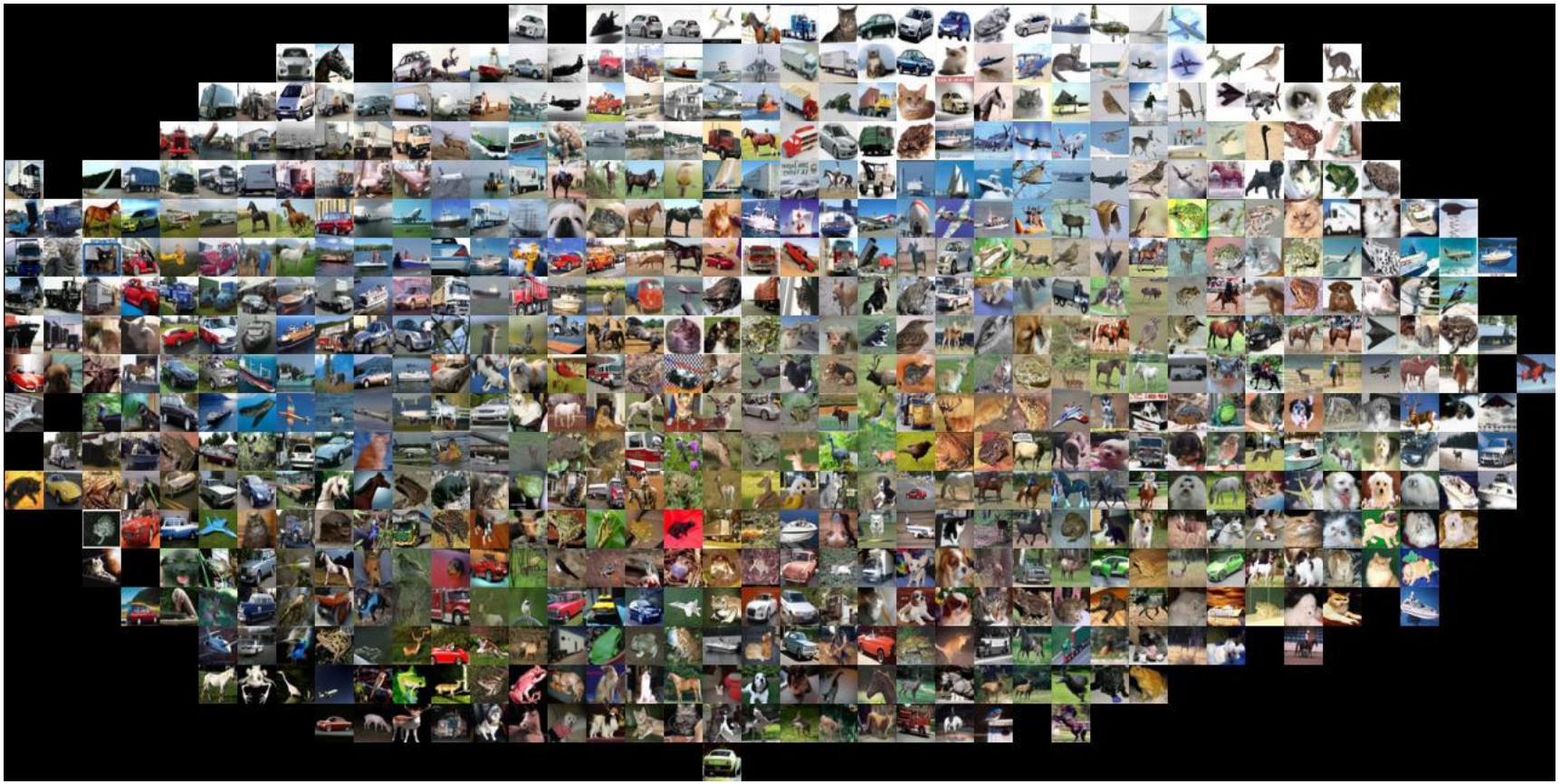
分类的线性模型
LINEAR MODELS FOR CLASSIFICATION

张玲玲
计算机学院
zhanglling@xjtu.edu.cn

主要内容

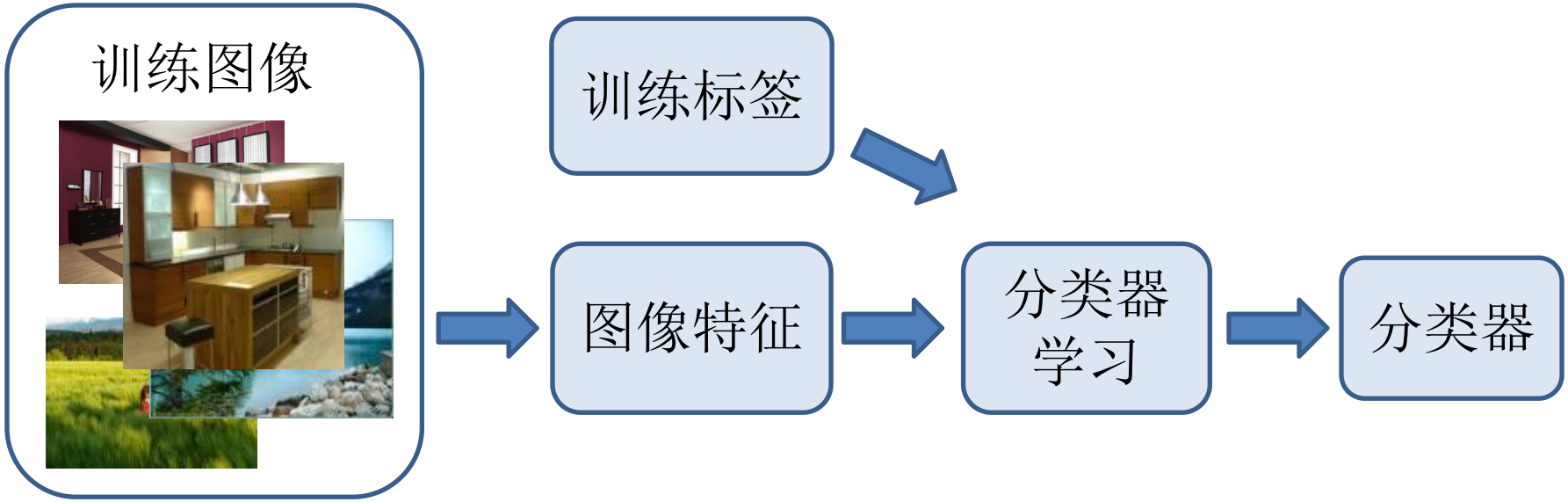
1. 分类举例：图像分类
2. 广义线性模型
3. 判别函数（Discriminant Functions）
4. 生成模型（Generative Models）
5. 判别模型（Discriminative Models）

分类举例：图像分类



分类举例：图像分类

训练/学习



分类举例：图像分类

- 分类的要素：
  1. 机器学习方法，例如：线性分类，深度学习
  2. 图像表征，例如：SIFT，HoG，深度特征
  3. 数据，例如：PASCAL，ImageNet，COCO，Objects-365

分类举例：图像分类

- 假设给定一组离散标签，例如：

  {cat, dog, cow, apple, tomato, truck, … }

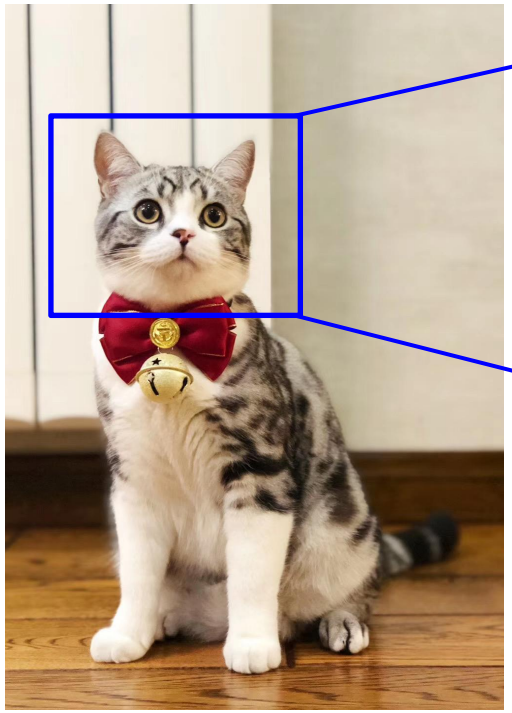
f() = “apple”

f() = “tomato”

f() = “cow”

分类举例：图像分类

- 问题：语义鸿沟



193	194	190	184	169	144	128	89	60	60	109	44	40	46	45	45	58	61	72	50
191	195	192	173	134	114	121	116	64	77	60	41	41	43	41	41	46	60	67	57
187	196	178	139	110	113	112	132	126	61	70	55	45	42	40	37	39	53	50	59
170	186	151	122	114	117	114	131	139	76	83	74	52	45	45	41	43	46	39	47
147	163	139	131	132	121	125	143	132	78	64	64	42	33	35	30	32	36	33	43
129	126	132	148	134	136	141	133	121	81	72	67	49	30	24	21	25	30	32	34
126	101	106	146	149	132	138	134	101	80	65	62	53	37	27	28	21	28	39	40
137	117	103	130	141	118	119	99	83	74	66	60	52	42	30	27	21	29	39	33
141	115	97	103	82	79	84	80	79	74	69	64	52	45	26	31	25	25	29	35
105	99	64	67	70	71	78	83	83	79	80	72	57	46	44	65	29	18	21	27
63	60	52	56	65	75	86	92	87	82	83	81	74	53	62	52	27	26	27	25
55	37	35	46	56	64	71	82	83	85	82	73	62	54	58	30	27	25	22	24

