几何解释。



- 随机梯度下降 (stochastic gradient descent 或 incremental gradient descent)

  观察，上文梯度下降法更新公式，

  $w_j \doteq w_j - \alpha \sum_{i=1}^{n} (\mathbf{w}^t \mathbf{x}_i - y_i) x_{ij}$，每次更新需要利用到全部样本，所以称为批量梯度下降法 (batch gradient descent)。

  随机梯度下降法更新规则：

  for $j = 1$ to d

  {

  $$w_j \doteq w_j - \alpha (\mathbf{w}^t \mathbf{x}_i - y_i) x_{ij} \quad (for \quad every \quad i) \tag{6}$$

  }

3. 线性回归的扩展：正规最小均方误差回归

在上一节课模型选择过程中，我们介绍了正则化的概念，即在误差函数 $E(\mathbf{w})$ 中加入对 $\mathbf{w}$ 的约束项，可得

$$E(\mathbf{w}) = \frac{1}{2}\sum_{i=1}^{n}(\mathbf{w}^t\mathbf{x}_i - y_i)^2 + \frac{\lambda}{2}\|\mathbf{w}\|_2^2 \tag{7}$$

其中 $\|\mathbf{w}\|_2^2 = \mathbf{w}^t\mathbf{w}$。采用最小二乘法来求解上述问题，可得：

$E(\mathbf{w}) = \frac{1}{2}(\mathbf{w}^t\mathbf{X} - \mathbf{y}^t)(\mathbf{w}^t\mathbf{X} - \mathbf{y}^t)^t + \frac{\lambda}{2}\mathbf{w}^t\mathbf{w}$，因此 $\mathbf{w}_{opt} = (\mathbf{X}\mathbf{X}^t + \lambda\mathbf{I})^{-1}\mathbf{X}\mathbf{y}$。（Exercise!!）

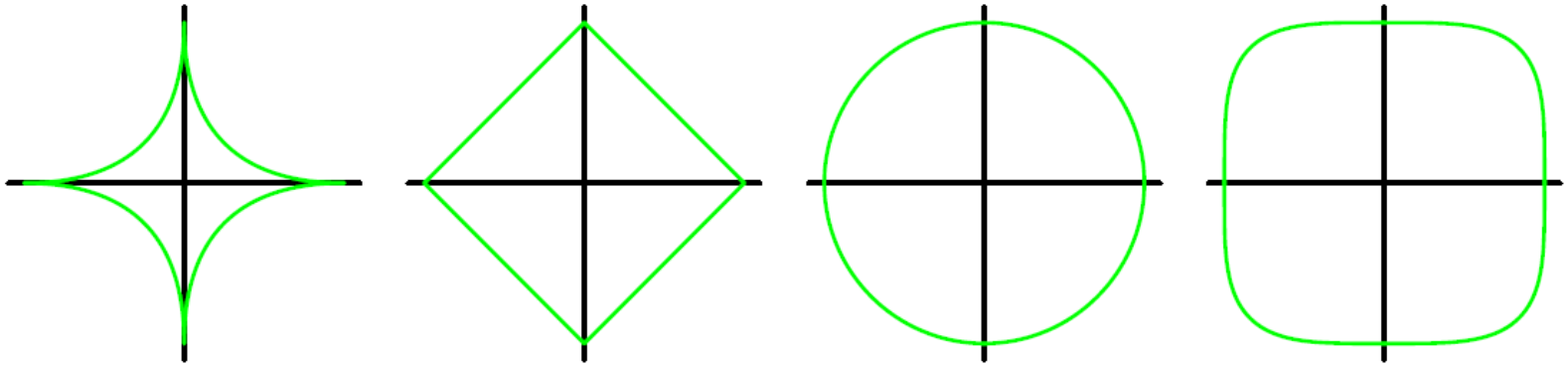
正规最小均方误差回归的泛化形式

$$E(\mathbf{w}) = \frac{1}{2}\sum_{i=1}^{n}(\mathbf{w}^t\mathbf{x}_i - y_i)^2 + \frac{\lambda}{2}\sum_{j=1}^{d}|w_j|^q \tag{8}$$

