第二章：线性模型

2.1 线性回归

2.2 梯度更新方式

2.3 线性回归矩阵形式

2.4 最大似然估计

2.5 分类指标

2.6 逻辑斯蒂回归

西安交通大学

XI'AN JIAOTONG UNIVERSITY

2.1 线性回归

2.1 线性回归

线性判别模型

判别模型

- 性质
  - 建模预测变量和观测变量之间的关系
  - 也称作条件模型 (Conditional Models)
- 分类
  - 确定性判别模型：$y = f_\theta(x)$
  - 概率判别模型：$p_\theta(y|x)$

本节集中介绍线性判别模型 (linear regression)

2.1 线性回归

线性判别模型

判别模型

- 性质
  - 建模预测变量和观测变量之间的关系
  - 也称作条件模型 (Conditional Models)
- 分类
  - 确定性判别模型：$y = f_\theta(x)$
  - 概率判别模型：$p_\theta(y|x)$

线性判别模型 (linear regression)

$$y = f_\theta(x) = \theta_0 + \sum_{j=1}^{d} \theta_j x_j = \theta^\top x$$

$$x = (1, x_1, x_2, \cdots, x_d)$$

2.1 线性回归

学习目标

- 使预测值和真实值的距离越近越好

$$\min_{\theta} \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(y_i, f_\theta(x_i))$$

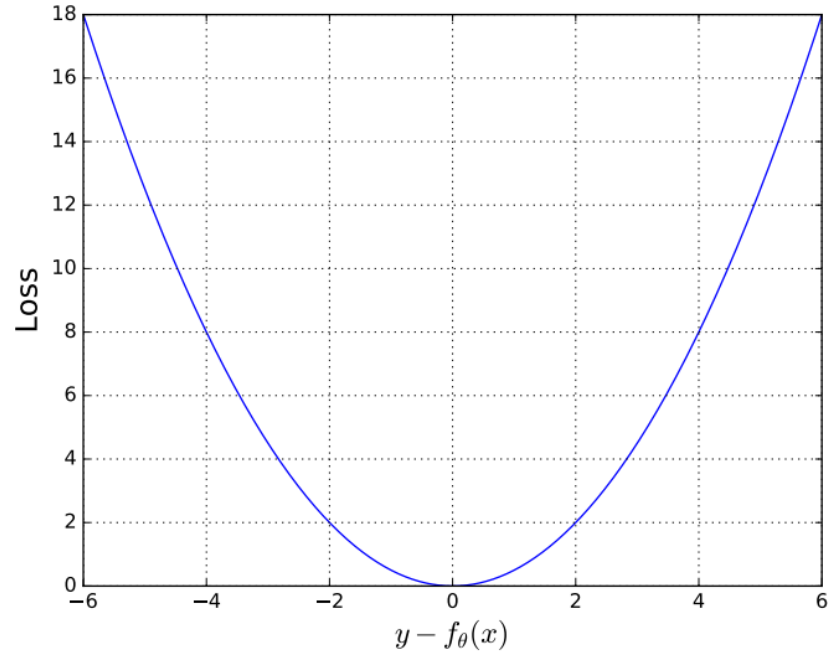
- 损失函数$\mathcal{L}(y_i, f_\theta(x_i))$测量预测值和真实值之间的误差，越小越好
- 具体损失函数的定义依赖于具体的数据和任务
- 最广泛使用的损失回归函数：平方误差(squared loss)

$$\mathcal{L}\big(y_i, f_\theta(x_i)\big) = \frac{1}{2}\big(y_i - f_\theta(x_i)\big)^2$$

2.1 线性回归

平方误差

$$\mathcal{L}\left(y_i, f_\theta(x_i)\right) = \frac{1}{2}\left(y_i - f_\theta(x_i)\right)^2$$



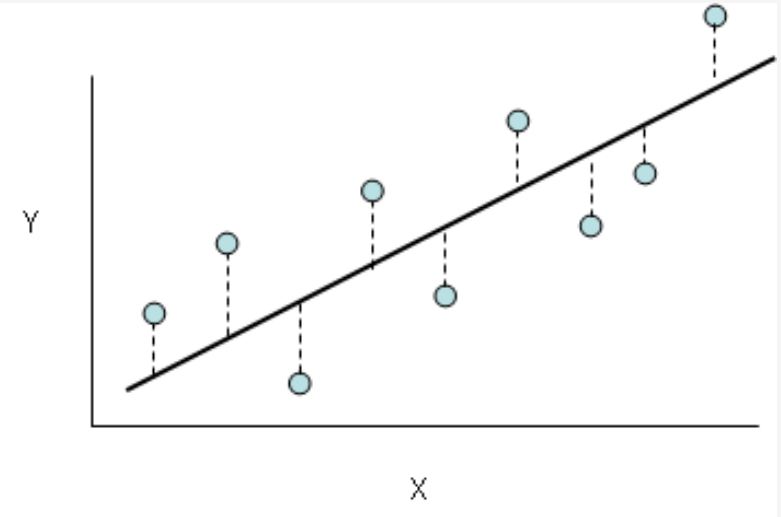
- 对预测误差大的有更大的惩罚
- 容忍很小的预测误差
  - 观测误差等
  - 提升模型的泛化能力

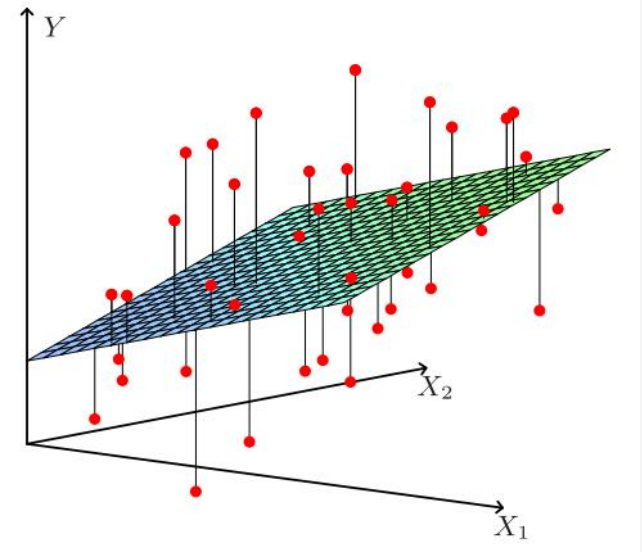
2.1 线性回归

最小均方误差回归

- 优化目标是最小化训练数据上的均方误差

$$J_\theta = \frac{1}{2N}\sum_{i=1}^{N}\left(y_i - f_\theta(x_i)\right)^2 \qquad \min_\theta J_\theta$$





2.2 梯度更新方式

批量梯度下降

- 优化目标

$$J(\theta)=\frac{1}{2N}\sum_{i=1}^{N}\left(y_i-f_\theta(x_i)\right)^2 \qquad \min_\theta J(\theta)$$

- 根据整个批量数据的梯度更新参数 $\theta_{\text{new}} \leftarrow \theta_{\text{old}}-\eta\frac{\partial J(\theta)}{\partial\theta}$

$$\frac{\partial J(\theta)}{\partial\theta}=-\frac{1}{N}\sum_{i=1}^{N}\left((y_i-f_\theta(x_i))\frac{\partial f_\theta(x_i)}{\partial\theta}\right)$$

$$=-\frac{1}{N}\sum_{i=1}^{N}(y_i-f_\theta(x_i))\,x_i$$

$$\theta_{\text{new}}=\theta_{\text{old}}+\eta\frac{1}{N}\sum_{i=1}^{N}\left(y_i-f_\theta(x_i)\right)x_i$$

2.2 梯度更新方式

随机梯度下降

- 优化目标

$$J^{(i)}(\theta) = \frac{1}{2}\left(y_i - f_\theta(x_i)\right)^2 \qquad \min_\theta \frac{1}{N}\sum_i J^{(i)}(\theta)$$

- 根据整个批量数据的梯度更新参数 $\theta_{\text{new}} = \theta_{\text{old}} - \eta \frac{\partial J^{(i)}(\theta)}{\partial \theta}$

$$\frac{\partial J^{(i)}(\theta)}{\partial \theta} = -\left(y_i - f_\theta(x_i)\right)\frac{\partial f_\theta(x_i)}{\partial \theta}$$
$$= -\left(y_i - f_\theta(x_i)\right)x_i$$
$$\theta_{\text{new}} = \theta_{\text{old}} + \eta\left(y_i - f_\theta(x_i)\right)x_i$$