计算机看到的

分类举例：图像分类

训练/学习

训练图像

训练标签

图像特征

分类器学习

分类器

测试

测试图像

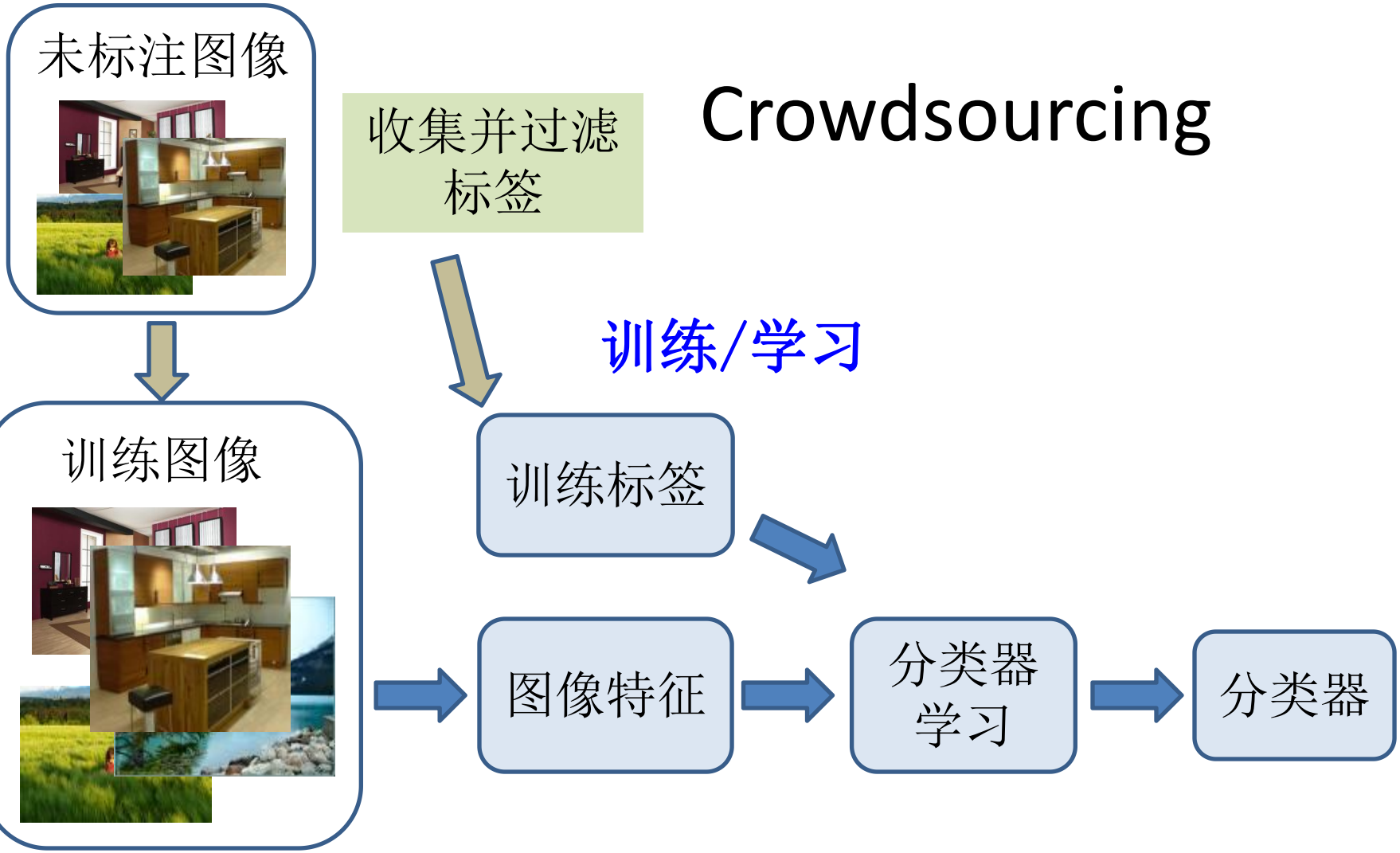
图像特征

分类器

预测

Outdoor

分类举例：图像分类

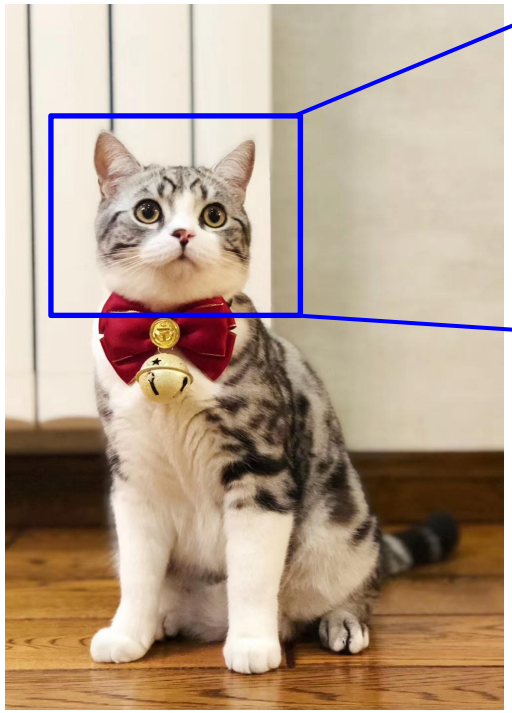


分类举例：图像分类

- 假设给定离散标签{dog, cat, truck, plane}

cat

分类举例：图像分类



193	194	190	184	169	144	128	89	60	60	109	44	40	46	45	45	58	61	72	50
191	195	192	173	134	114	121	116	64	77	60	41	41	43	41	41	46	60	67	57
187	196	178	139	110	113	112	132	126	61	70	55	45	42	40	37	39	53	50	59
170	186	151	122	114	117	114	131	139	76	83	74	52	45	45	41	43	46	39	47
147	163	139	131	132	121	125	143	132	78	64	64	42	33	35	30	32	36	33	43
129	126	132	148	134	136	141	133	121	81	72	67	49	30	24	21	25	30	32	34
126	101	106	146	149	132	138	134	101	80	65	62	53	37	27	28	21	28	39	40
137	117	103	130	141	118	119	99	83	74	66	60	52	42	30	27	21	29	39	33
141	115	97	103	82	79	84	80	79	74	69	64	52	45	26	31	25	25	29	35
105	99	64	67	70	71	78	83	83	79	80	72	57	46	44	65	29	18	21	27
63	60	52	56	65	75	86	92	87	82	83	81	74	53	62	52	27	26	27	25
55	37	35	46	56	64	71	82	83	85	82	73	62	54	58	30	27	25	22	24

计算机看到的

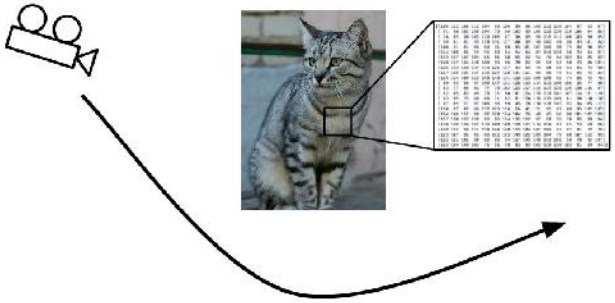
图像分类

85% cat
15% dog
2% hat
1% mug

分类举例：图像分类

- 图像分类面临的挑战

Viewpoint



Illumination

This image is CC0 1.0 public domain

Deformation

This image by Umberto Salvagnin is licensed under CC-BY 2.0

Occlusion

This image by jonsson is licensed under CC-BY 2.0

Clutter

This image is CC0 1.0 public domain

Intraclass Variation

This image is CC0 1.0 public domain

分类举例：图像分类

- 图像分类器

```
def classify_image(image):
  # Some magic here?
  return class_label
```

输入：图像

分类器

输出：类别标签

分类举例：图像分类

- 收集带有标签的图像数据集
- 使用机器学习方法训练图像分类器
- 在测试图像上评测学到的分类器

```
def train(train_images, train_labels):
    # build a model of images -> labels

def predict(image):
    # evaluate the model on the image
    return class_label
```

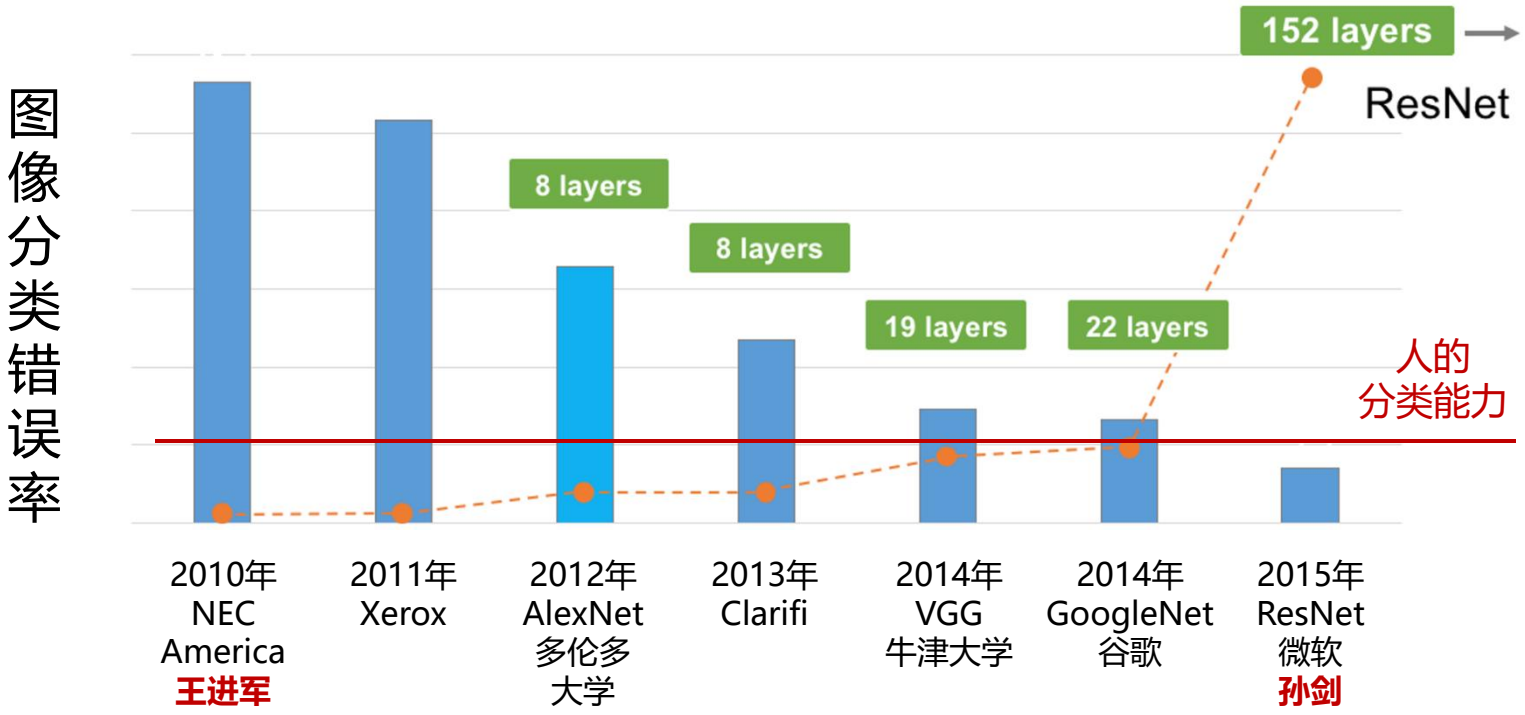
分类举例：图像分类

- 分类器
  - 最近邻算法（Neareast Neighbor）
  - kNN（ k-Neareast Neighbor ）
  - 线性分类器（Linear Classifier）
  - 神经网络（Neural Network）
  - 深度神经网络（Deep Neural Network）
  - …