其中，$q = 2$。

因此，可否改变 q 的值来泛化上式？常用的取值包括 $q = 0.5, 1$ 等等。

假设 $\sum_{j=1}^{d} |w_j|^q = 1$，则选择不同的 q，$\mathbf{w}$ 的定义域有什么不同？

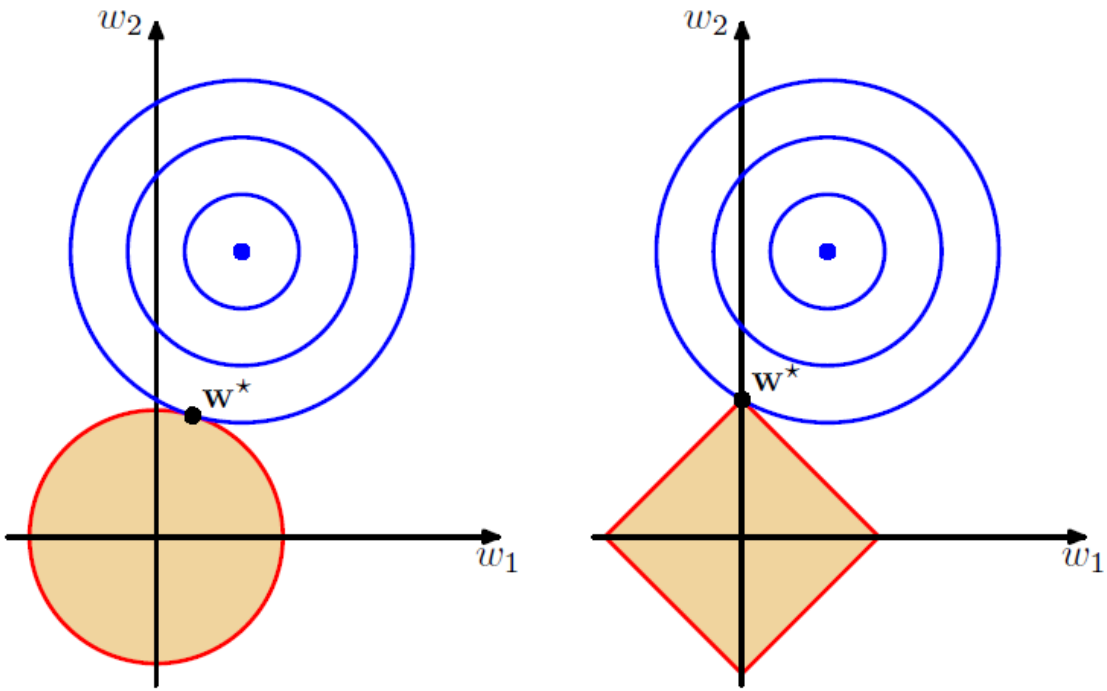


分析式 (9), 即：

$$E(\mathbf{w}) = \frac{1}{2}\sum_{i=1}^{n}(\mathbf{w}^t\mathbf{x}_i - y_i)^2 + \frac{\lambda}{2}\sum_{j=1}^{d}|w_j|^q$$

