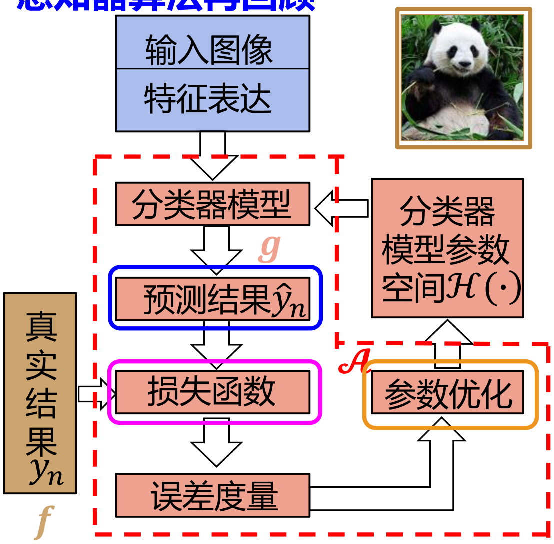
人工智能与自动化学院

模式识别与机器学习

第五讲　逻辑斯蒂回归（*Logistic Regression*）

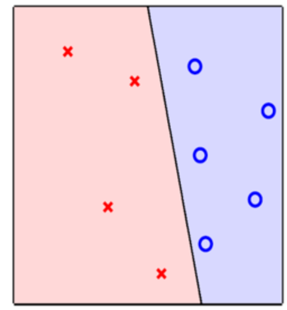
5.1 逻辑斯蒂回归问题

感知器算法再回顾



$$f(\mathbf{x}_n) = y_n \in \{1, -1\}$$

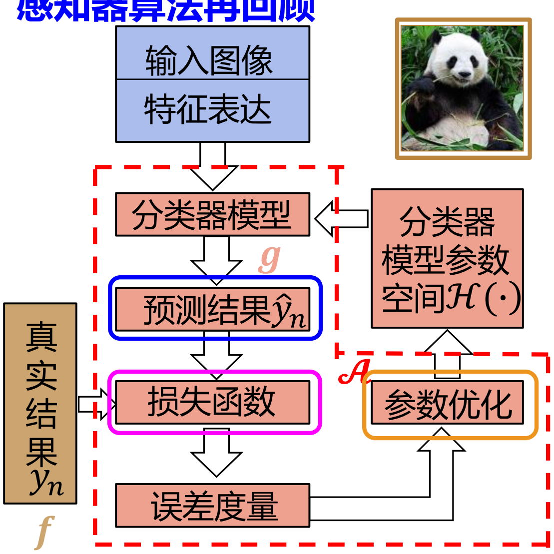
线性可分



- 设置初始分类面(权重)$\boldsymbol{w}_0$
- 如果有样本分错，就修正权重

5.1 逻辑斯蒂回归问题

感知器算法再回顾



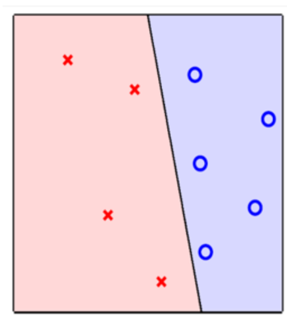
$$\hat{y}_{n(t)} = \mathrm{sign}(\boldsymbol{w}_t^T \mathbf{x}_{n(t)})$$

算法收敛：

$$L_{in} = \sum_{n=1}^{N} [\![y_n \neq \hat{y}_n]\!] = 0$$

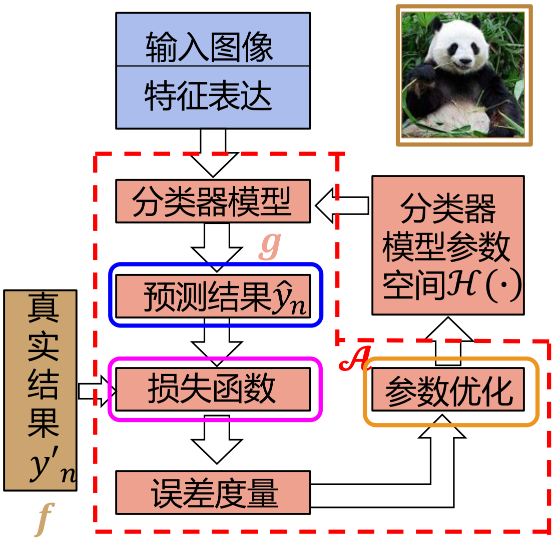
$$\boldsymbol{w}_{t+1} = \boldsymbol{w}_t + y_n \mathbf{x}_{n(t)}$$