无监督

有监督

分类举例：图像分类


图像分类错误率
152 layers
ResNet
8 layers
8 layers
19 layers
22 layers
人的
分类能力
2010年
NEC
America
王进军
2011年
Xerox
2012年
AlexNet
多伦多
大学
2013年
Clarifi
2014年
VGG
牛津大学
2014年
GoogleNet
谷歌
2015年
ResNet
微软
孙剑

分类举例：图像分类

- 使用线性分类器进行图像分类

- 需要：
  - 评分函数 (Score function)：原始数据到类别分数
  - 损失函数 (Loss function)：预测分数与真值标签之间的一致性

分类举例：图像分类

- 评分函数（Score function）

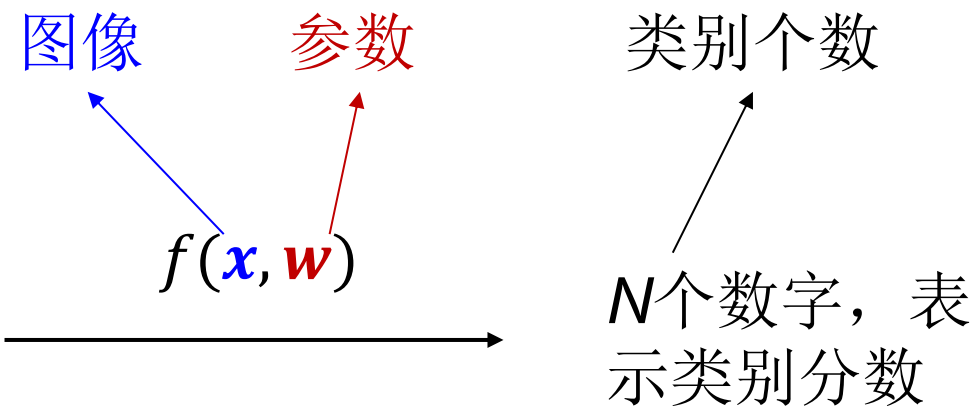
类别分数

85% cat
15% dog
2% hat
1% mug

分类举例：图像分类

- 评分函数（Score function）：f

[32 × 32 × 3]
数字数组
（总共3072个数字）



分类举例：图像分类

- 评分函数（Score function）：线性分类器f

$[32 \times 32 \times 3]$
数字数组
（总共3072个数字）

$N \times 1$　　$N \times 3072$　　3072×1

$f(x, w) = w^T x$

N个数字，表示类别分数

参数，或权重

分类举例：图像分类

- 评分函数（Score function）：线性分类器f

$[32 \times 32 \times 3]$
数字数组
（总共3072个数字）

$N \times 1$　$N \times 3072$　3072×1　$N \times 1$

$$f(\boldsymbol{x}, \boldsymbol{W}, w_0) = \boldsymbol{W}^{\mathrm{T}}\boldsymbol{x} + w_0$$

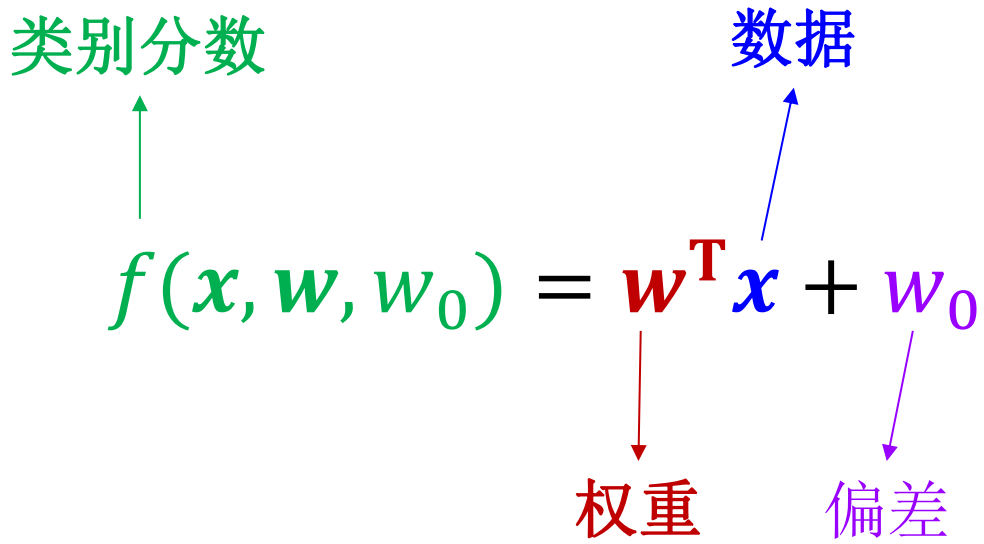
N个数字，表示类别分数

偏差（参数）

参数，或权重

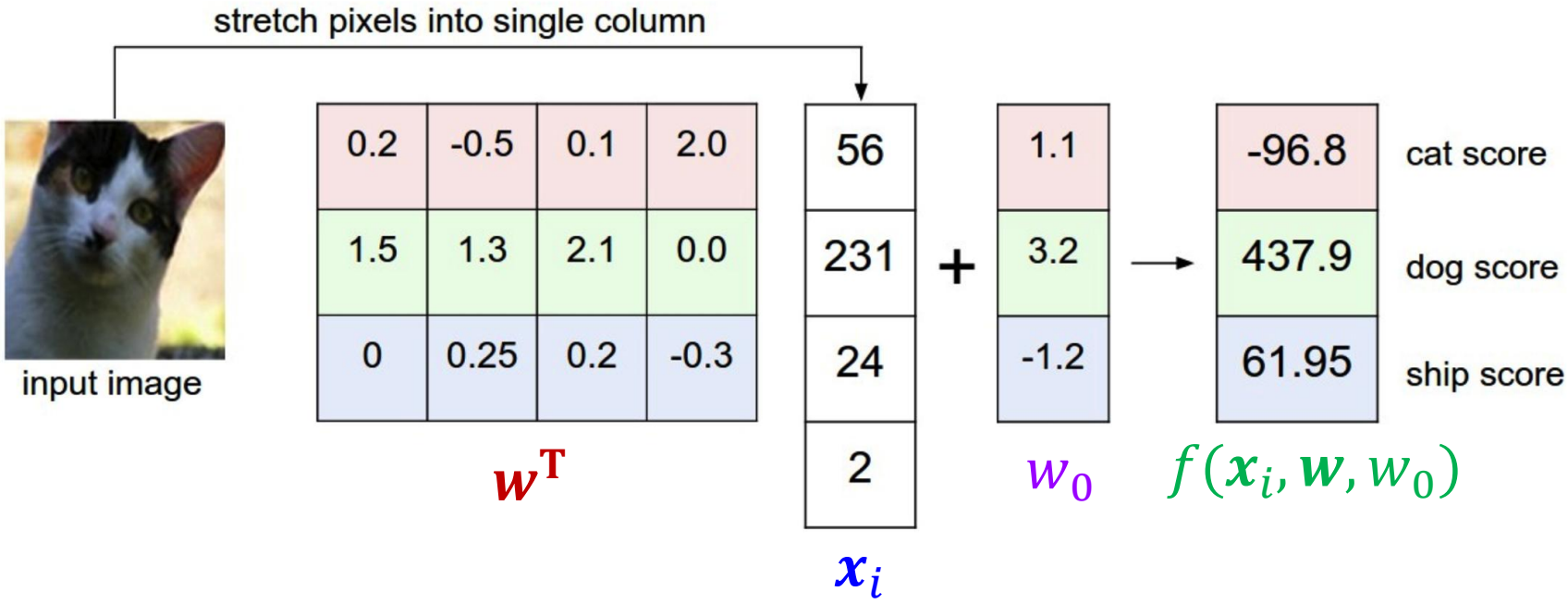
分类举例：图像分类

- 评分函数（Score function）：线性分类器f



分类举例：图像分类

- 一个只有4个像素的图像，3个类别（cat/dog/ship）



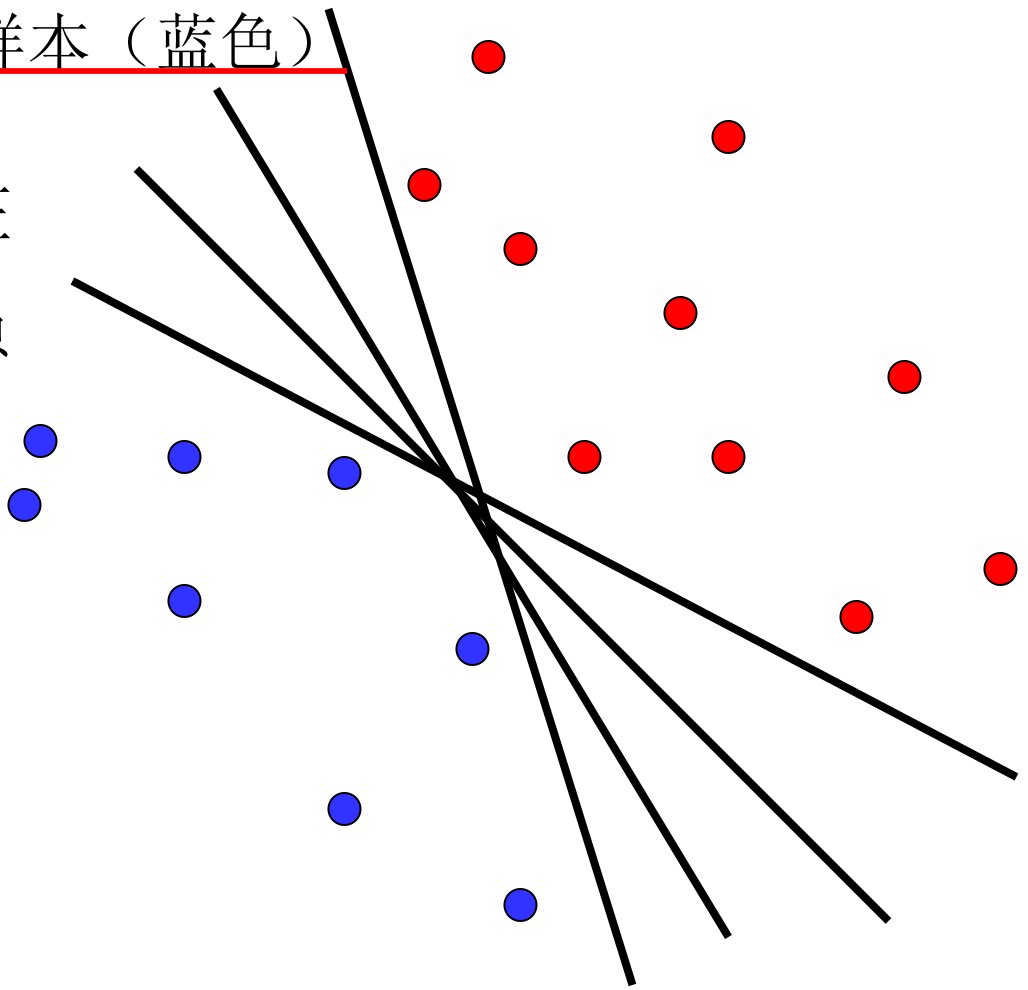
$$f(x, w, w_0) = w^T x + w_0$$

分类举例：图像分类

- 线性分类器：寻找一个线性函数（超平面）以分离开正样本（红色）和负样本（蓝色）

$\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + w_0 \geq 0$，　$\boldsymbol{x}_i$为正