- 对比批量梯度下降
  - 更快地更新参数(优点)
  - 学习中不确定性或震荡(缺点)

2.2 梯度更新方式

小批量梯度下降

算法思想

批量梯度下降和随机梯度下降的结合

训练步骤

- 将整个训练集分成 K个小批量（mini-batches）

$$\{1,2,3,\cdots,K\}$$

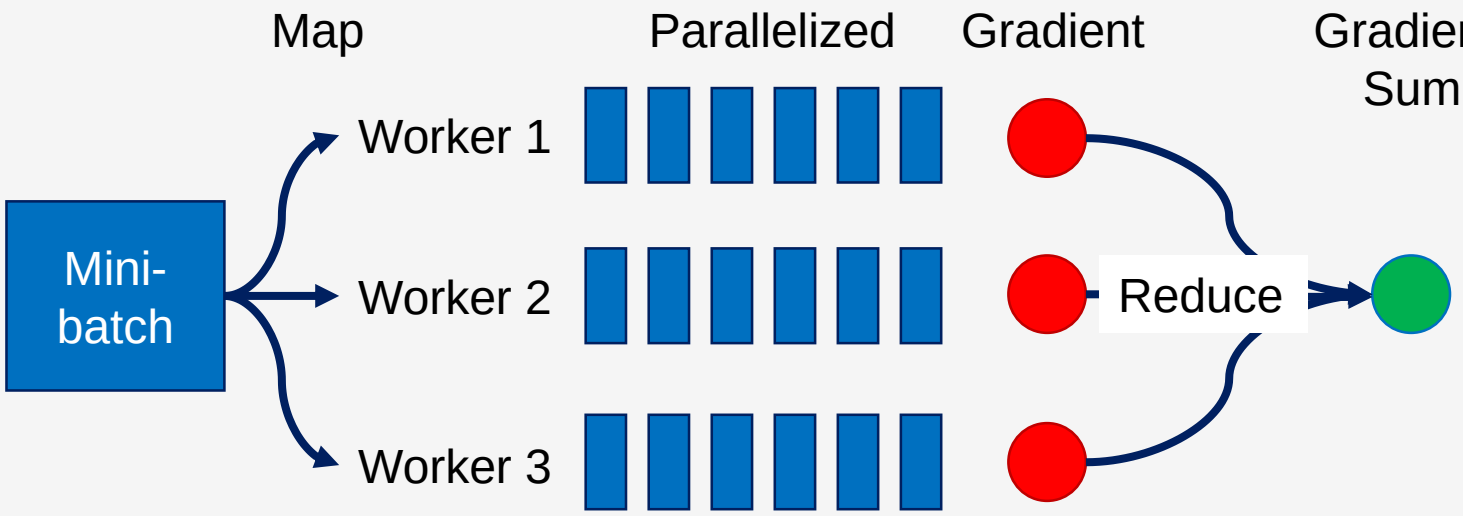
- 对于每一个小批量 k, 做一步批量下降来降低

$$J^{(k)}(\theta)=\frac{1}{2N_k}\sum_{i=1}^{N_k}\left(y_i-f_\theta(x_i)\right)^2$$

- 对于每一个小批量，更新参数 $\theta_{\text{new}} = \theta_{\text{old}}-\eta\frac{\partial J^{(k)}(\theta)}{\partial\theta}$

2.2 梯度更新方式

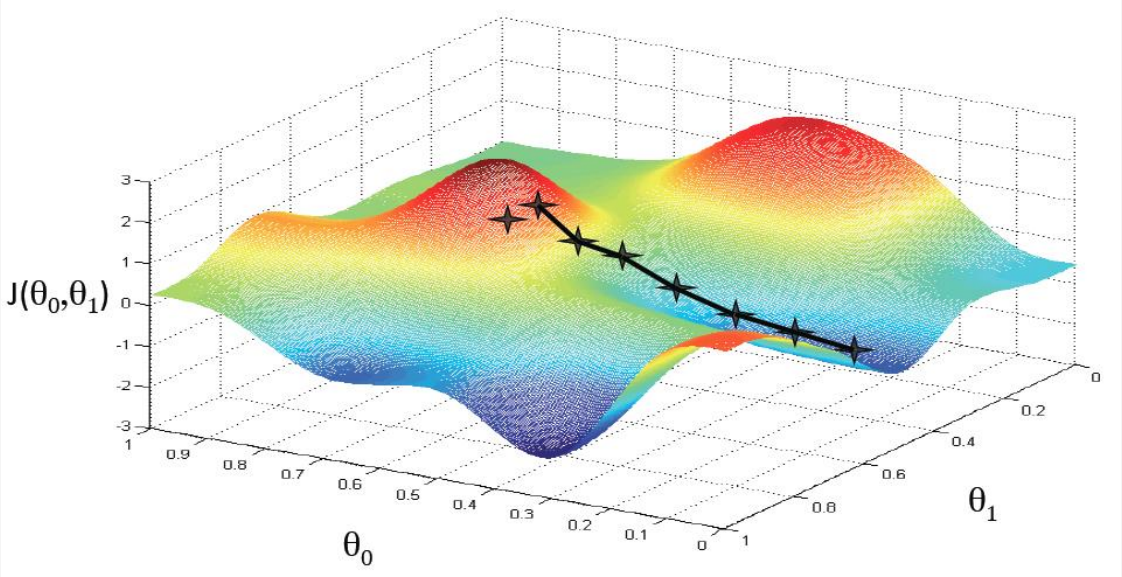
- 结合了批量梯度下降和随机梯度下降的优点
  - 批量梯度下降的优秀的稳定性
  - 随机梯度下降的快速更新
- 小批量梯度下降很适合使用在并行化计算中
  - 将每个小批量数据进一步切分，每个线程分别计算梯度，最后再加和这些梯度



2.2 梯度更新方式

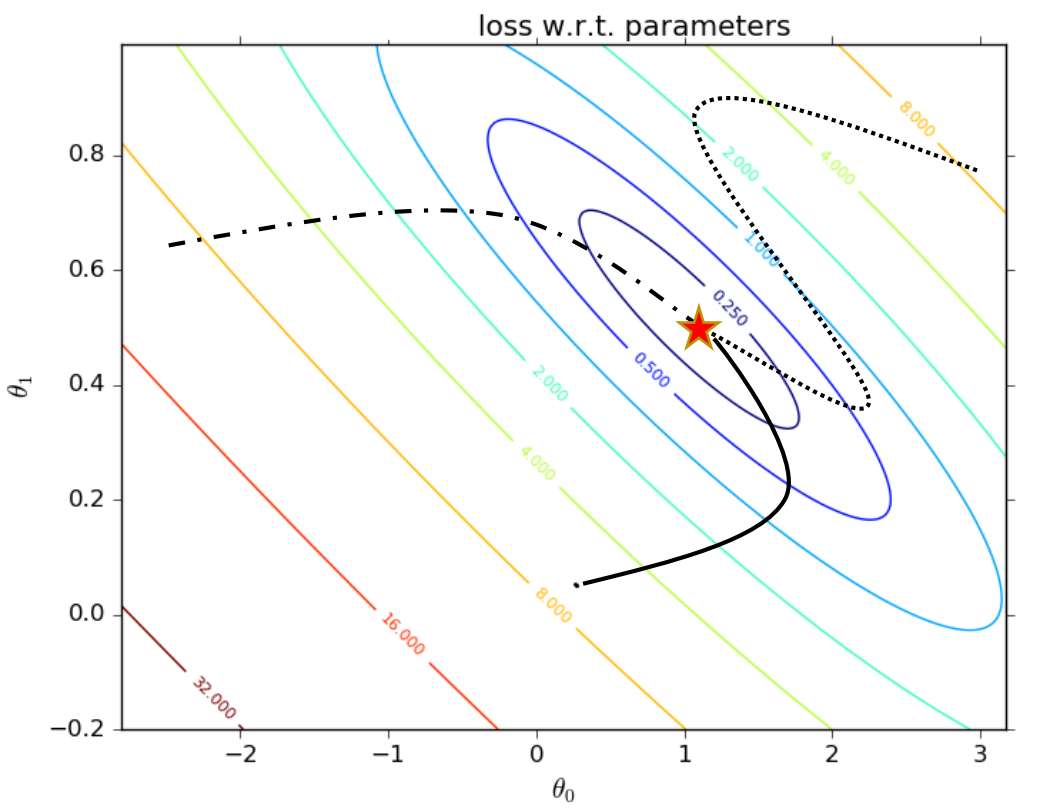
基本搜索步骤

- 随机选择一个参数初始化 θ
- 根据数据和梯度算法来更新 θ
- 直到走到局部一个最小区域(local minimum)