线性可分



- 设置初始分类面(权重)$\boldsymbol{w}_0$
- 如果有样本分错，就修正权重

5.1 逻辑斯蒂回归问题

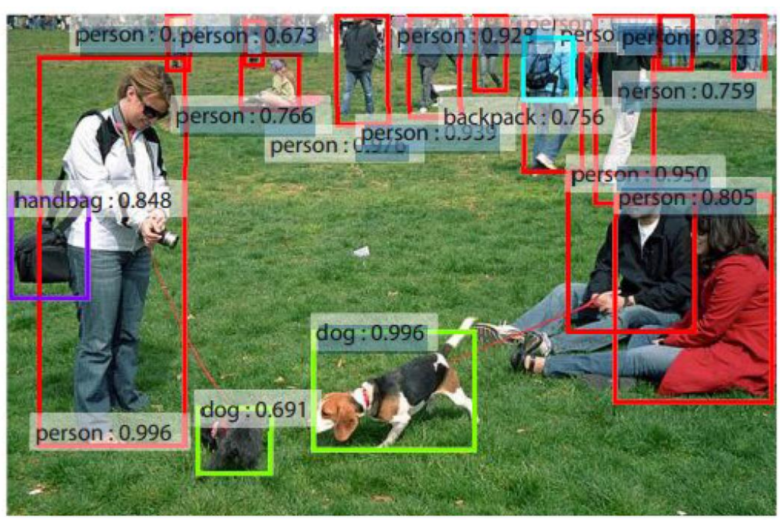


$$f(\mathbf{x}_n) = y'_n = P(+1|\mathbf{x}_n) \ \in [0,1]$$

$$\hat{y}_n = sign(P(+1|\mathbf{x}_n) - 0.5) \ \in \{1,-1\}$$

5.1 逻辑斯蒂回归问题

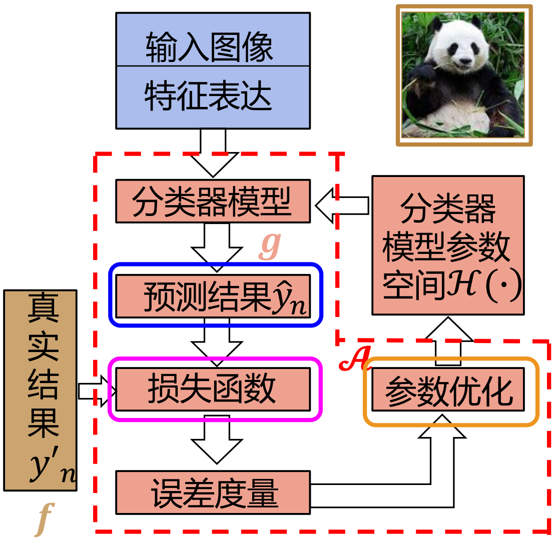
逻辑斯蒂回归应用示例



感知器算法：硬分类

逻辑斯蒂回归：软分类
（*"Soft" binary classification*）

5. 1 逻辑斯蒂回归问题



逻辑斯蒂回归：软分类

$$f(\mathbf{x}_n) = y'_n = P(+1|\mathbf{x}_n) \in [0,1]$$

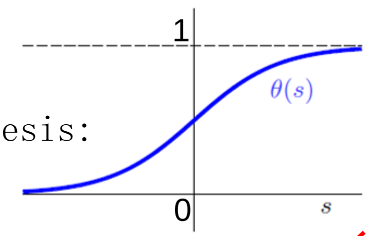
$$\hat{y}_n = sign(P(+1|\mathbf{x}_n) - 0.5) \in \{1,-1\}$$

$$\mathbf{x} = (x_1, x_2, \dots x_d)^T$$

$$s = \sum_{i=0}^{d} w_i x_i$$

logistic hypothesis:

$$h(\mathbf{x}) = \theta(\mathbf{w}^T\mathbf{x})$$



5. 1 逻辑斯蒂回归问题

逻辑斯蒂函数

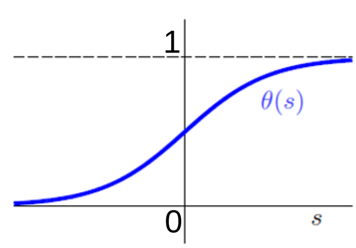
$\theta(-\infty) = 0$; $\theta(0) = \frac{1}{2}$; $\theta(\infty) = 1$

$$\theta(s) = \frac{e^s}{1+e^s} = \frac{1}{1+e^{-s}}$$

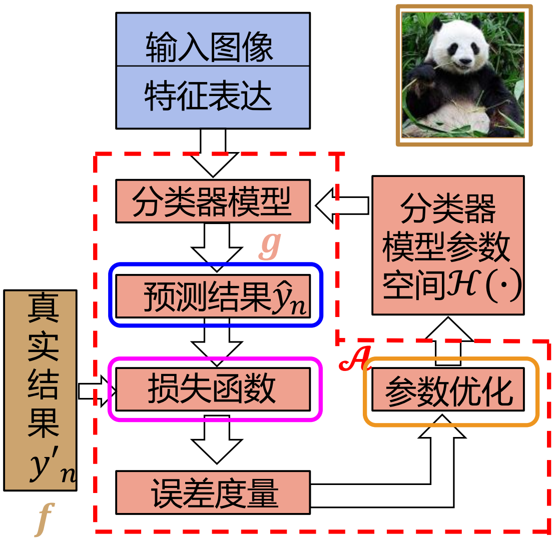
---- Sigmoid 函数：平滑 (*Smooth*)、单调 (*Monotonic*)

逻辑斯蒂回归用如下模型来估计$f(\mathbf{x}_n)$

$$h(\mathbf{x}) = \theta(\mathbf{w}^T\mathbf{x}) = \frac{1}{1 + exp(-\mathbf{w}^T\mathbf{x})}$$



5.1 逻辑斯蒂回归问题



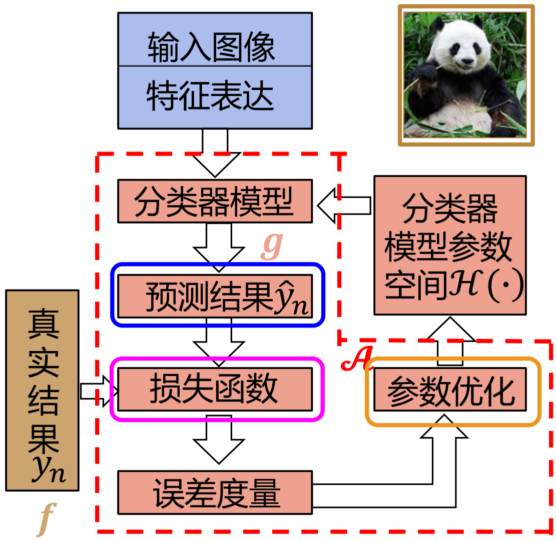
逻辑斯蒂回归：软分类

$$f(\mathbf{x}_n) = y'_n = P(+1|\mathbf{x}_n) \in [0,1]$$

$$\hat{y}_n = h(\mathbf{x}) = \theta(\mathbf{w}^T\mathbf{x}) = \frac{1}{1+exp(-\mathbf{w}^T\mathbf{x})}$$