$\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + w_0 < 0$，　$\boldsymbol{x}_i$为负



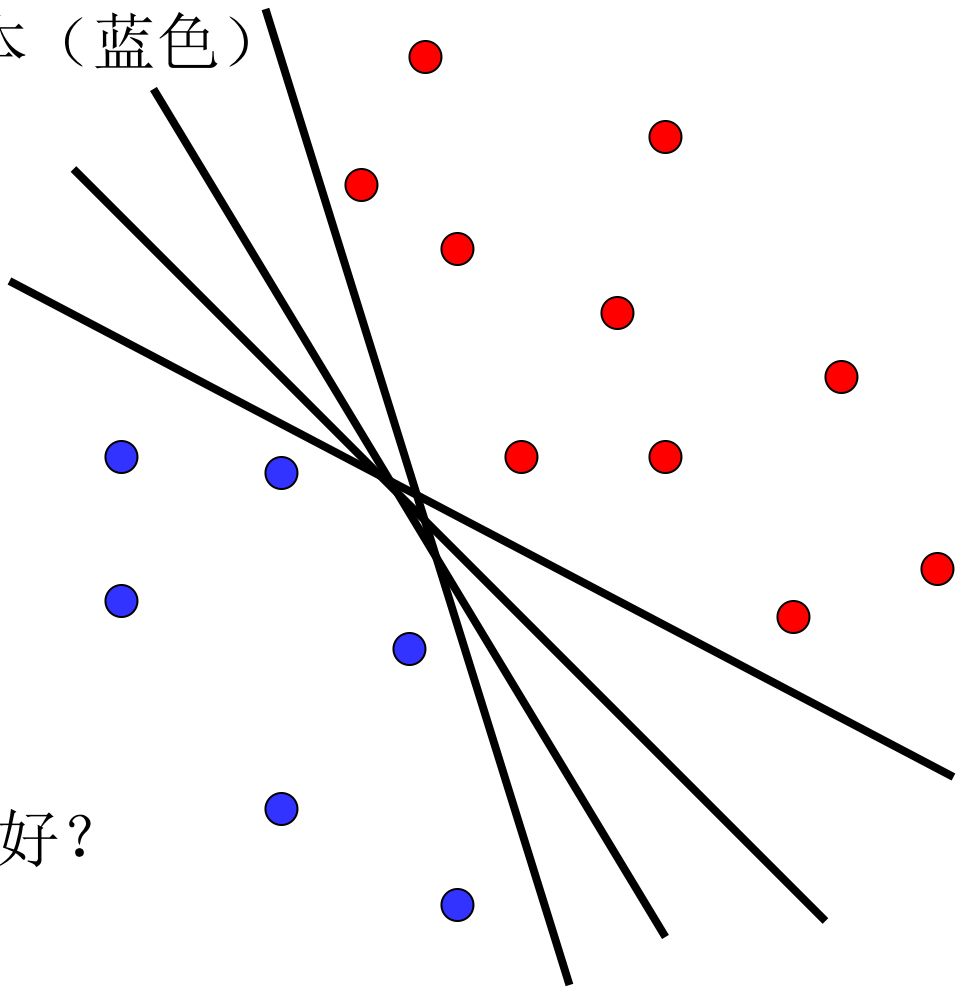
分类举例：图像分类

- 线性分类器：寻找一个线性函数（超平面）以分离开正样本（红色）和负样本（蓝色）

$\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + w_0 \geq 0$， $\boldsymbol{x}_i$为正

$\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + w_0 < 0$， $\boldsymbol{x}_i$为负

损失函数



哪个分类器（超平面）最好？

分类举例：图像分类

- 损失函数（Loss function）

  成本/目标函数（Cost/Objective function）
- 给定真值标签y_i，和类别分数$f(\boldsymbol{x}_i, \boldsymbol{W}, w_0)$
  - 评价对类别分数有多么“不满意”？

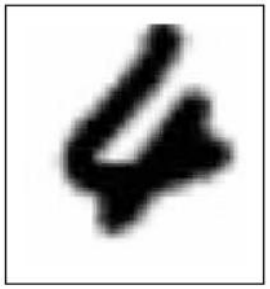
- 损失函数就是用来度量这种“不满意”的

- 在训练期间，要找到使损失函数最小的参数$\boldsymbol{W}, w_0$

主要内容

1. 分类举例：图像分类
2. 广义线性模型
3. 判别函数（Discriminant Functions）
4. 生成模型（Generative Models）
5. 判别模型（Discriminative Models）

分类：手写字符识别

$\boldsymbol{x}_i =$ 

28×28

$\boldsymbol{t}_i = (0,0,0,1,0,0,0,0,0,0)^{\mathrm{T}}$

10×1

- 每个输入向量被分为K个离散类别中的某一类。
  - 由$\mathcal{C}_k$表示类别
- 将输入图像表征为一个向量$\boldsymbol{x_i} \in \mathbb{R}^{784}$。
- 输出的目标向量是 $\boldsymbol{t}_i \in \{0,1\}^{10}$ 。
- 给定一个训练集 $\{(\boldsymbol{x}_1, \boldsymbol{t}_1), \ldots, (\boldsymbol{x}_N, \boldsymbol{t}_N)\}$，学习问题就是从训练集上构建一个“好”函数 $\boldsymbol{y}$，能对输入数据正确分类。
  - $\boldsymbol{y} : \mathbb{R}^{784} \rightarrow \mathbb{R}^{10}$

广义线性模型

- 与回归的线性模型一样，可以使用“线性”模型分类：

$$y(\boldsymbol{x}) = f(\boldsymbol{w}^T\boldsymbol{x} + w_0)$$

这个模型称作广义线性模型。

- $f(\cdot)$是固定的非线性函数，在机器学习中称作激活函数。
  - 例如：

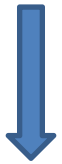
$$f(u) = \begin{cases} 1 \text{ if } u \geq 0 \\ 0 \text{ otherwise} \end{cases}$$

- 决策边界$y(\boldsymbol{x})$是$\boldsymbol{x}$的线性函数，因为$y(\boldsymbol{x})$的取值是[0,1]的常数。
- 决策边界可视作D维输入空间中的$D-1$维超平面。可以通过线性决策边界分类的数据集是线性可分的。

广义线性模型

- 我们也可以对输入变量$\boldsymbol{x}$进行非线性变换（特征提取），例如使用回归模型中的基函数$\boldsymbol{\phi}(\boldsymbol{x})$，这样就可以用来解决$\boldsymbol{x}$的非线性问题。

$$y(\boldsymbol{x}) = f(\boldsymbol{w}^T \boldsymbol{x} + w_0)$$



$$y(\boldsymbol{x}) = f(\boldsymbol{w}^T \boldsymbol{\phi}(x) + w_0)$$

广义线性模型

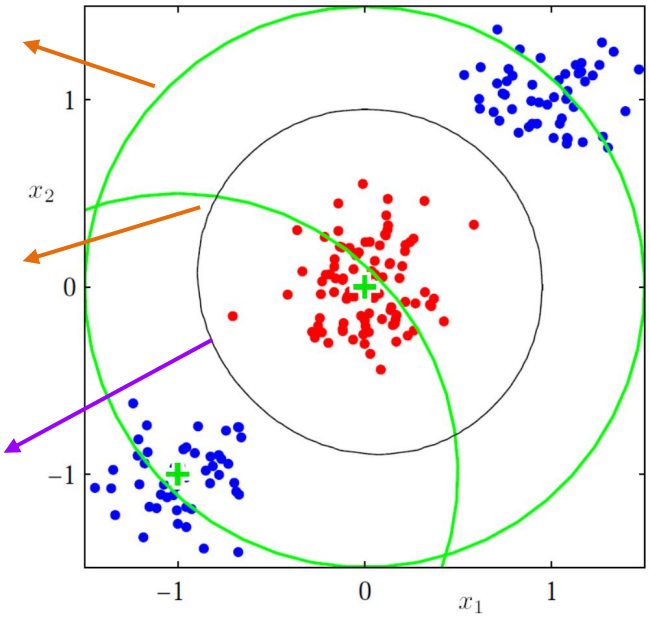
$$y(\boldsymbol{x}) = f(\boldsymbol{w}^T \boldsymbol{\phi}(x) + w_0)$$

$$y\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right) = f\left(\begin{bmatrix} w_1 & w_2 \end{bmatrix} \begin{bmatrix} \phi_1(x_1, x_2) \\ \phi_2(x_1, x_2) \end{bmatrix} + w_0\right)$$

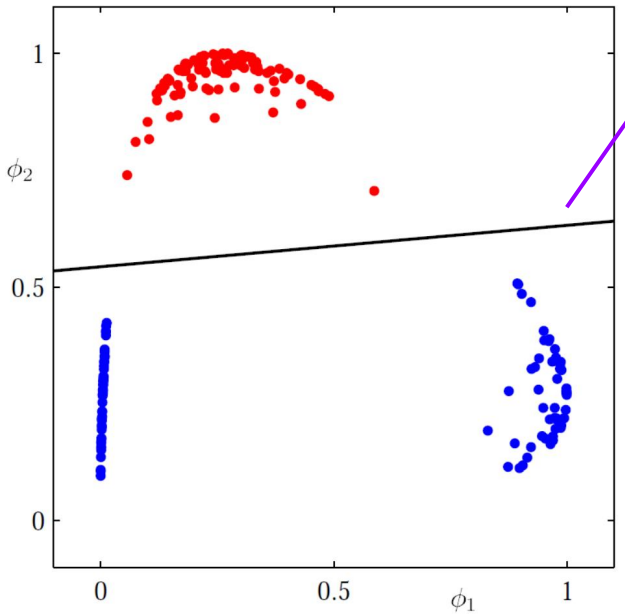
$\phi_1(x_1, x_2)$

$\phi_2(x_1, x_2)$

非线性
决策边界



(x_1, x_2)表示的输入空间



线性
决策边界

(ϕ_1, ϕ_2)表示的特征空间

主要内容

1. 分类举例：图像分类
2. 广义线性模型
3. 判别函数（Discriminant Functions）
4. 生成模型（Generative Models）
5. 判别模型（Discriminative Models）

两类判别函数

- 从两类问题开始，也就是目标值$t_i \in \{0,1\}$。
- 使用最简单的线性判别函数

$$y(\boldsymbol{x}) = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + w_0$$