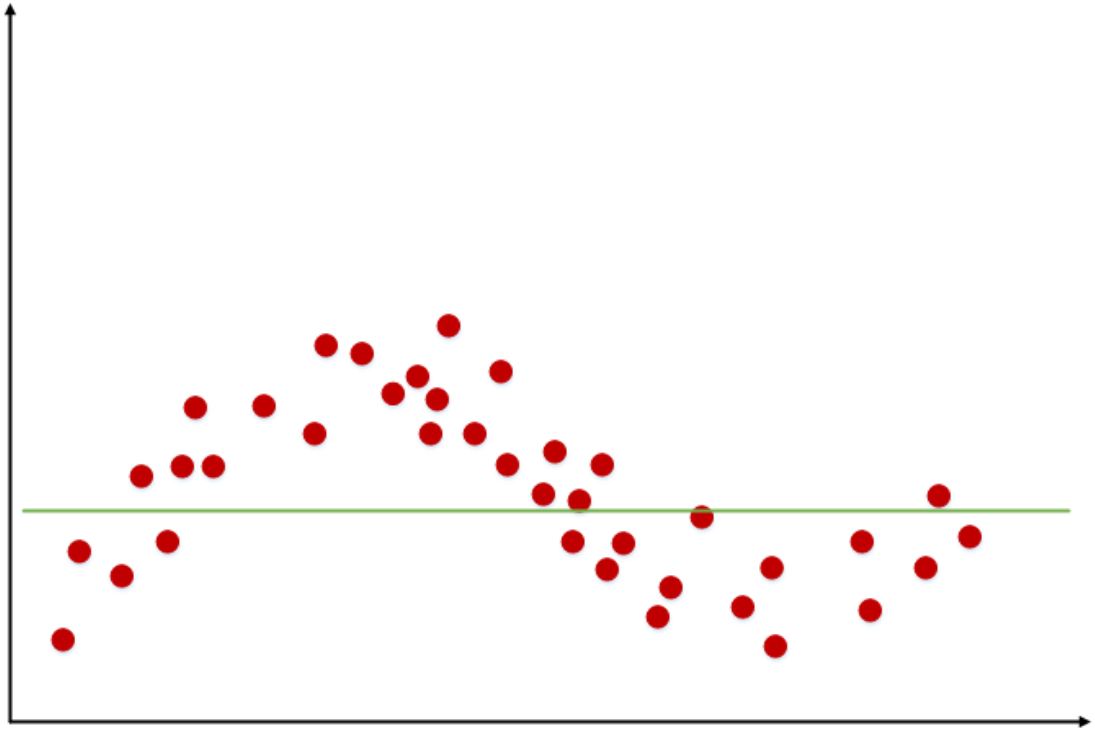
可以发现最小化上述问题，可以转变成在限定 $\sum_{j=1}^{d}|w_j|^q$ 小于某一个常数项的情况下，最小化 $\frac{1}{2}\sum_{i=1}^{n}(\mathbf{w}^t\mathbf{x}_i - y_i)^2$。假设现在限定 $\sum_{j=1}^{d}|w_j|^q \leq 1$，则 $\mathbf{w}$ 的最优解有什么不同？

3. 线性回归的扩展：正规最小均方误差回归



可见，当 $q=1$ 时，$\mathbf{w}_{opt}$ 中 $w_1=0$，即只有一个非零项。因此其要比 $q=2$ 时，求得的解更为稀疏。事实上，当 $q=1$ 时，正规最小均方误差回归也叫 **Least absolute shrinkage and selection operator(LASSO)**。

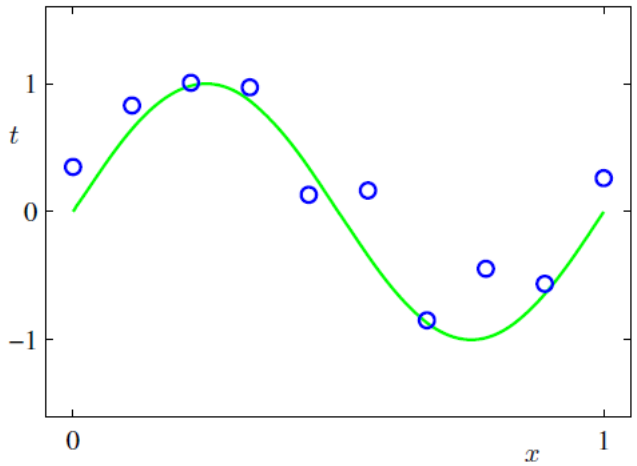
3. 线性回归的扩展：局部带权线性回归



对于线性回归来说，一旦得到 $\mathbf{w}$，那么对于新的数据 $\mathbf{x}_*$，则其预测值 $y_* = \mathbf{w}^t\mathbf{x}_*$。