第五讲 逻辑斯蒂回归（*Logistic Regression*）

5. 2 逻辑斯蒂回归损失



感知器(线性分类)

$$\hat{y}_{n(t)} = \mathrm{sign}(\mathbf{w}_t^T \mathbf{x}_{n(t)})$$

$$L_{in} = \sum_{n=1}^{N} [\![y_n \neq \hat{y}_n]\!]$$

线性回归

$$\hat{y}_n = \mathbf{w}^T \mathbf{x}_n$$

$$L_{in}(\mathbf{w}) = \frac{1}{N} \sum_{n=1}^{N} (\mathbf{w}^T \mathbf{x}_n - y_n)^2$$

逻辑斯蒂回归

$$\hat{y}_n = h(\mathbf{x}) = \theta(\mathbf{w}^T \mathbf{x}) = \frac{1}{1 + exp(-\mathbf{w}^T \mathbf{x})}$$

$$L_{in} = ?$$

5. 2 逻辑斯蒂回归损失

逻辑斯蒂回归可以使用平方误差作为损失函数吗?

$$L_{in}(\mathbf{w}) = (\mathbf{w}^T \mathbf{x} - y)^2$$

5.2 逻辑斯蒂回归损失

逻辑斯蒂回归可以使用平方误差作为损失函数吗?

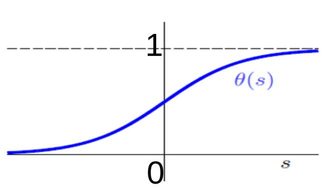
$$L_{in}(\boldsymbol{w}) = (\theta(\boldsymbol{w}^T\mathbf{x}) - y')^2$$

5.2 逻辑斯蒂回归损失

逻辑斯蒂回归可以使用平方误差作为损失函数吗?

$$L_{in}(\boldsymbol{w}) = (\theta(y\boldsymbol{w}^T\mathbf{x}) - 1)^2$$

$$\frac{\partial L_{in}(\boldsymbol{w},\mathbf{x},y)}{\partial w_i} = 2(\theta(y\boldsymbol{w}^T\mathbf{x}) - 1)\underbrace{\theta(y\boldsymbol{w}^T\mathbf{x})(1\text{-}\theta(y\boldsymbol{w}^T\mathbf{x}))}_{\frac{\partial\theta(z)}{\partial z}}yx_i$$



$$if\ \ (y\boldsymbol{w}^T\mathbf{x})) > 0 \quad \nabla L_{in}(\boldsymbol{w},\mathbf{x},y) = 0$$

$$\boxed{if\ \ (y\boldsymbol{w}^T\mathbf{x})) < 0 \quad \nabla L_{in}(\boldsymbol{w},\mathbf{x},y) = 0}$$

5.2 逻辑斯蒂回归损失

逻辑斯蒂回归

$f(\mathbf{x}_n) = y'_n = P(+1|\mathbf{x}_n) \in [0,1]$

(理想)训练样本:

$(\mathbf{x}_1,\ y'_1 = 0.9 = P(+1|\mathbf{x}_1))$

$(\mathbf{x}_2,\ y'_2 = 0.2 = P(+1|\mathbf{x}_2))$

$\vdots$

$(\mathbf{x}_N,\ y'_N = 0.6 = P(+1|\mathbf{x}_N))$

实际训练样本(含噪标签):

$(\mathbf{x}_1,\ y_1 = \text{o} = 1 \sim P(+1|\mathbf{x}_1))$

$(\mathbf{x}_2,\ y_2 = \times = -1 \sim P(+1|\mathbf{x}_2))$

$\vdots$

$(\mathbf{x}_N,\ y_N = \times = -1 \sim P(+1|\mathbf{x}_N))$

$L_{in} = ?$

5.2 逻辑斯蒂回归损失

逻辑斯蒂回归的最佳解:

$$g = \underset{\mathbf{w}}{argmin} \frac{1}{N}\sum_{n=1}^{N} -ln\theta(y_n \mathbf{w}^T \mathbf{x}_n)$$

$$\theta(y_n \mathbf{w}^T \mathbf{x}_n) = \frac{1}{1+\exp(-y_n \mathbf{w}^T \mathbf{x}_n)}$$

$$g = \underset{\mathbf{w}}{argmin} \frac{1}{N}\sum_{n=1}^{N} \ln(1+exp(-y_n \mathbf{w}^T \mathbf{x}_n))$$

交叉熵损失
(*Cross-Entropy Loss*)

$$L_{in} = \frac{1}{N}\sum_{n=1}^{N} \ln(1+exp(-y_n \mathbf{w}^T \mathbf{x}_n))$$

5.2 逻辑斯蒂回归损失

交叉熵介绍：

Distribution p(x)

$p(x=1)=f(x_n)$

$p(x=0)=1-f(x_n)$

cross entropy

Distribution q(x)