2.2 梯度更新方式

凸优化目标函数具有唯一最小点



- 不同的初始化参数最终也会学习到相同的最优值

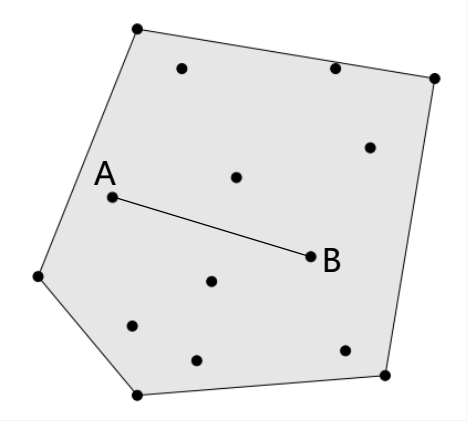
2.2 梯度更新方式

凸集 (Convex Set)

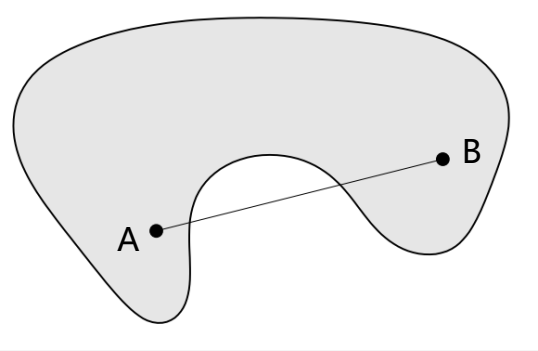
一个点集S被称为凸集，当且仅当该S里的任意两点A和B的连线上任意一点同样属于S

$$tx_1 + (1-t)x_2 \in S$$

$$\text{for all } x_1, x_2 \in S, 0 \leq t \leq 1$$



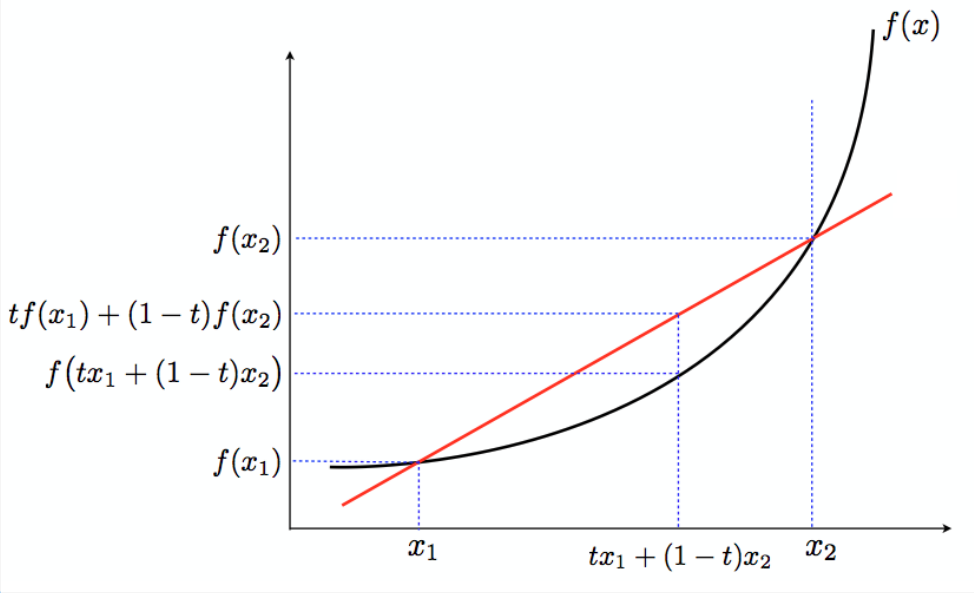
凸集



非凸集

2.2 梯度更新方式

凸函数 (Convex Function)



凸函数的定义

$f:\mathbb{R}^n \to \mathbb{R}$ 是凸函数: $\mathbf{dom}\ f$ 是一个凸集，并且满足

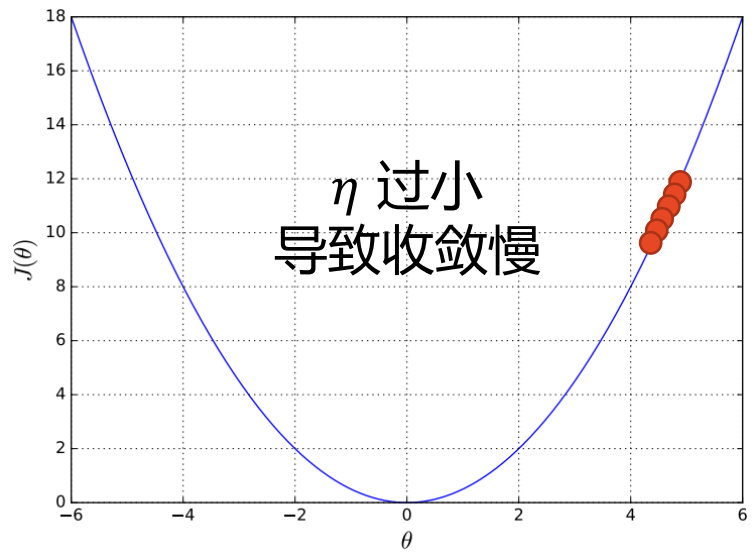
$$f(tx_1+(1-t)x_2) \le tf(x_1)+(1-t)f(x_2)$$

$$\forall x_1, x_2 \in \mathbf{dom} f, 0 \le t \le 1$$

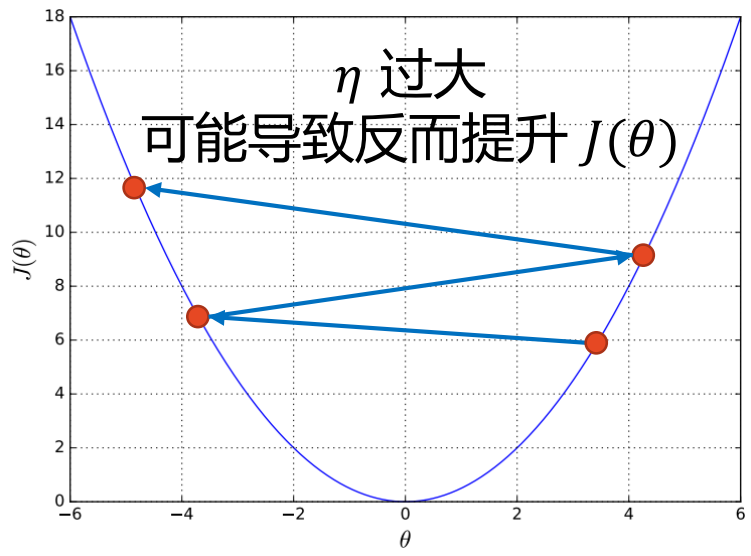
2.2 梯度更新方式

学习率的选择

$$\theta_{\text{new}} = \theta_{\text{old}} - \eta \frac{\partial J(\theta)}{\partial \theta}$$



- 初始点可能距离最优点太远，从而导致收敛速度慢



- 可能越过最优点
- 可能无法收敛
- 甚至可能发散

- 要检查梯度下降是否有效工作，可以打印出每几个迭代得到的损失 $J(\theta)$, 如果发现 $J(\theta)$ 并没有正常地下降，调整学习率 η

2.3 线性回归矩阵形式

从代数视角来看线性回归

训练数据矩阵

$$\boldsymbol{X}=\begin{pmatrix}\boldsymbol{x}^{(1)}\\ \boldsymbol{x}^{(2)}\\ \vdots\\ \boldsymbol{x}^{(n)}\end{pmatrix}=\begin{pmatrix}x_1^{(1)} & x_2^{(1)} & \cdots & x_d^{(1)}\\ x_1^{(2)} & x_2^{(2)} & & x_d^{(2)}\\ \vdots & & \ddots & \vdots\\ x_1^{(n)} & x_2^{(n)} & \cdots & x_d^{(n)}\end{pmatrix}$$