3. 线性回归的扩展：局部带权线性回归

回顾上一节课的多项式拟合问题，

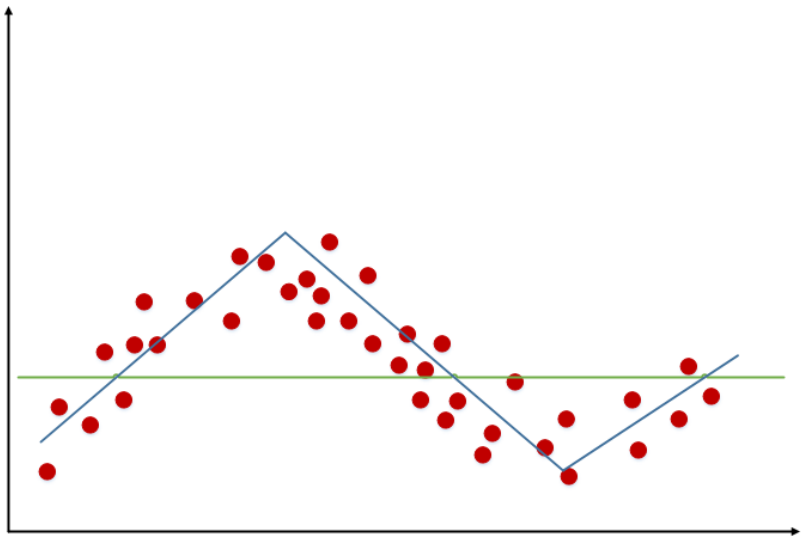


$$
\begin{aligned}
f(x, \mathbf{w}) &= w_0 + w_1 x^1 + w_2 x^2 + \cdots + w_M x^M = \sum_{j=1}^{M} w_j x^j \\
&= \overline{\mathbf{w}}^t \overline{\mathbf{x}}
\end{aligned} \tag{9}
$$

这里，$\overline{\mathbf{w}} = (w_0, w_1, \cdots, w_M)^t$，$\overline{\mathbf{x}} = (1, x, x^2, \cdots, x^M)^t$。由此可见，**在低维空间中的一个非线性问题，可以在高维空间中转变成线性问题。**

3. 线性回归的扩展：局部带权线性回归

对非线性数据拟合问题的另一种解法：



可见图中的蓝色折线段也能较好的拟合非线性数据集。对于线性回归来说，误差函数 $E(\mathbf{w}) = \frac{1}{2}\sum_{i=1}^{n}(\mathbf{w}^t\mathbf{x}_i - y_i)^2$，局部带权线性回归误差定义为：$E(\mathbf{w}) = \frac{1}{2}\sum_{i=1}^{n}\gamma_i(\mathbf{w}^t\mathbf{x}_i - y_i)^2$