$q(x=1)=\theta(x_n)$

$q(x=0)=1-\theta(x_n)$

$$H(p,q)=-\sum_{x}p(x)\ln(q(x))$$

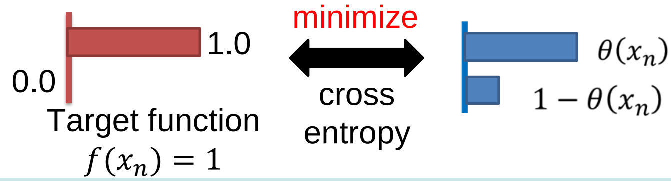
$$s=w^Tx$$

$$\theta(s)=\frac{1}{1+\exp(-s)}$$

$$H(f(x_n),\theta(x_n)))=\sum_{n}-[f(x_n)ln\theta(x_n)+(1-f(x_n))ln(1-ln\theta(x_n))]$$

$$H(f(x_n),\theta(x_n)))=\sum_{n}-[ln\theta(x_n)]$$

$$H(f(x_n),\theta(x_n)))=\sum_{n}[\ln(1+\exp(-ys)]$$



5.2 逻辑斯蒂回归损失

逻辑斯蒂回归的最佳解：

$$g=\underset{w}{argmin}\frac{1}{N}\sum_{n=1}^{N}-ln\theta(y_n\mathbf{w}^T\mathbf{x}_n)$$

$$\theta(y_n\mathbf{w}^T\mathbf{x}_n)=\frac{1}{1+\exp(-y_n\mathbf{w}^T\mathbf{x}_n)}$$

$$g=\underset{w}{argmin}\frac{1}{N}\sum_{n=1}^{N}\ln(1+exp(-y_n\mathbf{w}^T\mathbf{x}_n))$$

交叉熵损失

(*Cross-Entropy Loss*)

$$\boxed{L_{in}=\frac{1}{N}\sum_{n=1}^{N}\ln(1+exp(-y_n\mathbf{w}^T\mathbf{x}_n))}$$

5.2 逻辑斯蒂回归损失

逻辑斯蒂回归的最佳解：

$$g = \underset{\mathbf{w}}{argmin} \frac{1}{N} \sum_{n=1}^{N} -ln\theta(y_n \mathbf{w}^T \mathbf{x}_n)$$

$$\theta(y_n \mathbf{w}^T \mathbf{x}_n) = \frac{1}{1 + \exp(-y_n \mathbf{w}^T \mathbf{x}_n)}$$

$$g = \underset{\mathbf{w}}{argmin} \frac{1}{N} \sum_{n=1}^{N} \ln(1 + exp(-y_n \mathbf{w}^T \mathbf{x}_n))$$

交叉熵损失
(Cross-Entropy Loss)

$$\boxed{L_{in} = \frac{1}{N} \sum_{n=1}^{N} \ln(1 + exp(-ys))}$$

令：$s = \mathbf{w}^T \mathbf{x}$

5.2 逻辑斯蒂回归损失

逻辑斯蒂回归的最佳解：

$$g = \underset{\mathbf{w}}{argmin} \frac{1}{N} \sum_{n=1}^{N} -ln\theta(y_n \mathbf{w}^T \mathbf{x}_n)$$

$$\theta(y_n \mathbf{w}^T \mathbf{x}_n) = \frac{1}{1 + \exp(-y_n \mathbf{w}^T \mathbf{x}_n)}$$

$$g = \underset{\mathbf{w}}{argmin} \frac{1}{N} \sum_{n=1}^{N} \ln(1 + exp(-y_n \mathbf{w}^T \mathbf{x}_n))$$

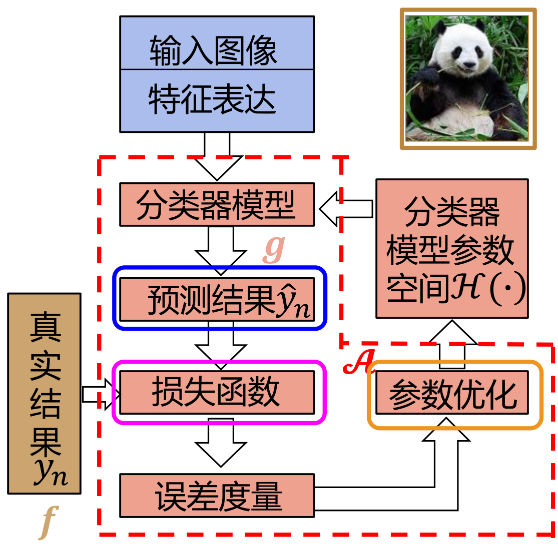
交叉熵损失
(Cross-Entropy Loss)

$$\boxed{L_{in} = \frac{1}{N} \sum_{n=1}^{N} \ln(1 + exp(-ys))}$$

$$H(f(x_n), \theta(x_n))) = \sum_{n} [\ln(1 + \exp(-ys)]$$

令：$s = \mathbf{w}^T \mathbf{x}$

5.2 逻辑斯蒂回归损失



感知器(线性分类)

$$\hat{y}_{n(t)} = \text{sign}(\boldsymbol{w}_t^T \mathbf{x}_{n(t)})$$

$$L_{in} = \sum_{n=1}^{N} [\![y_n \neq \hat{y}_n]\!]$$

线性回归

$$\hat{y}_n = \boldsymbol{w}^T \mathbf{x}_n$$

$$L_{in}(\boldsymbol{w}) = \frac{1}{N}\sum_{n=1}^{N} (\boldsymbol{w}^T \mathbf{x}_n - y_n)^2$$

逻辑斯蒂回归

$$\hat{y}_n = h(\mathbf{x}) = \theta(\boldsymbol{w}^T \mathbf{x}) = \frac{1}{1 + exp(-\boldsymbol{w}^T \mathbf{x})}$$

$$L_{in} = \frac{1}{N}\sum_{n=1}^{N} \ln(1 + exp(-y_n \boldsymbol{w}^T \mathbf{x}_n))$$