参数 $\boldsymbol{\theta}=\begin{pmatrix}\theta_1\\ \theta_2\\ \vdots\\ \theta_d\end{pmatrix}$ 标签 $\boldsymbol{y}=\begin{pmatrix}y_1\\ y_2\\ \vdots\\ y_n\end{pmatrix}$

预测

$$\hat{\boldsymbol{y}}=\boldsymbol{X\theta}=\begin{pmatrix}\boldsymbol{x}^{(1)}\boldsymbol{\theta}\\ \boldsymbol{x}^{(2)}\boldsymbol{\theta}\\ \vdots\\ \boldsymbol{x}^{(n)}\boldsymbol{\theta}\end{pmatrix}$$

目标函数

$$J(\boldsymbol{\theta})=\frac{1}{2}(\boldsymbol{y}-\hat{\boldsymbol{y}})^{\mathsf{T}}(\boldsymbol{y}-\hat{\boldsymbol{y}})=\frac{1}{2}(\boldsymbol{y}-\boldsymbol{X\theta})^{\mathsf{T}}(\boldsymbol{y}-\boldsymbol{X\theta})$$

2.3 线性回归矩阵形式

目标函数

$$J(\boldsymbol{\theta}) = \frac{1}{2}(\boldsymbol{y} - \boldsymbol{X\theta})^{\mathsf{T}}(\boldsymbol{y} - \boldsymbol{X\theta}) \qquad \min_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$$

对参数向量的梯度

$$\frac{\partial J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = -\boldsymbol{X}^{\mathsf{T}}(\boldsymbol{y} - \boldsymbol{X\theta})$$

最优参数求解

$$\frac{\partial J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = 0 \rightarrow \boldsymbol{X}^{\mathsf{T}}(\boldsymbol{y} - \boldsymbol{X\theta}) = 0$$

$$\rightarrow \boldsymbol{X}^{\mathsf{T}}\boldsymbol{y} = \boldsymbol{X}^{\mathsf{T}}\boldsymbol{X\theta}$$

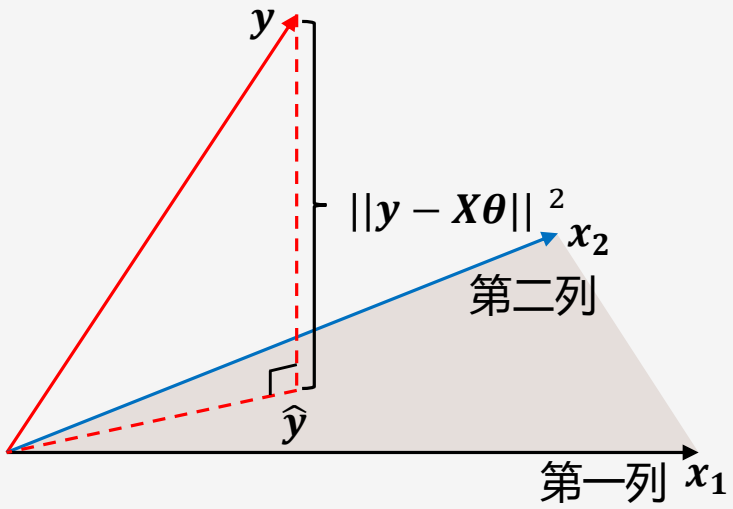
$$\rightarrow \widehat{\boldsymbol{\theta}} = (\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathsf{T}}\boldsymbol{y}$$

2.3 线性回归矩阵形式

预测值

$$\hat{y} = X(X^{\top}X)^{-1}X^{\top}y = Hy$$

H:帽子矩阵



几何解释

- 数据矩阵的列向量$[x_1, x_2, \ldots, x_d]$张成一个 $\mathbb{R}^n$上的子空间
- H就是将标签向量y投影到该子空间的映射

$$X = \begin{pmatrix} x_1^{(1)} & x_2^{(1)} & \cdots & x_d^{(1)} \\ x_1^{(2)} & x_2^{(2)} & & x_d^{(1)} \\ \vdots & & \ddots & \vdots \\ x_1^{(n)} & x_2^{(n)} & \cdots & x_d^{(n)} \end{pmatrix} = [x_1, x_2, \ldots, x_d] \qquad y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

2.3 线性回归矩阵形式

$X^T X$为奇异矩阵的情况

- 当数据矩阵的一些列向量线性相关时
  - 例如$x_2 = 3x_1$
- $X^\top X$为奇异矩阵，所以$\hat{\theta} = (X^\top X)^{-1} X^\top y$无法被直接计算。

解决方案

- 正则化（Regularization）
- $J(\mu) = \frac{1}{2}(y - X\theta)^\top (y - X\theta) + \frac{\lambda}{2}||\theta||_2^2$

2.3 线性回归矩阵形式

带正则项的线性回归矩阵形式

优化目标

$$J(\boldsymbol{\theta}) = \frac{1}{2}(\boldsymbol{y} - \boldsymbol{X\theta})^{\top}(\boldsymbol{y} - \boldsymbol{X\theta}) + \frac{\lambda}{2}||\boldsymbol{\theta}||_2^2 \qquad \min_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$$

对参数向量的梯度

$$\frac{\partial J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = -\boldsymbol{X}^{\top}(\boldsymbol{y} - \boldsymbol{X\theta}) + \lambda\boldsymbol{\theta}$$

最优参数求解

$$\frac{\partial J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = 0 \rightarrow -\boldsymbol{X}^{\top}(\boldsymbol{y} - \boldsymbol{X\theta}) + \lambda\boldsymbol{\theta} = 0$$

$$\rightarrow \boldsymbol{X}^{\top}\boldsymbol{y} = (\boldsymbol{X}^{\top}\boldsymbol{X} + \lambda\boldsymbol{I})\boldsymbol{\theta}$$

$$\rightarrow \widehat{\boldsymbol{\theta}} = (\boldsymbol{X}^{\top}\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^{\top}\boldsymbol{y}$$

2.4 最大似然估计

判别模型

- 建模预测变量和观测变量之间的关系
- 又名条件模型（Conditional Models）
- 确定性判别模型：$y = f_\theta(x)$
- **概率判别模型**：$p_\theta(y|x)$

带高斯白噪声的线性拟合

$$y = f_\theta(x) + \epsilon = \theta_0 + \sum_{j=1}^{d} \theta_j x_j + \epsilon = \theta^\top x + \epsilon$$

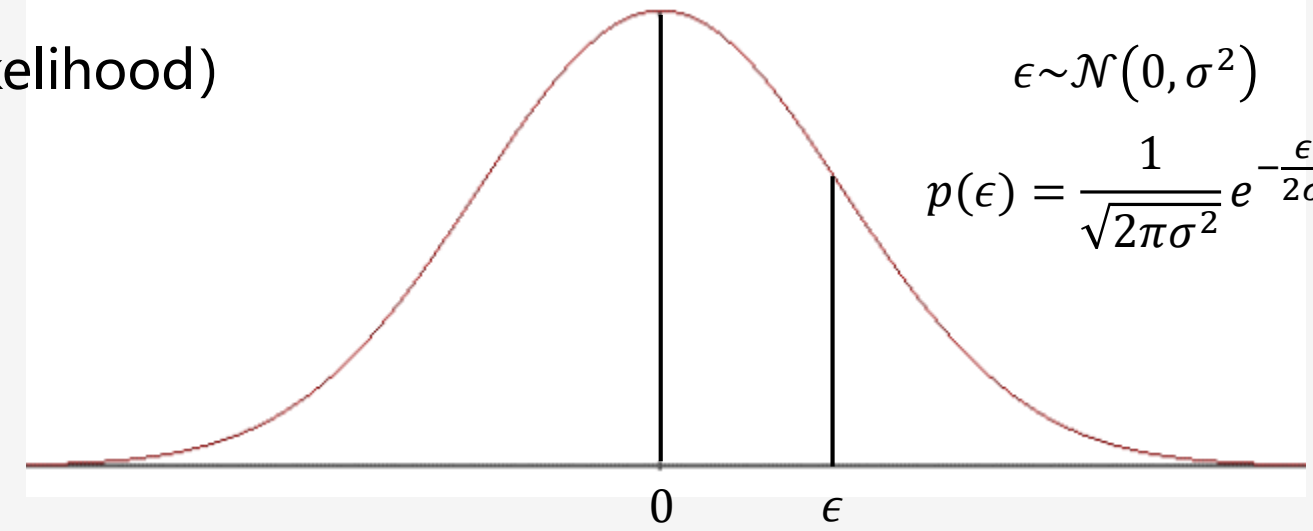
$$\epsilon \sim \mathcal{N}(0, \sigma^2)$$

$$x = (1, x_1, x_2, \ldots, x_d)$$

2.4 最大似然估计

优化目标

最大似然（likelihood）



$$\epsilon \sim \mathcal{N}(0, \sigma^2)$$

$$p(\epsilon) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\epsilon^2}{2\sigma^2}}$$