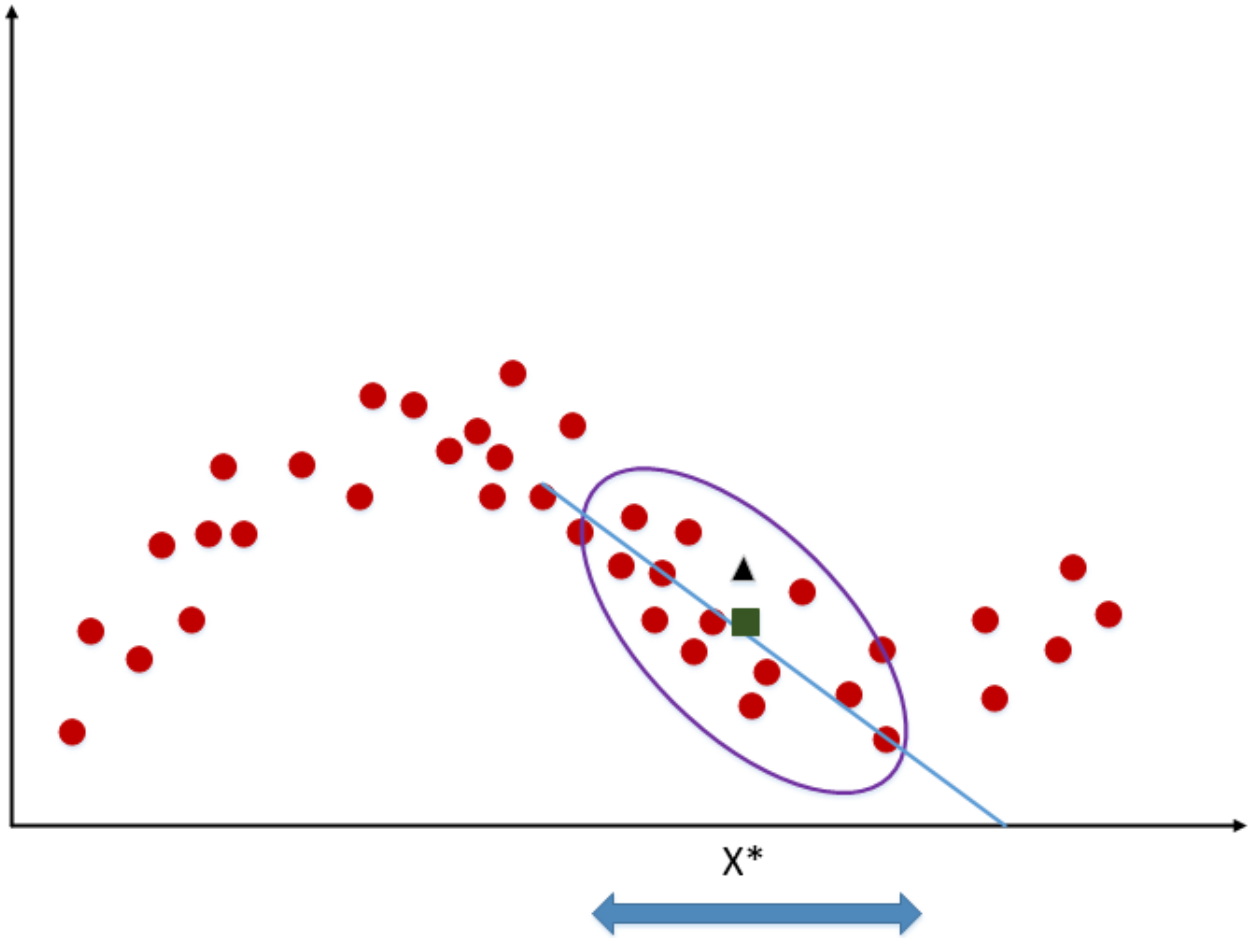
上式中 γ_i 称为权重，取值为非负的实数值。可以看到，如果 γ_i 较大，由于要求最小化 $E(\mathbf{w})$，则 $(\mathbf{w}^t\mathbf{x}_i - y_i)^2$ 就要求小。而如果 γ_i 较小，$(\mathbf{w}^t\mathbf{x}_i - y_i)^2$ 就会被忽略掉。

合理的 γ_i 的定义可以有如下形式:

$$\gamma_i = \exp\left(-\frac{(\mathbf{x}_* - \mathbf{x}_i)^2}{2\tau^2}\right) \tag{10}$$

这里，τ 是个非负参数，称为带宽 (bandwidth)。这样，假设 $|\mathbf{x}_* - \mathbf{x}_i|$ 非常小，则 γ_i 接近于 1，否则，γ_i 接近于 0.