5.2 逻辑斯蒂回归损失

逻辑斯蒂回归的最佳解：

$$g = \underset{\boldsymbol{w}}{argmin} \frac{1}{N}\sum_{n=1}^{N} -ln\theta(y_n \boldsymbol{w}^T \mathbf{x}_n)$$

$$\theta(y_n \boldsymbol{w}^T \mathbf{x}_n) = \frac{1}{1 + \exp(-y_n \boldsymbol{w}^T \mathbf{x}_n)}$$

$$g = \underset{\boldsymbol{w}}{argmin} \frac{1}{N}\sum_{n=1}^{N} \ln(1 + exp(-y_n \boldsymbol{w}^T \mathbf{x}_n))$$

交叉熵损失

(*Cross-Entropy Loss*)

$$L_{in} = \frac{1}{N}\sum_{n=1}^{N} \ln(1 + exp(-y_n \boldsymbol{w}^T \mathbf{x}_n))$$

令： $s = \boldsymbol{w}^T \mathbf{x}$

$$L_{in} = \ln(1 + exp(-ys))$$

5.2 逻辑斯蒂回归损失

交叉熵损失梯度:

$$L_{in} = \ln(\underbrace{1 + exp(\overbrace{-y_n \mathbf{w}^T \mathbf{x}_n}^{\bullet}}_{\blacksquare}))$$

$$\frac{\partial L_{in}(\mathbf{w}, \mathbf{x}, y)}{\partial w_i} = \frac{\partial \ln(\blacksquare)}{\partial \blacksquare} \frac{\partial (1+\exp(\bullet))}{\partial \bullet} \frac{\partial (-y\mathbf{w}^T\mathbf{x})}{\partial w_i}$$

$$= \frac{1}{\blacksquare} \exp(\bullet)\,(-yx_i) = \frac{\exp(\bullet)}{1+\exp(\bullet)}(-yx_i)$$

$$\nabla L_{in}(\mathbf{w}, \mathbf{x}, y) = \theta(-y\mathbf{w}^T\mathbf{x})(-y\mathbf{x})$$

5.2 逻辑斯蒂回归损失

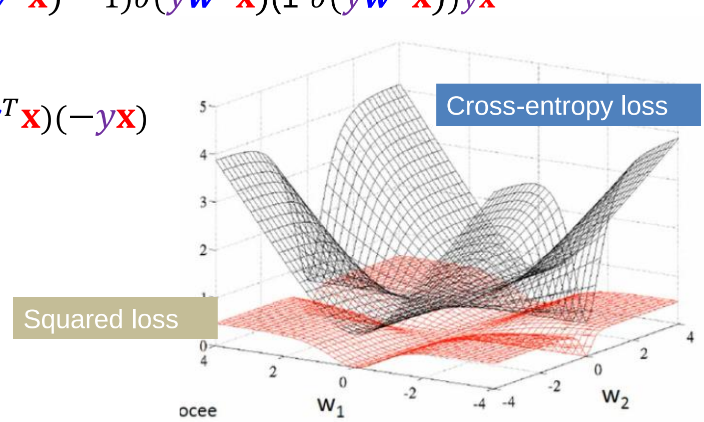
交叉熵损失与平方损失的梯度对比:

平方损失的梯度:

$$\nabla L_{in}(\mathbf{w}, \mathbf{x}, y) = 2(\theta(y\mathbf{w}^T\mathbf{x}) - 1)\theta(y\mathbf{w}^T\mathbf{x})(1\text{-}\theta(y\mathbf{w}^T\mathbf{x}))y\mathbf{x}$$

交叉熵损失的梯度:

$$\nabla L_{in}(\mathbf{w}, \mathbf{x}, y) = \theta(-y\mathbf{w}^T\mathbf{x})(-y\mathbf{x})$$



3.3 梯度下降法（引）

训练(Training)

输入图像

特征表达

回归模型

g

预测结果$\hat{y}_n$

损失函数

误差度量

回归模型参数空间 $\mathcal{H}(\cdot)$

$\mathcal{A}$ 参数优化

真实结果 y_n

f

随机梯度下降法(SGD):

$$\nabla L_{in}(\mathbf{w}) = \sum_{n=1}^{B} (\mathbf{w}^T\mathbf{x}_n - y_n)\mathbf{x}_n$$

$$\boldsymbol{m}_{i,t+1} = \lambda \boldsymbol{m}_{i,t} - \frac{\eta}{\sigma_{i,t}} \frac{\partial L_{in}}{\partial w_{i,t}}$$

$$\boldsymbol{w}_{i,t+1} \leftarrow \boldsymbol{w}_{i,t} + \boldsymbol{m}_{i,t+1}$$

- *问题1：学习率η如何取值？*
- *问题2：梯度为0就能得到最佳解？*
- *问题3：训练样本批量大小的影响？*
- *问题4：损失函数的影响？*

3.3 梯度下降法（引）

训练(Training)

输入图像

特征表达

回归模型

g

预测结果$\hat{y}_n$

损失函数

误差度量

回归模型参数空间 $\mathcal{H}(\cdot)$

$\mathcal{A}$ 参数优化

真实结果 y_n

f

随机梯度下降法(SGD):

$$\nabla L_{in}(\mathbf{w}) = \sum_{n=1}^{B} (\mathbf{w}^T\mathbf{x}_n - y_n)\mathbf{x}_n$$