一个数据点的标签预测似然

$$p(y|x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\theta^T x)^2}{2\sigma^2}}$$

2.4 最大似然估计

概率判别模型的学习

最大化训练数据的似然

$$\max_{\theta} \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i-\theta^\top x_i)^2}{2\sigma^2}}$$

最大化训练数据的对数似然

$$\log \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i-\theta^\top x_i)^2}{2\sigma^2}} = \sum_{i=1}^{N} \log \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i-\theta^\top x_i)^2}{2\sigma^2}} = -\sum_{i=1}^{N} \frac{\left(y_i-\theta^\top x_i\right)^2}{2\sigma^2} + \text{const}$$

$$\min_{\theta} \sum_{i=1}^{N} \left(y_i-\theta^\top x_i\right)^2$$

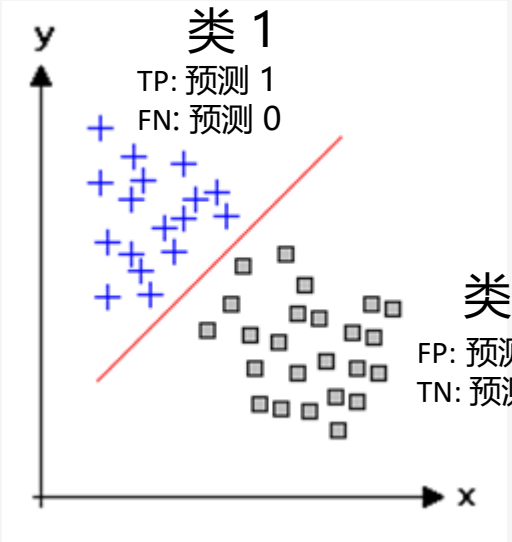
等价于最小均方误差学习

2.5 分类指标

评估指标

		预测	
		1	0
标签	1	True Positive	False Negative
	0	False Positive	True Negative

- True / False
  - True：预测 = 标签
  - False：预测 ≠ 标签
- Positive / Negative
  - Positive：预测$y = 1$
  - Negative：预测$y = 0$



2.5 分类指标

评估指标

		预测 1	预测 0
标签	1	True Positive	False Negative
	0	False Positive	True Negative

精度(Accuracy)

- 分类正确的样本占样本总数的比例

$$\mathrm{Acc} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}$$

2.5 分类指标

评估指标

预测

标签	1	0
1	True Positive	False Negative
0	False Positive	True Negative

精确率(Precision)

- 预测为1的样本中标签为1的比例

$$\mathrm{Prec} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

预测

标签	1	0
1	True Positive	False Negative
0	False Positive	True Negative

召回率(Recall)

- 标签为1的样本中预测为1的比例

$$\mathrm{Rec} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

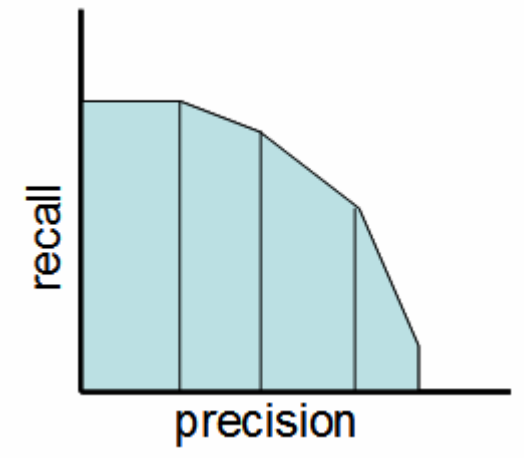
2.5 分类指标

评估指标

- 精确率和召回率的权衡

$$\hat{y} = \begin{cases} 1, & p_\theta(y=1|x) > h \\ 0, & \text{otherwise} \end{cases}$$

  - 阈值越高，精确率越高，召回率越低
    - 极端情况：阈值=0.99
  - 阈值越低，精确率越低，召回率越高
    - 极端情况：阈值=0



- F1分数

$$\mathrm{F1} = \frac{2 \times \mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}}$$