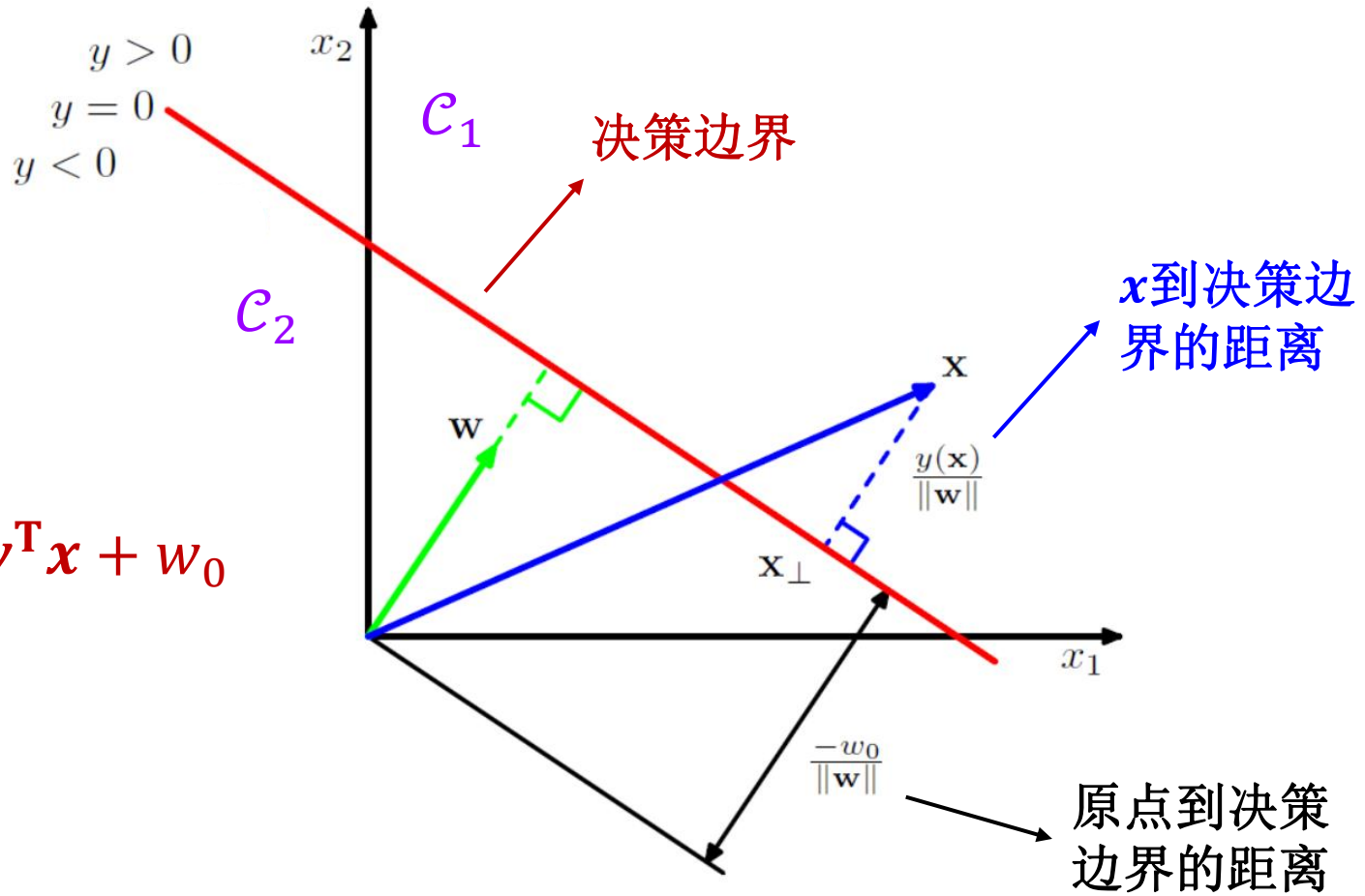
其中$\boldsymbol{w}$是权重向量，w_0是偏差，$-w_0$通常称作阈值。

- 如果$y(\boldsymbol{x}) > 0$，$\boldsymbol{x}$被分为类别$\mathcal{C}_1$；
- 如果$y(\boldsymbol{x}) < 0$，$\boldsymbol{x}$被分为类别$\mathcal{C}_2$；
- $y(\boldsymbol{x}) = 0$表示决策边界，由于决策边界可视作D维输入空间中的$D-1$维超平面，这里是一维超平面。

两类判别函数

- 对于决策边界上的两个点$\boldsymbol{x}_A$和$\boldsymbol{x}_B$，由于$y(\boldsymbol{x}_A) = y(\boldsymbol{x}_B) = 0$，有$\boldsymbol{w}^{\mathrm{T}}(\boldsymbol{x}_A - \boldsymbol{x}_B) = 0$。
- 因此，参数向量$\boldsymbol{w}$与决策边界上的每个向量正交，$\boldsymbol{w}$决定了决策边界的方向。
- $\boldsymbol{x}$与决策边界的有符号正交距离是$\frac{y(\boldsymbol{x})}{\|\boldsymbol{w}\|}$，$\|\boldsymbol{w}\|$是$L_2$范数。
- 如果$\boldsymbol{x}$在决策边界上，那么$y(\boldsymbol{x}) = 0$。因此，从原点到决策面的距离是$-\frac{w_0}{\|\boldsymbol{w}\|}$。
- 因此，偏差w_0决定了决策边界的位置。

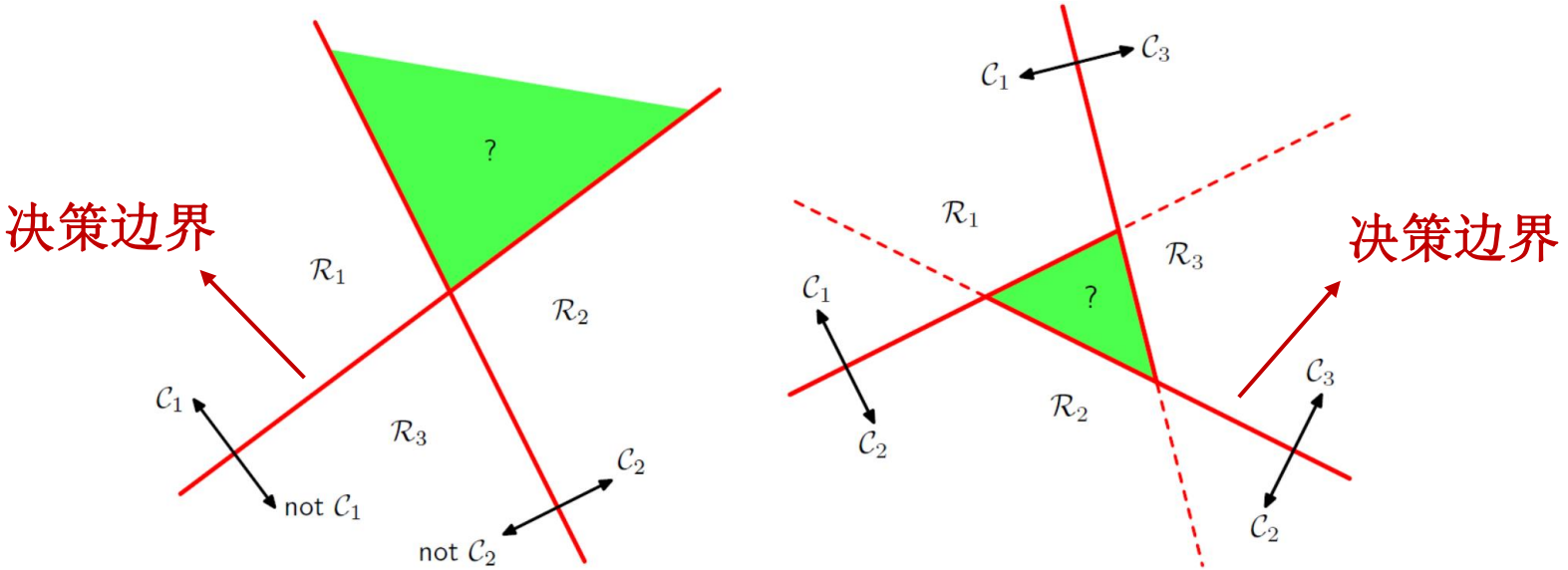
两类判别函数



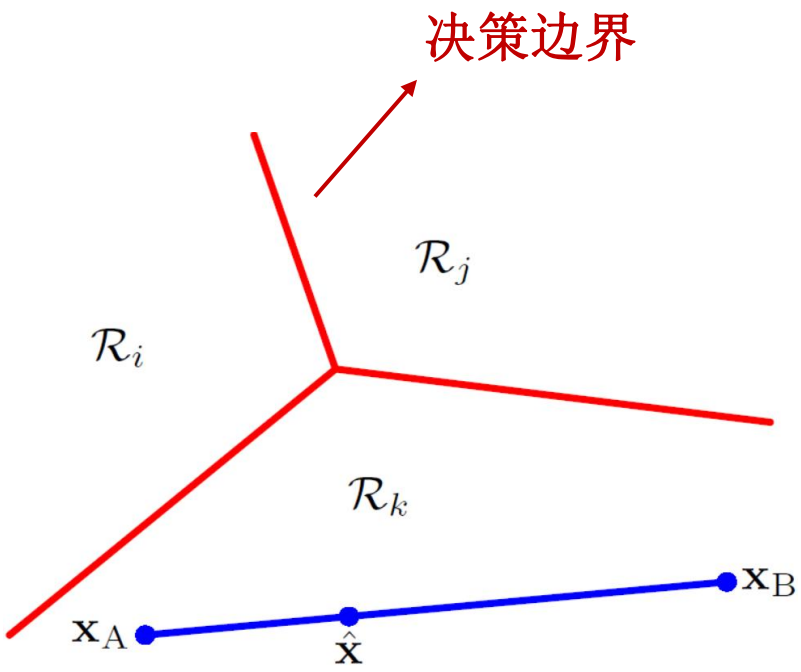
$$y(\boldsymbol{x}) = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + w_0$$

多类分类

- 两类之间的线性判别函数可以将一个超平面分开。
- 如何使用两类线性判别函数用来进行多类分类？
  - One-versus-the-rest方法：在$\mathcal{C}_k$类和其它类之间构建$K-1$个分类器
  - One-versus-one方法：在所有类别之间构建$K(K-1)/2$个分类器