3. 线性回归的扩展：局部带权线性回归



局部带权回归是一种非参数方法。

3. 线性回归的扩展：对数几率回归

前面所讨论的都是利用线性模型进行回归学习，现在考虑利用线性模型进行分类学习。

线性模型 $f(\mathbf{x}) = \mathbf{w}^t\mathbf{x}$，如果其输出 $f(\mathbf{x})$ 的取值为 0 或 1，那么就可以用来进行分类任务了。而前面所述，线性回归模型输出的是实值连续量，因此如果能够设计一个由实值连续量到 0/1 的转换函数，那么问题就解决了。

3. 线性回归的扩展：对数几率回归

单位阶跃函数 (unit-step function):

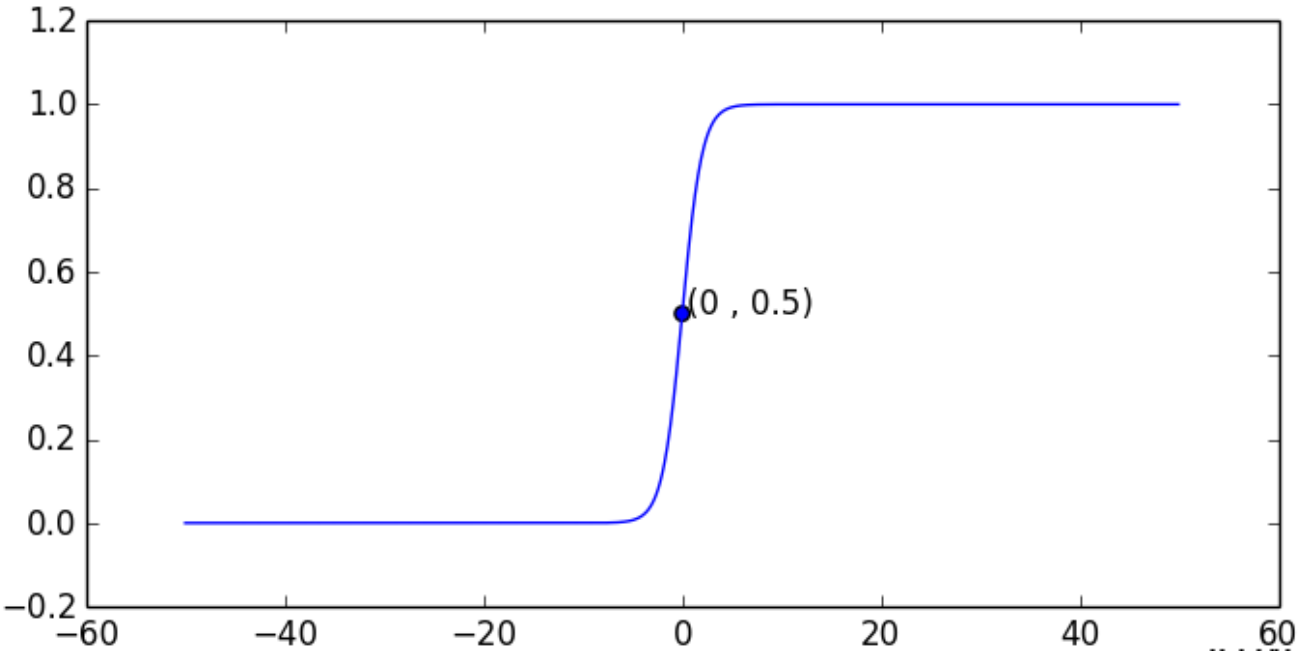
$$y = \begin{cases} 0, & \text{z<0;} \\ 0.5, & \text{z=0;} \\ 1, & \text{z>0.} \end{cases} \tag{11}$$

单位阶跃函数不连续，不利用数学处理，因此设计对数几率函数 (logistic function):

$$y = \frac{1}{1 + e^{-z}} = \frac{1}{1 + e^{-\mathbf{w}^t \mathbf{x}}} \tag{12}$$

3. 线性回归的扩展：对数几率回归

对数几率函数示意图：



由式（12）可得，$\ln \frac{y}{1-y} = \mathbf{w}^t\mathbf{x}$，其中 y 可以视为样本 $\mathbf{x}$ 为正例的比例，$1-y$ 视为样本 $\mathbf{x}$ 为反例的比例，两者比值

$$\frac{y}{1-y} \tag{13}$$

称为几率，

$$\ln \frac{y}{1-y} \tag{14}$$

称为对数几率。

Thanks !