2.5 分类指标

评估指标

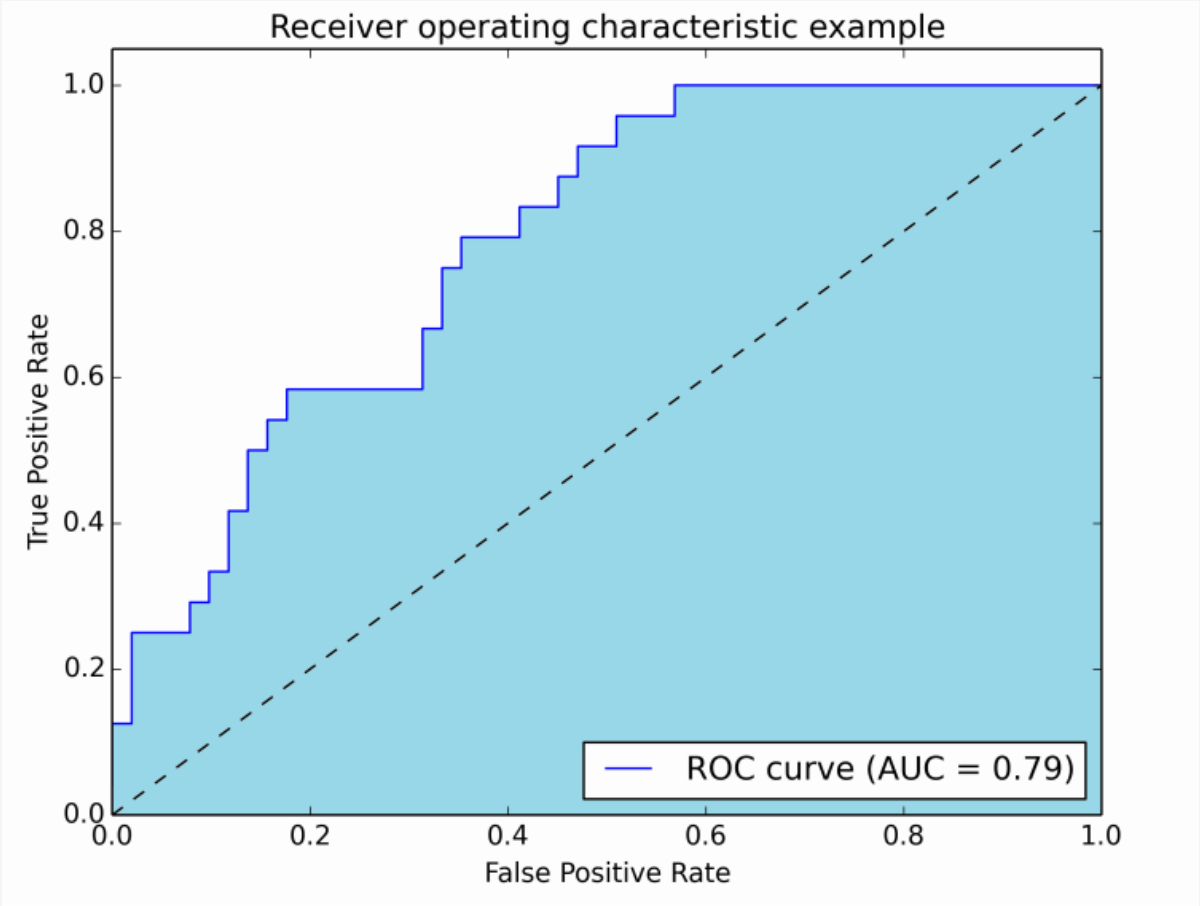
- 基于排序的度量：ROC曲线下面积（AUC）

- True Positive Rate

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

- False Positive Rate

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}}$$

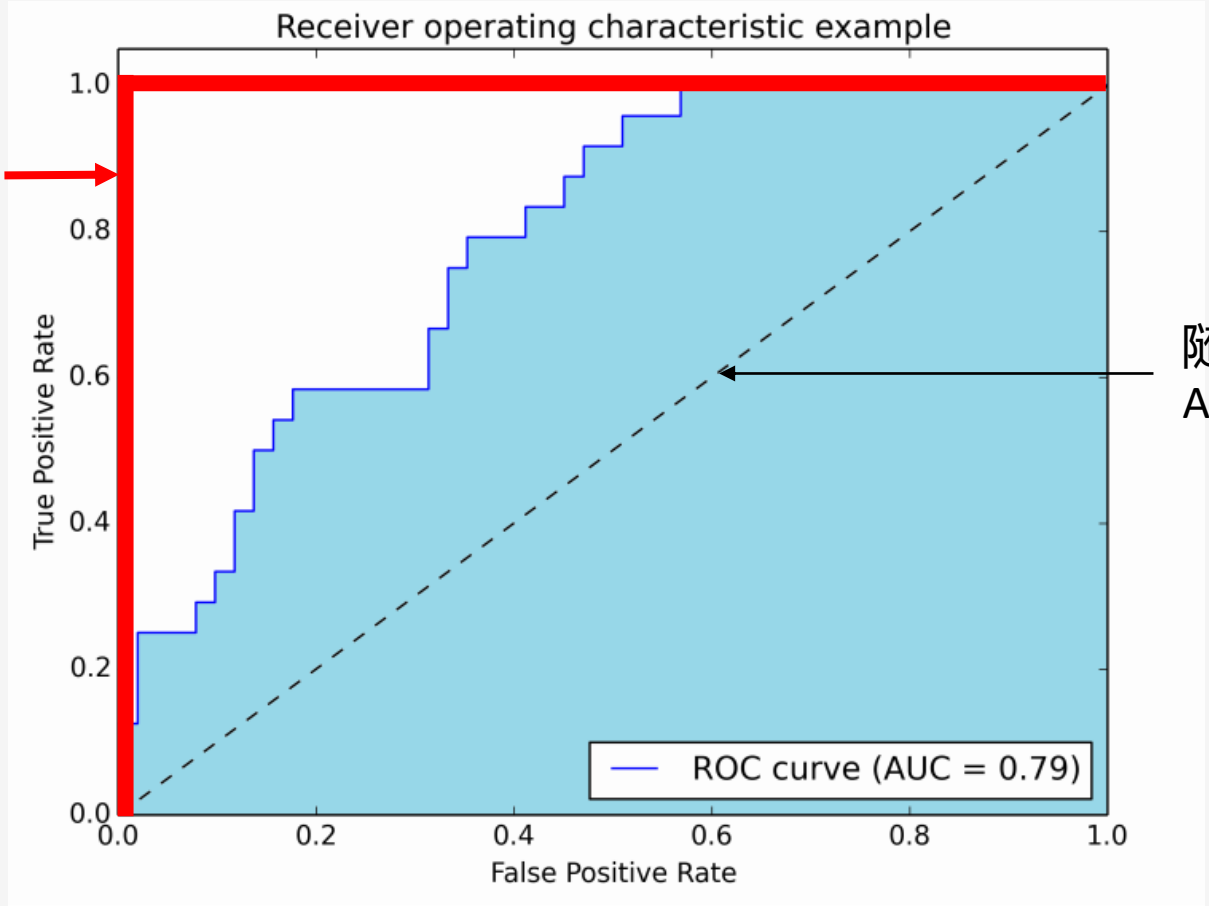


2.5 分类指标

评估指标

- 基于排序的度量：ROC曲线下面积（AUC）

完美预测
AUC=1

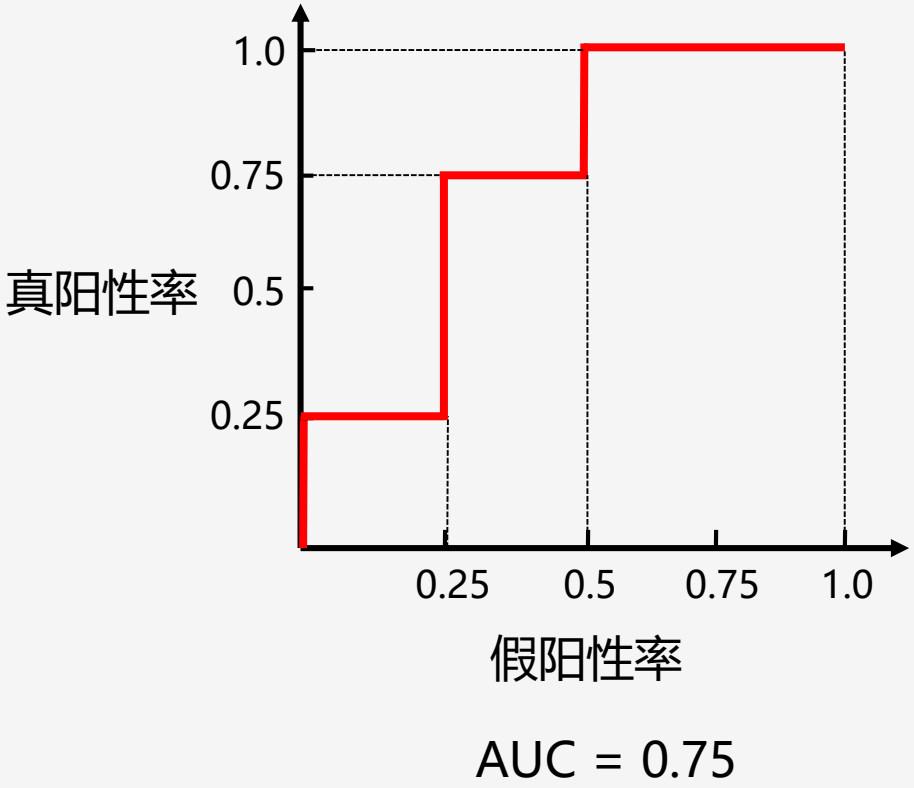


随机预测
AUC = 0.5

2.5 分类指标

评估指标

- AUC计算例子



AUC = 0.75

Prediction	Label
0.91	1
0.85	0
0.77	1
0.72	1
0.61	0
0.48	1
0.42	0
0.33	0

2.6 逻辑斯谛回归

分类问题

给定

- 样本空间$\mathbb{X}$中一个样本x $(x \in \mathbb{X})$的描述
- 一个固定的类别集：$C = \{c_1, c_2, \cdots, c_m\}$

求解

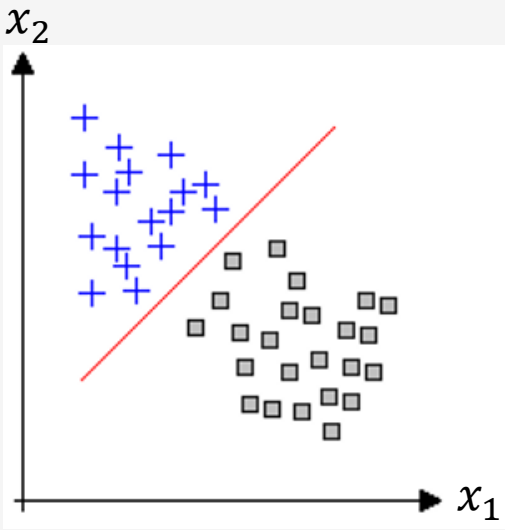
- x的类别:$f(x) \in C$，其中$f(x)$是一个定义域为$\mathbb{X}$，值域为C的类别函数