多类分类



- 一种解决方案是构建一个包含K个线性函数的K类判别函数：

$$y_k(\boldsymbol{x}) = \boldsymbol{w}_k{}^{\mathrm{T}}\boldsymbol{x} + w_{k0}$$

将$\boldsymbol{x}$分配给分类值最大的类别，即$\mathrm{argmax}_k y_k(\boldsymbol{x})$。

- 决策区域始终是连通且凸的。

$$\hat{\boldsymbol{x}} = \lambda \boldsymbol{x}_A + (1-\lambda)\boldsymbol{x}_B$$

$$0 \leqslant \lambda \leqslant 1$$

$$y_k(\hat{\boldsymbol{x}}) = \lambda y_k(\boldsymbol{x}_A) + (1-\lambda) y_k(\boldsymbol{x}_B)$$

$$\Rightarrow y_k(\hat{\boldsymbol{x}}) > y_j(\hat{\boldsymbol{x}}), \forall j \neq k$$

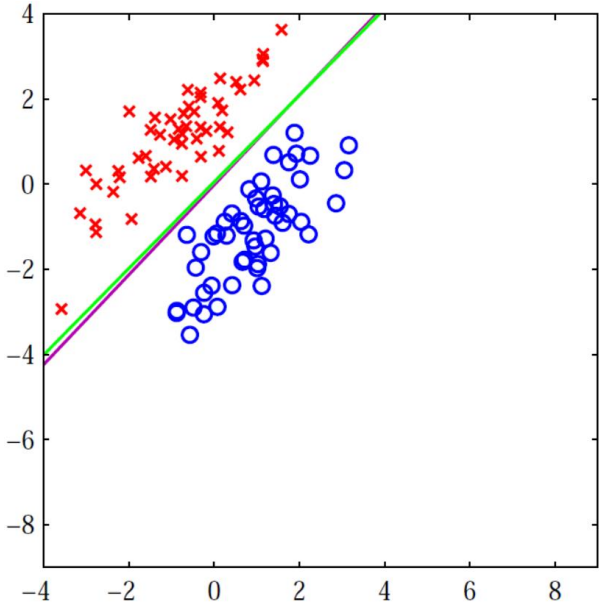
$$\begin{aligned} y_k(\mathbf{x}_\mathrm{A}) &> y_j(\mathbf{x}_\mathrm{A}) \\ y_k(\mathbf{x}_\mathrm{B}) &> y_j(\mathbf{x}_\mathrm{B}) \end{aligned}$$

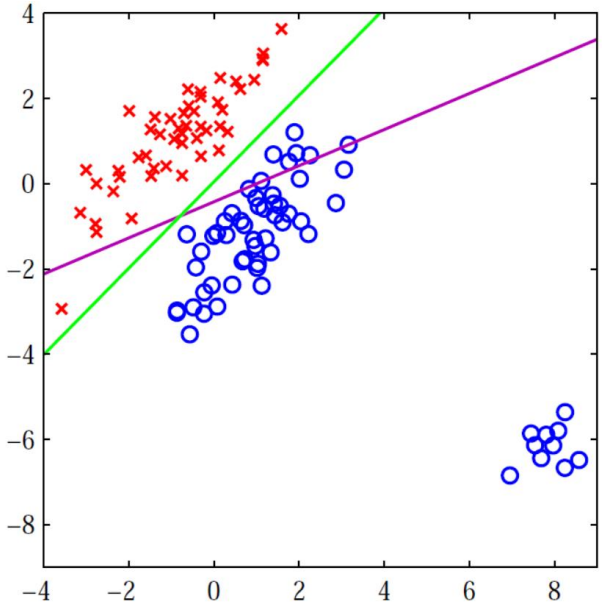
分类的最小二乘

- 如何学习决策边界 $(\boldsymbol{w}_k, w_{k0})$？
- 一种方法是使用最小二乘，与回归类似。
- 寻找$\boldsymbol{W}$使得在所有样本和标签的所有分量上的平方误差最小化：

$$E(\boldsymbol{W}) = \frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{K}(y_k(\boldsymbol{x}_n) - t_{nk})^2$$

最小二乘的问题

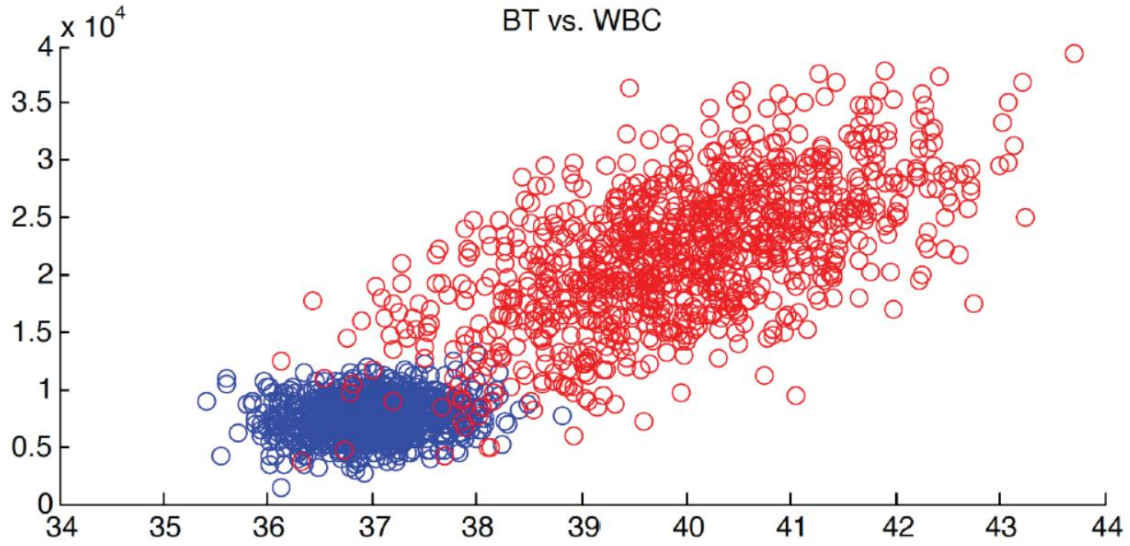




- 品红色线表示使用最小二乘得到的决策边界。
- 绿色线表示使用逻辑回归得到的决策边界。（后面介绍）
- 这两个决策边界类似，都不错。

- 添加数据点（异常值）
- 品红色的最小二乘决策边界变糟了。
- 绿色的逻辑回归决策边界依然可以分离出这些异常值。

Fisher线性判别



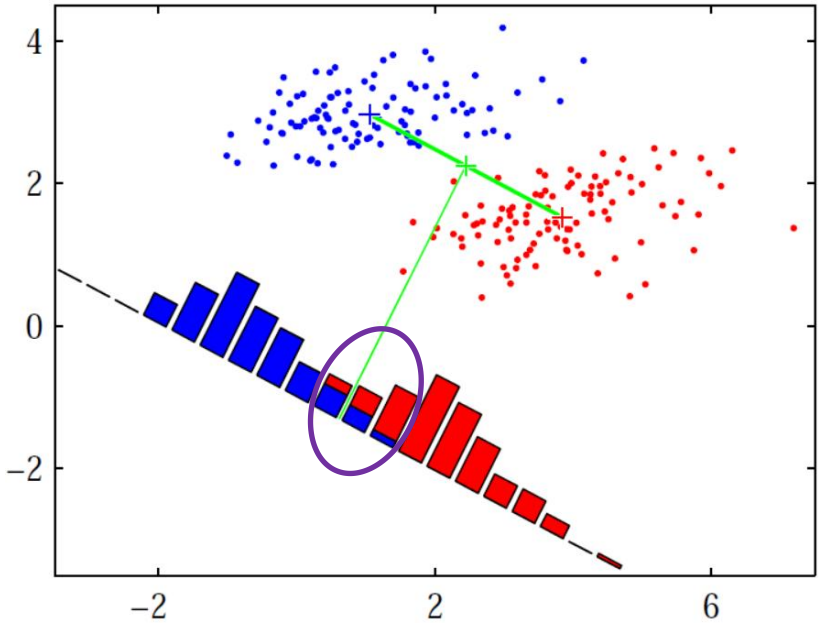
- 两类线性判别实际上是投影：

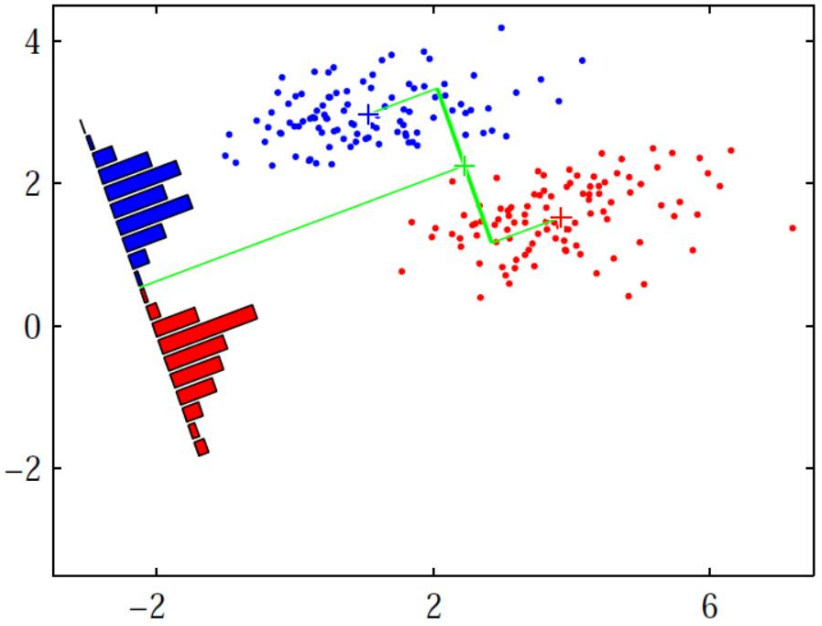
$$y = \boldsymbol{w}^T \boldsymbol{x} \geq -w_0$$

外加一个阈值。

- 我们应该将 $\boldsymbol{w}$ 往哪个方向投影？
- 一个可以将类别分开的方向。

Fisher线性判别





- 一个自然的想法是向连接各类均值的线的方向投影。
- 但是，如果各类在此方向上有分歧，则会出现问题。
- Fisher准则：最大化类间分割与类内方差的比率。

Fisher线性判别

- 投影公式：$y_n = \boldsymbol{w}^T \boldsymbol{x}_n$
- 将属于类别$\mathcal{C}_k$的数据投影到$\boldsymbol{w}$上后的均值：

$$m_k = \frac{1}{N_k} \sum_{n \in \mathcal{C}_k} \boldsymbol{w}^T \boldsymbol{x}_n$$