$\boldsymbol{m}_{i,t+1} = \lambda \boldsymbol{m}_{i,t}$ …

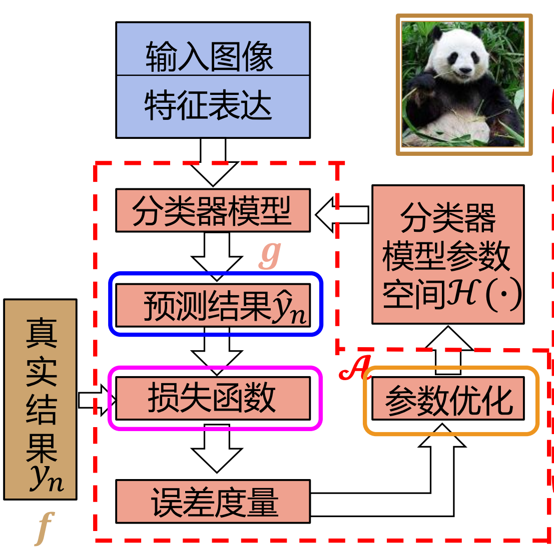
$\boldsymbol{w}_{i,t+1} \leftarrow \boldsymbol{w}$ …

不同损失函数在梯度下降时的速度不同

- *问题1：学习率…*
- *问题2：梯度为…佳解？*
- *问题3：训练样本批量大小的影响？*
- *问题4：损失函数的影响？*

第五讲　逻辑斯蒂回归（*Logistic Regression*）

5.2 逻辑斯蒂回归损失



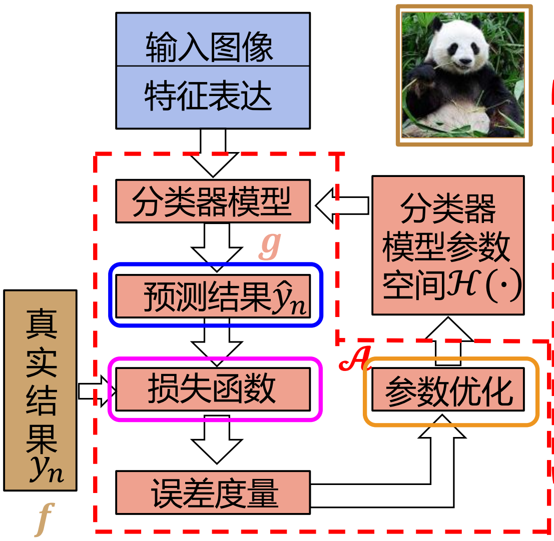
逻辑斯蒂回归

$$\hat{y}_n = h(\mathbf{x}) = \theta(\mathbf{w}^T\mathbf{x}) = \frac{1}{1+exp(-\mathbf{w}^T\mathbf{x})}$$

$$L_{in} = \frac{1}{N}\sum_{n=1}^{N}\ln(1+exp(-y_n\mathbf{w}^T\mathbf{x}_n))$$

$$\nabla L_{in}(\mathbf{w},\mathbf{x},y) = \theta(-y\mathbf{w}^T\mathbf{x})(-y\mathbf{x})$$

5.2 逻辑斯蒂回归损失



逻辑斯蒂回归

$$\hat{y}_n = h(\mathbf{x}) = \theta(\boldsymbol{w}^T\mathbf{x}) = \frac{1}{1 + exp(-\boldsymbol{w}^T\mathbf{x})}$$

$$L_{in} = \frac{1}{N}\sum_{n=1}^{N} \ln(1 + exp(-y_n \boldsymbol{w}^T \mathbf{x}_n))$$

$$\nabla L_{in}(\boldsymbol{w}, \mathbf{x}, y) = \frac{1}{N}\sum_{n=1}^{N} \theta(-y_n \boldsymbol{w}^T \mathbf{x}_n)(-y_n \mathbf{x}_n)$$

3.2 线性回归算法

梯度下降法实现逻辑斯蒂回归

- 初始化权向量$\boldsymbol{w}_0$
- $\boldsymbol{for}\ t = 0,1,2, \ldots$ (t 代表迭代次数)

  ① 计算梯度：$\nabla L_{in}(\boldsymbol{w}_t) = \frac{1}{N}\sum_{n=1}^{N} \theta(-y_n \boldsymbol{w}_t^T \mathbf{x}_n)(-y_n \mathbf{x}_n)$

  ② 对权向量$\boldsymbol{w}_t$进行更新：$\boldsymbol{w}_{t+1} \leftarrow \boldsymbol{w}_t - \eta \nabla L_{in}(\boldsymbol{w}_t)$

…直到$\nabla L_{in}(\boldsymbol{w}) = \boldsymbol{0}$，或者迭代足够多次数

返回最终的$\boldsymbol{w}_{t+1}$作为学到的g

3.2 线性回归算法

梯度下降法实现逻辑斯蒂回归

Stochastic Gradient Descent(SGD)

- 初始化权向量w_0
- $for\ t = 0,1,2,\ldots$ (t 代表迭代次数)

① 计算梯度：$\nabla L_{in}(\mathbf{w}_t) = \frac{1}{B}\sum_{n=1}^{B}\theta(-y_n \mathbf{w}_t^T \mathbf{x}_n)(-y_n \mathbf{x}_n)$

② 对权向量$\mathbf{w}_t$进行更新：$\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t - \eta\nabla L_{in}(\mathbf{w}_t)$

…直到$\nabla L_{in}(\mathbf{w}) = \mathbf{0}$，或者迭代足够多次数

返回最终的$\mathbf{w}_{t+1}$作为学到的g

第五讲 逻辑斯蒂回归（*Logistic Regression*）

5.1 逻辑斯蒂回归问题（*Logistic Regression Problem*）

5.2 逻辑斯蒂回归损失（*Logistic Regression Loss*）

5.3 逻辑斯蒂回归算法（*Logistic Regression Algorithm*）

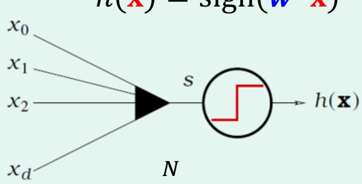
5.4 二元分类线性模型讨论（*Linear Models for Binary Classification*）