二分类

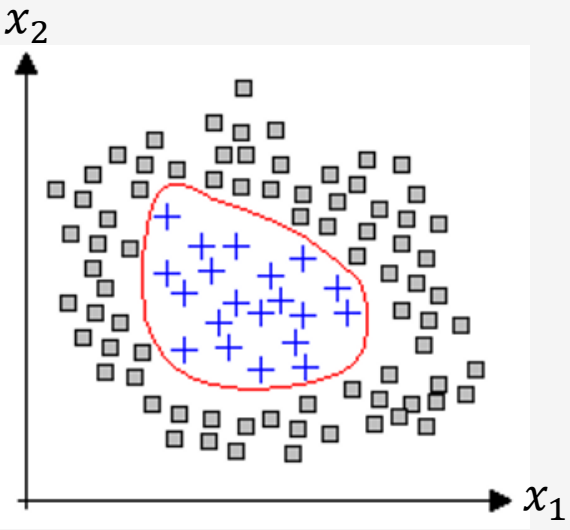
- 假如类别集是二元的，即$C = \{0,1\}$({错误，正确}，{负，正})，那么这就是二分类问题

2.6 逻辑斯谛回归

二分类



线性可分



线性不可分

线性可分性：是否存在 $ax_1 + bx_2 + c = 0$

使得对于所有的正例：$ax_1 + bx_2 + c > 0$

对于所有的负例：$ax_1 + bx_2 + c < 0$

2.6 逻辑斯谛回归

线性判别模型

判别模型

- 性质
  - 建模预测变量和观测变量之间的关系
  - 也称作条件模型(Conditional Models)
- 分类
  - 确定性判别模型：$y = f_\theta(x)$
    - 对于分类任务不可微分
  - 概率判别模型：$p_\theta(y|x)$
    - 对于分类任务可微分

二分类

$$p_\theta(y = 1|x)$$

$$p_\theta(y = 0|x) = 1 - p_\theta(y = 1|x)$$

熵（Entropy）

- 在信息论中，熵用来衡量一个随机事件的不确定性

- 自信息（Self Information）

$$I(x) = -\log\big(p(x)\big)$$

  - x表示一个事件
  - $p(x)$表示x发生的概率
  - 信息量， x越不可能发生时，它一旦发生后的信息量就越大

熵（Entropy）

- 在信息论中，熵用来衡量一个随机事件的不确定性
- 自信息（Self Information）

$$I(x) = -\log\big(p(x)\big)$$

- 熵的计算

$$\begin{aligned} H(X) &= \mathbb{E}_X[\mathrm{I}(x)] \\ &= \mathbb{E}_X[-\log(p(x))] \\ &= -\sum_{x\in\mathcal{X}} p(x)\log p(x) \end{aligned}$$

熵（Entropy）

- 假设对于这门课程，我们有三种可能的情况发生

事件编号	事件	概率 p	信息量 I
x_1	优秀	$p=0.7$	$I=-\ln(0.7)=0.36$
x_2	及格	$p=0.2$	$I=-\ln(0.2)=1.61$
x_3	不及格	$p=0.1$	$I=-\ln(0.1)=2.30$

- 某某同学不及格！好大的信息量！相比较来说，“优秀”事件的信息量反而小了很多。上面的问题的熵是：

$$
\begin{aligned}
H(p) &= -[p(x_1)\ln p(x_1) + p(x_2)\ln p(x_2) + p(x_3)\ln p(x_3)] \\
&= 0.7 \times 0.36 + 0.2 \times 1.61 + 0.1 \times 2.30 \\
&= 0.804
\end{aligned}
$$

2.6 逻辑斯谛回归

损失函数

交叉熵损失

- 离散的情况 $H(p,q) = -\sum_x p(x)\log q(x)$
- 连续的情况 $H(p,q) = -\int_x p(x)\log q(x)\,dx$

分类问题计算交叉熵损失

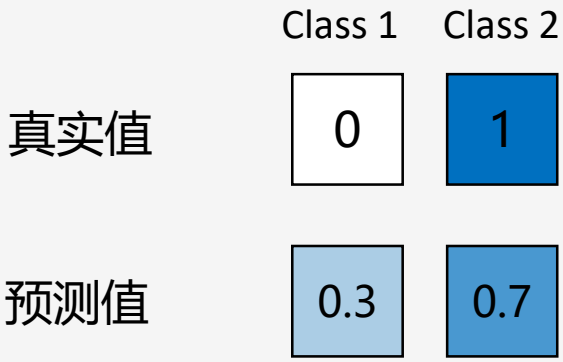
Ground Truth	0	1	0	0	0
Prediction	0.1	0.6	0.05	0.05	0.2

$$\mathcal{L}(y,x,p_\theta) = -\sum_k \delta(y=c_k)\log p_\theta(y=c_k|x)$$

$$\delta(z) = \begin{cases} 1, & z \text{ is true} \\ 0, & \text{otherwise} \end{cases}$$

2.6 逻辑斯谛回归

二分类的交叉熵



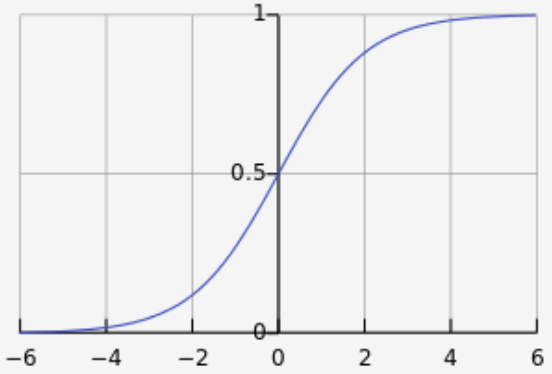
- 损失函数

$$\mathcal{L}(y, x, p_\theta) = -\delta(y=1)\log p_\theta(y=1|x) - \delta(y=0)\log p_\theta(y=0|x)$$
$$= -y\log p_\theta(y=1|x) - (1-y)\log\big(1 - p_\theta(y=1|x)\big)$$

2.6 逻辑斯谛回归

逻辑斯谛(Logistic)回归

Sigmoid函数



- 逻辑斯谛回归是一个二分类模型

$$p_\theta(y=1|x)=\sigma(\theta^\top x)=\frac{1}{1+e^{-\theta^\top x}}$$

$$p_\theta(y=0|x)=\frac{e^{-\theta^\top x}}{1+e^{-\theta^\top x}}$$

- 交叉熵损失函数

$$\mathcal{L}(y,x,p_\theta)=-y\log\sigma(\theta^\top x)-(1-y)\log(1-\sigma(\theta^\top x))$$

- 梯度

$$\frac{\partial\mathcal{L}(y,x,p_\theta)}{\partial\theta}=-y\frac{1}{\sigma(\theta^\top x)}\sigma(z)(1-\sigma(z))x-(1-y)\frac{-1}{1-\sigma(\theta^\top x)}\sigma(z)(1-\sigma(z))x$$

$$=(\sigma(\theta^\top x)-y)x$$

$$\theta\leftarrow\theta+\eta\left(y-\sigma(\theta^\top x)\right)x$$

线性回归：$\theta_{\text{new}}=\theta_{\text{old}}+\eta(y_i-f_\theta(x_i))x_i$

$$\frac{\partial\sigma(z)}{\partial z}=\sigma(z)(1-\sigma(z))$$

标签的决定

- 逻辑斯谛回归求出的概率

$$p_\theta(y=1|x)=\delta\left(\theta^\top x\right)=\frac{1}{1+e^{-\theta^\top x}}$$

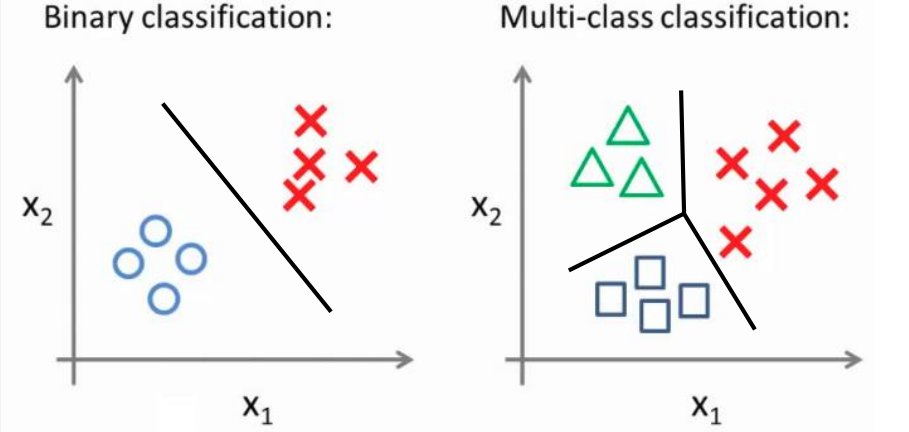
$$p_\theta(y=0|x)=\frac{e^{-\theta^\top x}}{1+e^{-\theta^\top x}}$$