类间分割是投影数据类均值之间的距离（越大越好）。

- 投影数据类内方差（越小越好）：

$$s_k^2 = \sum_{n \in \mathcal{C}_k} (y_n - m_k)^2$$

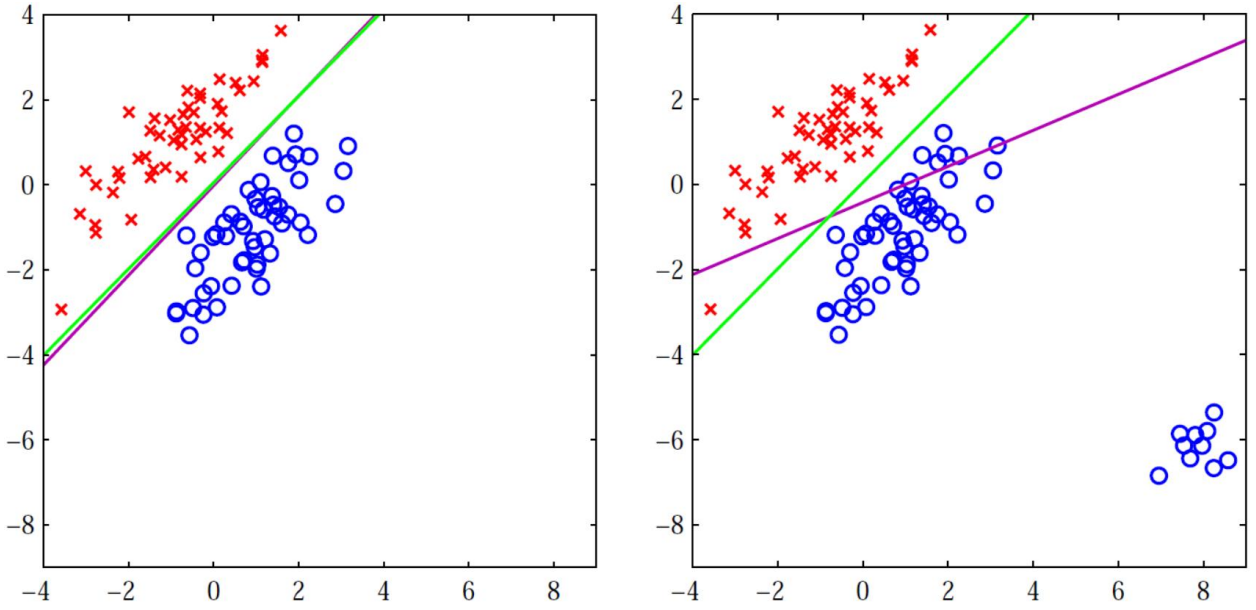
- Fisher准则：

$$J(\boldsymbol{w}) = \frac{(m_2 - m_1)^2}{s_1^2 + s_2^2}$$

关于$\boldsymbol{w}$求最大化。

Fisher线性判别小结

- Fisher线性判别是一种维数降低技术，因为投影是将多维的数据点转换为一维空间中的目标值。
- Fisher线性判别是一种基于类标签选择投影的准则。
  - 仍然受到异常值的影响（例如，最小二乘示例）。



感知机（Perceptron）

- 感知机是两类分类的线性分类模型，其输入是数据点的特征向量，输出是数据点的类别，取+1和−1二值。
- 感知机对应于输入空间（特征空间）中将数据点划分为正负两类的分离超平面，属于判别模型。
- 感知机学习旨在求出将训练数据进行线性划分的分类超平面。
  - 通过导入基于误分类的损失函数，利用梯度下降法对损失函数进行极小化，求得感知机模型。

感知机（Perceptron）

- 感知机的定义：输入$\boldsymbol{\phi(x)}$表示数据点$\boldsymbol{x}$的特征向量，对应于输入空间的数据点$\boldsymbol{x}$；输出$y(\boldsymbol{x}) \in \{1, -1\}$表示数据点的类别。由输入到输出由以下函数表示：

$$y(\boldsymbol{x}) = f(\boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x}) + w_0)$$

称为感知机。其中$\boldsymbol{w}$和w_0是感知机参数，$\boldsymbol{w}$称作权值向量，w_0称作偏差。$f(\cdot)$是符号函数：

$$f(a) = \begin{cases} +1, & a \geq 0 \\ -1, & a < 0 \end{cases}$$

感知机学习策略

- 数据集的线性可分性：给定一个数据集

$$\mathbf{T} = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \cdots, (\boldsymbol{x}_N, y_N)\}$$

其中，$y_i \in \{+1, -1\}$，$i = 1, \cdots, N$。若存在某个超平面S，

$$\boldsymbol{w}^T \phi(\boldsymbol{x}) + w_0 = 0$$

能够将数据集的正样本和负样本完全正确地划分到超平面的两侧，即对所有$y_i = +1$的数据点$\boldsymbol{x}_i$有$\boldsymbol{w}^T \phi(\boldsymbol{x}_i) + w_0 > 0$，对所有$y_i = -1$的数据点$\boldsymbol{x}_i$有$\boldsymbol{w}^T \phi(\boldsymbol{x}_i) + w_0 < 0$，则称数据集$\mathbf{T}$是线性可分数据集；否则，称数据集$\mathbf{T}$线性不可分。

感知机学习策略

- 假设训练数据集是线性可分的，感知机的学习目标是求得一个能够将训练集正样本和负样本正确分开的分离超平面。为了找到这个超平面，即确定感知机模型参数$\boldsymbol{w}$和w_0，需要确定一个学习策略，即定义损失函数并将损失函数极小化。
- 感知机采用的损失函数是误分类点到超平面S的总距离。输入空间中任意数据点到超平面的距离是：

$$\frac{|\boldsymbol{w}^T\phi(\boldsymbol{x}) + w_0|}{\|\boldsymbol{w}\|}$$

其中，$\|\boldsymbol{w}\|$是L_2范数

感知机学习策略

- 对于误分类的数据点$\boldsymbol{x}_i$来说，

$$-y_i \cdot (\boldsymbol{w}^T \phi(\boldsymbol{x}_i) + w_0) > 0$$

因为在误分类情况下，当$\boldsymbol{w}^T \phi(\boldsymbol{x}_i) + w_0 > 0$时，$y_i = -1$；

当$\boldsymbol{w}^T \phi(\boldsymbol{x}_i) + w_0 < 0$时，$y_i = +1$。

- 因此，误分类点$\boldsymbol{x}_i$到超平面S的距离是

$$-\frac{1}{\|\boldsymbol{w}\|} y_i \cdot (\boldsymbol{w}^T \phi(\boldsymbol{x}_i) + w_0)$$

感知机学习策略

- 假设超平面S的误分类点集合为$\mathcal{M}$，那么所有误分类点到超平面S的总距离为：

$$-\frac{1}{\|\boldsymbol{w}\|}\sum_{\boldsymbol{x}_i\in\mathcal{M}} y_i\cdot(\boldsymbol{w}^T\phi(\boldsymbol{x}_i)+w_0)$$

- 感知机的损失函数是：

$$L(\boldsymbol{w},w_0)=-\sum_{\boldsymbol{x}_i\in\mathcal{M}} y_i\cdot(\boldsymbol{w}^T\phi(\boldsymbol{x}_i)+w_0)$$

仅在分错的样本上加和。

感知机学习策略

- 求$\boldsymbol{w}$和w_0，使其为损失函数极小化问题的解：

$$\min_{\boldsymbol{w},w_0} L(\boldsymbol{w},w_0) = -\sum_{\boldsymbol{x}_i \in \mathcal{M}} y_i \cdot (\boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x}_i) + w_0)$$

- 使用随机梯度下降法使误差函数最小化：

① 任意选取一个超平面，使用梯度下降法不断极小化损失函数。

② 随机选取一个误分类点$(\boldsymbol{x}_i, y_i)$，对$\boldsymbol{w}$和w_0进行更新：

$$\boldsymbol{w} \leftarrow \boldsymbol{w} + \eta y_i \boldsymbol{\phi}(\boldsymbol{x}_i)$$

$$w_0 \leftarrow w_0 + \eta y_i$$

其中，η（$0 < \eta \leq 1$）是步长，又称作学习率。

– 通过迭代，损失函数$L(\boldsymbol{w}, w_0)$不断减小，直到为0。

主要内容

1. 分类举例：图像分类
2. 广义线性模型
3. 判别函数（Discriminant Functions）
4. 生成模型（Generative Models）
5. 判别模型（Discriminative Models）

概率生成模型

- 生成模型是由数据学习联合概率分布$p(\mathcal{C}_k, \boldsymbol{x})$，然后求出数据点$\boldsymbol{x}$属于$\mathcal{C}_k$的后验分布$p(\mathcal{C}_k|\boldsymbol{x})$。假设有$\mathcal{C}_1$和$\mathcal{C}_2$两类，$\mathcal{C}_1$类的后验分布为：

$$p(\mathcal{C}_1|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\mathcal{C}_1)p(\mathcal{C}_1)}{p(\boldsymbol{x})} \quad \text{Bayes' Rule}$$

$$p(\mathcal{C}_1|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\mathcal{C}_1)p(\mathcal{C}_1)}{p(\boldsymbol{x}, \mathcal{C}_1) + p(\boldsymbol{x}, \mathcal{C}_2)} \quad \text{Sum rule}$$

$$p(\mathcal{C}_1|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\mathcal{C}_1)p(\mathcal{C}_1)}{p(\boldsymbol{x}|\mathcal{C}_1)p(\mathcal{C}_1) + p(\boldsymbol{x}|\mathcal{C}_2)p(\mathcal{C}_2)} \quad \text{Product rule}$$

$\mathcal{C}_1$类条件分布

$\mathcal{C}_1$类先验

$\boldsymbol{x}$的分布

$\boldsymbol{x}$和$\mathcal{C}_k$的联合分布

- 之所以称作生成模型，是因为生成模型$p(\mathcal{C}_k|\boldsymbol{x})$可以在给定输入$\boldsymbol{x}$时，产生输出$\mathcal{C}_k$的概率分布，可用来生成数据。

概率生成模型 – 例子

- 假设我们观察到x是当前温度。
- 需要确定我们是在西安（$\mathcal{C}_1$）还是三亚（$\mathcal{C}_2$）。
- 生成模型：

$$p(\mathcal{C}_1|x) = \frac{p(x|\mathcal{C}_1)p(\mathcal{C}_1)}{p(x|\mathcal{C}_1)p(\mathcal{C}_1) + p(x|\mathcal{C}_2)p(\mathcal{C}_2)}$$

  - $p(x|\mathcal{C}_1)$ 是西安的典型温度分布，例如 $p(x|\mathcal{C}_1) = \mathcal{N}(x; 10, 5)$
  - $p(x|\mathcal{C}_2)$ 是三亚的典型温度分布，例如 $p(x|\mathcal{C}_2) = \mathcal{N}(x; 25, 5)$
  - 类先验是 $p(\mathcal{C}_1) = 0.1, p(\mathcal{C}_2) = 0.9$
- $p(\mathcal{C}_1|x = 15) = \frac{0.0484 \cdot 0.1}{0.0484 \cdot 0.1 + 0.0108 \cdot 0.9} \approx 0.33$