5.4 二元分类线性模型讨论

三种线性模型比较

线性分类(感知器):

$$h(\mathbf{x}) = \text{sign}(\boldsymbol{w}^T\mathbf{x})$$

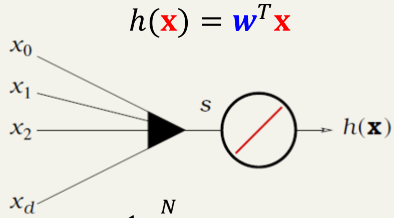


$$L_{in} = \sum_{n=1}^{N} [\![y_n \neq \hat{y}_n]\!]$$

NP难问题

线性回归:

$$h(\mathbf{x}) = \boldsymbol{w}^T\mathbf{x}$$

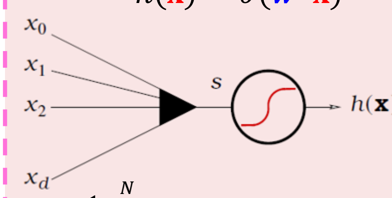


$$L_{in} = \frac{1}{N}\sum_{n=1}^{N} (\boldsymbol{w}^T\mathbf{x}_n - y_n)^2$$

凸函数、易优化、解析解

逻辑斯蒂回归:

$$h(\mathbf{x}) = \theta(\boldsymbol{w}^T\mathbf{x})$$



$$L_{in} = \frac{1}{N}\sum_{n=1}^{N} \ln(1 + exp(-y_n\boldsymbol{w}^T\mathbf{x}_n))$$

凸函数、平滑、梯度下降

5.4 二元分类线性模型讨论

三种线性模型用于二元分类时(即：$y \in \{+1, -1\}$)损失函数比较

样本特征向量 $\mathbf{x}$ 与模型的权向量 $\boldsymbol{w}$ 的内积用 s 表示： $s = \boldsymbol{w}^T\mathbf{x}$

线性分类(感知器):

$$h(\mathbf{x}) = \text{sign}(s)$$

$$L_{in} = [\![h(\mathbf{x}) \neq y]\!]$$

$$L_{0/1}(s, y) = [\![\text{sign}(s) \neq y]\!] = [\![\text{sign}(ys) \neq 1]\!]$$

线性回归:

$$h(\mathbf{x}) = s$$

$$L_{in} = (h(\mathbf{x}) - y)^2$$

$$L_{sqr}(s, y) = (s - y)^2 = (ys - 1)^2$$

逻辑斯蒂回归:

$$h(\mathbf{x}) = \theta(s)$$

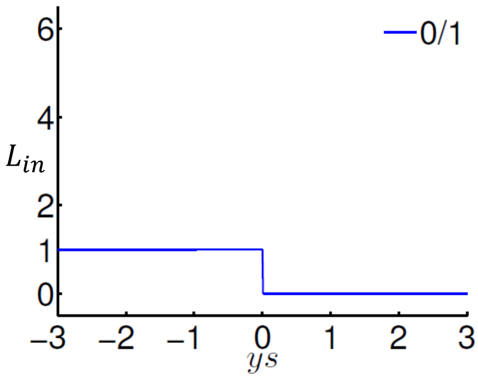
$$L_{in} = -\ln(h(y\mathbf{x}))$$

$$L_{ce} = \ln(1 + exp(-ys))$$

5.4 二元分类线性模型讨论

三种线性模型用于二元分类时(即： $y \in \{+1, -1\}$)损失函数比较

样本特征向量 $\mathbf{x}$ 与模型的权向量 $\boldsymbol{w}$ 的内积用 s 表示： $s = \boldsymbol{w}^T\mathbf{x}$

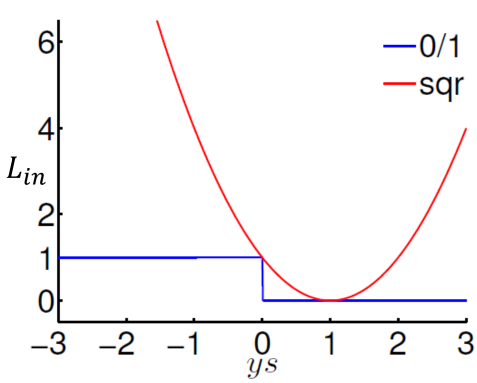


0/1 $L_{0/1}(s, y) = [\![sign(ys) \neq 1]\!]$

5.4 二元分类线性模型讨论

三种线性模型用于二元分类时(即： $y \in \{+1, -1\}$)损失函数比较

样本特征向量 $\mathbf{x}$ 与模型的权向量 $\boldsymbol{w}$ 的内积用 s 表示： $s = \boldsymbol{w}^T\mathbf{x}$



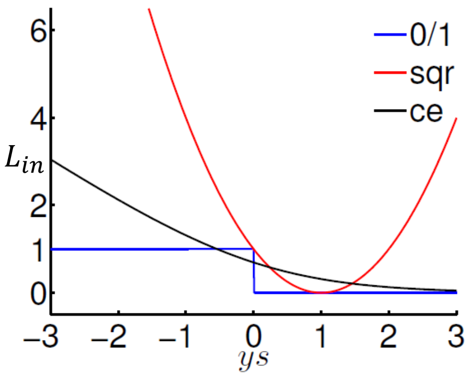
0/1 $L_{0/1}(s, y) = [\![sign(ys) \neq 1]\!]$

sqr $L_{sqr}(s, y) = (ys - 1)^2$

5. 4 二元分类线性模型讨论

三种线性模型用于二元分类时(即： $y \in \{+1, -1\}$)损失函数比较

样本特征向量 $\mathbf{x}$ 与模型的权向量 $\boldsymbol{w}$ 的内积用 s 表示： $s = \boldsymbol{w}^T\mathbf{x}$