- 设置阈值(threshold)h决定示例最终标签

$$\hat{y}=\begin{cases}1, & p_\theta(y=1|x)>h\\ 0, & \text{otherwise}\end{cases}$$

2.6 逻辑斯谛回归

多分类



多分类交叉熵

$$\mathcal{L}(y, x, p_\theta) = -\sum_k \delta(y = c_k) \log p_\theta(y = c_k | x)$$

$$\delta(z) = \begin{cases} 1, & z \text{ is true} \\ 0, & \text{otherwise} \end{cases}$$

真实值	0	1	0
预测值	0.1	0.7	0.2

多类别逻辑斯谛回归

- 类别集

$$C = \{c_1, c_2, \cdots, c_m\}$$

- 预测$p_\theta(y = c_j|x)$的概率

$$p_\theta(y = c_j|x) = \frac{e^{\theta_j^\top x}}{\sum_{k=1}^{m} e^{\theta_k^\top x}} \quad \text{for } j = 1, \cdots, m$$

- Softmax
  - 参数$\theta = \{\theta_1, \theta_2, \cdots, \theta_m\}$
  - 可以标准化成$m - 1$组参数

多类别逻辑斯谛回归

- 对一个示例的学习 $(x, y = c_j)$
  - 最大对数化似然(log-likelihood)

$$\max_{\theta} \log p_\theta(y = c_j|x)$$

  - 梯度

$$\frac{\partial \log p_\theta(y = c_j|x)}{\partial \theta_j} = \frac{\partial}{\partial \theta_j} \log \frac{e^{\theta_j^\top x}}{\sum_{k=1}^m e^{\theta_k^\top x}}$$

$$= x - \frac{\partial}{\partial \theta_j} \log \sum_{k=1}^m e^{\theta_k^\top x}$$

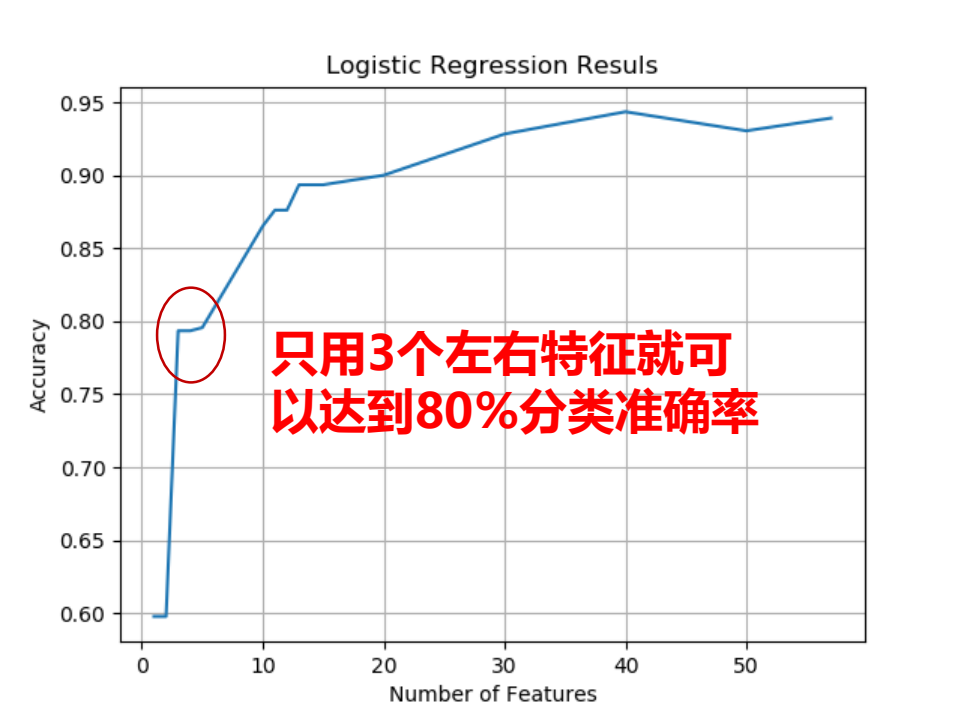
$$= x - \frac{e^{\theta_j^\top x}}{\sum_{k=1}^m e^{\theta_k^\top x}} x = \left(1 - p_\theta(y = c_j|x)\right) x$$

线性模型应用举例

垃圾邮件分类

二分类任务：

- 只用绝对值排名靠前的k个特征，重新训练和评估模型性能



- 是否可以显式地训练参数，以优先构建一个“稀疏”的模型？为什么？
  - 低复杂度的模型不容易过拟合

- 模型训练的目的是为了最小化损失函数：

$$\hat{\theta} = \min_{\theta} Loss(\theta)$$

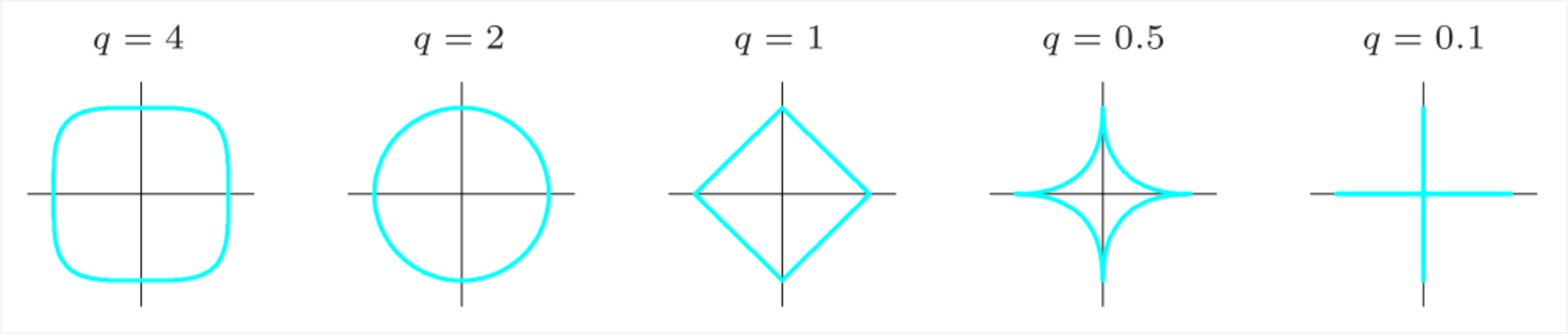
如何改变这个损失函数以降低训练出来的模型的复杂度？

线性模型应用举例

垃圾邮件分类

二分类任务：

- 在损失函数中引入正则化项
- 经典正则化方法$\|\theta\|_q$的数值分布图



Ridge　　LASSO

“正则化”后的损失函数

$$\hat{\theta} = \min_{\theta}\{Loss(\theta) + c\|\theta\|_0\}$$

c控制正则化项的相对重要性

线性模型总结

- 线性回归是机器学习中最基础的参数化学习模型，线性回归任务是机器学习中最基础的有监督学习任务。

- 逻辑斯谛回归，虽然其名字包含“回归”二字，但它是最具有代表性的机器学习分类模型，至今还在学术研究和工业落地场景中被广泛使用。

	激活函数	损失函数	优化方法
线性回归	-	$(y - \mathbf{w}^{\mathrm{T}}\mathbf{x})^2$	最小二乘、梯度下降
Logistic 回归	$\sigma(\mathbf{w}^{\mathrm{T}}\mathbf{x})$	$\mathbf{y} \log \sigma(\mathbf{w}^{\mathrm{T}}\mathbf{x})$	梯度下降
Softmax 回归	$\mathrm{softmax}(W^{\mathrm{T}}\mathbf{x})$	$\mathbf{y} \log \mathrm{softmax}(W^{\mathrm{T}}\mathbf{x})$	梯度下降