0/1 $L_{0/1}(s, y) = [\![sign(ys) \neq 1]\!]$

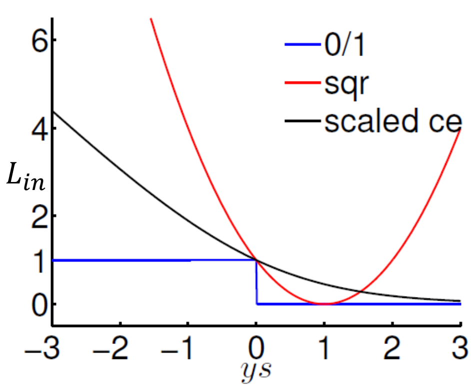
sqr $L_{sqr}(s, y) = (ys - 1)^2$

ce $L_{ce}\ (s, y) = ln(1 + \exp(-ys))$

5. 4 二元分类线性模型讨论

三种线性模型用于二元分类时(即： $y \in \{+1, -1\}$)损失函数比较

样本特征向量 $\mathbf{x}$ 与模型的权向量 $\boldsymbol{w}$ 的内积用 s 表示： $s = \boldsymbol{w}^T\mathbf{x}$



0/1 $L_{0/1}(s, y) = [\![sign(ys) \neq 1]\!]$

sqr $L_{sqr}(s, y) = (ys - 1)^2$

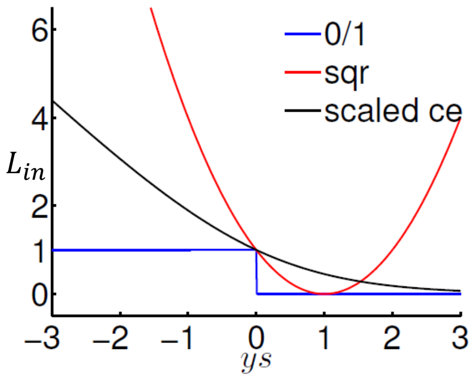
ce $L_{ce}\ (s, y) = ln(1 + \exp(-ys))$

Scaled ce $L_{sce}\ (s, y) = log_2(1 + \exp(-ys))$

5.4 二元分类线性模型讨论

三种线性模型用于二元分类时(即：$y \in \{+1,-1\}$)损失函数比较

样本特征向量 $\mathbf{x}$ 与模型的权向量 $\boldsymbol{w}$ 的内积用 s 表示： $s = \boldsymbol{w}^T\mathbf{x}$



$L_{0/1}(s,y) \le L_{sqr}(s,y)$

$L_{0/1}(s,y) \le L_{sce}(s,y)$

$L_{0/1}(s,y) \le L_{ce}(s,y)$

训练或测试时，只要做到 $L_{sqr}(s,y)$或者 $L_{ce}(s,y)$ 很小， $L_{0/1}(s,y)$也会很小

线性回归与逻辑斯蒂回归可用于线性分类

5.4 二元分类线性模型讨论

① 在标签为{+1，-1}的训练样本集 $\mathcal{D}$ 上运行线性回归/逻辑斯蒂回归算法，得到$\boldsymbol{w}^*$

② 返回分类结果：$g(\mathbf{x}) = sign({\boldsymbol{w}^*}^T\mathbf{x})$

线性分类(感知器)：

优点：样本线性可分时，算法收敛有理论保障

不足：样本非线性可分时NP难问题，可用Pocket算法实现

线性回归：

优点：凸函数，最容易优化，有解析解

不足：当$|ys|$很大时，$L_{0/1}(s,y)$的上界过于宽松

逻辑斯蒂回归：

优点：凸函数，易于优化

不足：当$ys \ll 0$时，$L_{0/1}(s,y)$的上界过于宽松

第五讲 逻辑斯蒂回归（*Logistic Regression*）

5.1 逻辑斯蒂回归问题

模型的理论输出为概率值，分类面假设空间模型用Sigmoid函数

5.2 逻辑斯蒂回归损失

用交叉熵(cross-entropy)作为损失函数

5.3 逻辑斯蒂回归算法

用梯度下降法迭代实现参数更新

5.4 二元分类线性模型讨论

三个线性模型的特点及用途

Taylor公式

$$f(x)=f(x_0)+\frac{f'(x_0)}{1!}(x-x_0)+\frac{f''(x_0)}{2!}(x-x_0)^2+\cdots+\frac{f^{(n)}(x_0)}{n!}(x-x_0)^n+R_n$$

$$R_n(x)=o[(x-x_0)^n]$$

$$f(x)=f(0)+\frac{f'(0)}{1!}x+\frac{f''(0)}{2!}x^2+\cdots+\frac{f^{(n)}(0)}{n!}x^n+R_n$$

$$e^x=1+\frac{1}{1!}x+\frac{1}{2!}x^2+\frac{1}{3!}x^3+o\left(x^3\right)$$

$$\ln(1+x)=x-\frac{1}{2}x^2+\frac{1}{3}x^3+o\left(x^3\right)$$

$$\sin x=x-\frac{1}{3!}x^3+\frac{1}{5!}x^5+o\left(x^5\right)$$

$$\arcsin x=x+\frac{1}{2}\times\frac{x^3}{3}+\frac{1\times 3}{2\times 4}\times\frac{x^5}{5}+\frac{1\times 3\times 5}{2\times 4\times 6}\times\frac{x^7}{7}+o\left(x^7\right)$$

$$\cos x=1-\frac{1}{2!}x^2+\frac{1}{4!}x^4+o\left(x^4\right)$$

$$\frac{1}{1-x}=1+x+x^2+x^3+o\left(x^3\right)$$