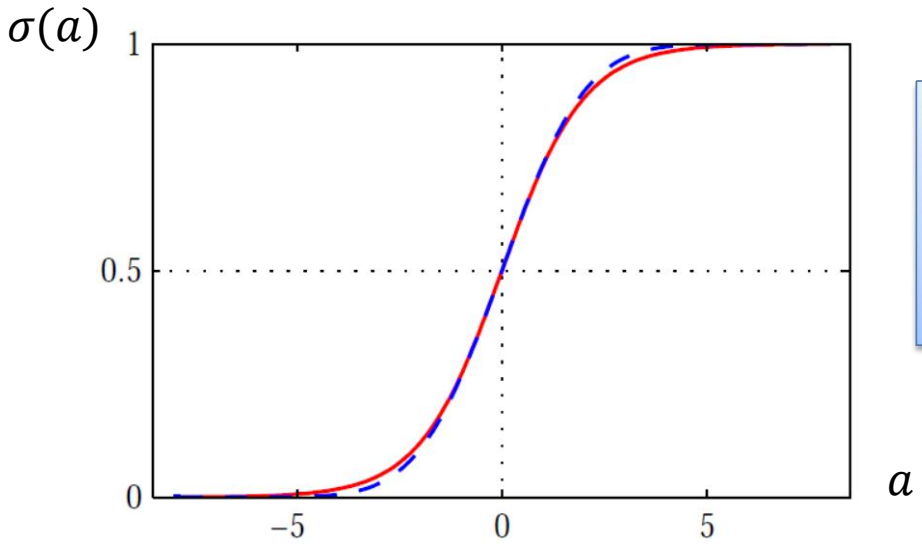
广义线性模型

- 可以将生成模型写成如下形式：

$$
\begin{aligned}
p(\mathcal{C}_1|\boldsymbol{x}) &= \frac{p(\boldsymbol{x}|\mathcal{C}_1)p(\mathcal{C}_1)}{p(\boldsymbol{x}|\mathcal{C}_1)p(\mathcal{C}_1) + p(\boldsymbol{x}|\mathcal{C}_2)p(\mathcal{C}_2)} \\
&= \frac{1}{1+\exp(-a)} \equiv \sigma(a) \\
\text{where} \quad a &= \ln \frac{p(\boldsymbol{x}|\mathcal{C}_1)p(\mathcal{C}_1)}{p(\boldsymbol{x}|\mathcal{C}_2)p(\mathcal{C}_2)}
\end{aligned}
$$

Logistic Sigmoid

- 函数$\sigma(a)=\frac{1}{1+\exp(-a)}$，被称作Logistic sigmoid。
- 它将实轴压缩到[0,1]，和概率的取值范围一致。
- 它是连续且可微的。

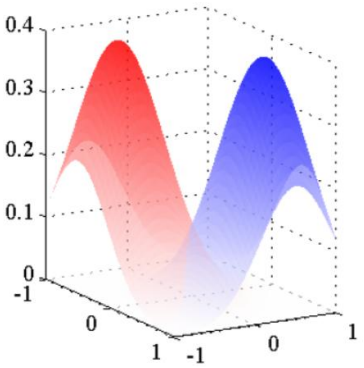


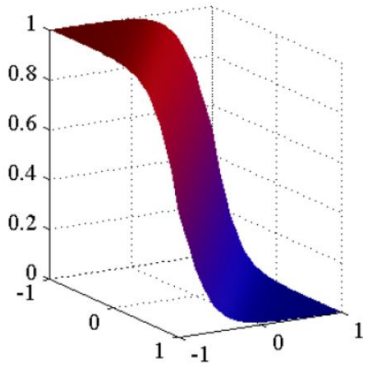
$$\sigma(a)=p(\mathcal{C}_1|\boldsymbol{x})$$

$$a=\ln\frac{p(\boldsymbol{x}|\mathcal{C}_1)p(\mathcal{C}_1)}{p(\boldsymbol{x}|\mathcal{C}_2)p(\mathcal{C}_2)}=\ln\frac{p(\mathcal{C}_1|\boldsymbol{x})}{p(\mathcal{C}_2|\boldsymbol{x})}$$

高斯类条件密度

$p(\boldsymbol{x}|\mathcal{C}_k)$





$$p(\mathcal{C}_1|\boldsymbol{x}) = \sigma(a) = \frac{1}{1+\exp(-a)}$$

$$p(\mathcal{C}_2|\boldsymbol{x}) = 1 - p(\mathcal{C}_1|\boldsymbol{x})$$

- 假设类条件密度（似然函数）$p(\boldsymbol{x}|\mathcal{C}_k)$是高斯分布，并且具有相同的协方差矩阵$\sum$，那么写做：

$$p(\boldsymbol{x}|\mathcal{C}_k) = \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k)\right\}$$

$$a = \ln\frac{p(\boldsymbol{x}|\mathcal{C}_1)p(\mathcal{C}_1)}{p(\boldsymbol{x}|\mathcal{C}_2)p(\mathcal{C}_2)} = \boldsymbol{w}^T\boldsymbol{x} + w_0$$

$$\mathbf{w} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$$

$$w_0 = -\frac{1}{2}\boldsymbol{\mu}_1^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 + \frac{1}{2}\boldsymbol{\mu}_2^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2 + \ln\frac{p(\mathcal{C}_1)}{p(\mathcal{C}_2)}$$

高斯类条件密度

$$a = \ln \frac{p(\boldsymbol{x}|\mathcal{C}_1)p(\mathcal{C}_1)}{p(\boldsymbol{x}|\mathcal{C}_2)p(\mathcal{C}_2)} = \ln \frac{p(\boldsymbol{x}|\mathcal{C}_1)}{p(\boldsymbol{x}|\mathcal{C}_2)} + \ln \frac{p(\mathcal{C}_1)}{p(\mathcal{C}_2)}$$

$$\ln \frac{p(\boldsymbol{x}|\mathcal{C}_1)}{p(\boldsymbol{x}|\mathcal{C}_2)} = -\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_1)^{\mathrm{T}}\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1) + \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_2)^{\mathrm{T}}\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2)$$

$$= -\frac{1}{2}(\boldsymbol{x}^{\mathrm{T}}-\boldsymbol{\mu}_1^{\mathrm{T}})\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1) + \frac{1}{2}(\boldsymbol{x}^{\mathrm{T}}-\boldsymbol{\mu}_2^{\mathrm{T}})\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2)$$

$$= -\frac{1}{2}(\boldsymbol{x}^{\mathrm{T}}\Sigma^{-1}\boldsymbol{x} - \boldsymbol{x}^{\mathrm{T}}\Sigma^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_1^{\mathrm{T}}\Sigma^{-1}\boldsymbol{x} + \boldsymbol{\mu}_1^{\mathrm{T}}\Sigma^{-1}\boldsymbol{\mu}_1) + \frac{1}{2}(\boldsymbol{x}^{\mathrm{T}}\Sigma^{-1}\boldsymbol{x} - \boldsymbol{x}^{\mathrm{T}}\Sigma^{-1}\boldsymbol{\mu}_2 - \boldsymbol{\mu}_2^{\mathrm{T}}\Sigma^{-1}\boldsymbol{x} + \boldsymbol{\mu}_2^{\mathrm{T}}\Sigma^{-1}\boldsymbol{\mu}_2)$$

$$= -\frac{1}{2}(\boldsymbol{x}^{\mathrm{T}}\Sigma^{-1}\boldsymbol{x} - 2\boldsymbol{\mu}_1^{\mathrm{T}}\Sigma^{-1}\boldsymbol{x} + \boldsymbol{\mu}_1^{\mathrm{T}}\Sigma^{-1}\boldsymbol{\mu}_1) + \frac{1}{2}(\boldsymbol{x}^{\mathrm{T}}\Sigma^{-1}\boldsymbol{x} - 2\boldsymbol{\mu}_2^{\mathrm{T}}\Sigma^{-1}\boldsymbol{x} + \boldsymbol{\mu}_2^{\mathrm{T}}\Sigma^{-1}\boldsymbol{\mu}_2)$$

$$= \boldsymbol{\mu}_1^{\mathrm{T}}\Sigma^{-1}\boldsymbol{x} - \frac{1}{2}\boldsymbol{\mu}_1^{\mathrm{T}}\Sigma^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2^{\mathrm{T}}\Sigma^{-1}\boldsymbol{x} + \frac{1}{2}\boldsymbol{\mu}_2^{\mathrm{T}}\Sigma^{-1}\boldsymbol{\mu}_2$$

$$= (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^{\mathrm{T}}\Sigma^{-1}\boldsymbol{x} - \frac{1}{2}\boldsymbol{\mu}_1^{\mathrm{T}}\Sigma^{-1}\boldsymbol{\mu}_1 + \frac{1}{2}\boldsymbol{\mu}_2^{\mathrm{T}}\Sigma^{-1}\boldsymbol{\mu}_2$$

高斯类条件密度

$$a = \ln \frac{p(\boldsymbol{x}|\mathcal{C}_1)p(\mathcal{C}_1)}{p(\boldsymbol{x}|\mathcal{C}_2)p(\mathcal{C}_2)} = \ln \frac{p(\boldsymbol{x}|\mathcal{C}_1)}{p(\boldsymbol{x}|\mathcal{C}_2)} + \ln \frac{p(\mathcal{C}_1)}{p(\mathcal{C}_2)}$$

$$= (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^{\mathrm{T}} \Sigma^{-1} \boldsymbol{x} - \frac{1}{2} \boldsymbol{\mu}_1^{\mathrm{T}} \Sigma^{-1} \boldsymbol{\mu}_1 + \frac{1}{2} \boldsymbol{\mu}_2^{\mathrm{T}} \Sigma^{-1} \boldsymbol{\mu}_2 + \ln \frac{p(\mathcal{C}_1)}{p(\mathcal{C}_2)}$$

$$= \boldsymbol{w}^T \boldsymbol{x} + w_0$$

$$\boldsymbol{w} = \Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$$

$$w_0 = -\frac{1}{2} \boldsymbol{\mu}_1^{\mathrm{T}} \Sigma^{-1} \boldsymbol{\mu}_1 + \frac{1}{2} \boldsymbol{\mu}_2^{\mathrm{T}} \Sigma^{-1} \boldsymbol{\mu}_2 + \ln \frac{p(\mathcal{C}_1)}{p(\mathcal{C}_2)}$$

最大似然学习

- 指定类条件密度$p(\boldsymbol{x}|\mathcal{C}_k)$的参数函数形式（高斯分布）后，可以使用最大似然学习确定参数值$\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$, $\sum$，以及类先验$p(\mathcal{C}_k)$，其中$p(\mathcal{C}_1) = \pi$，$p(\mathcal{C}_2) = 1 - \pi$。

- 对于一个来自类$\mathcal{C}_1$的数据点$\boldsymbol{x}_n$，其目标值$t_n = 1$，有：

$$p(\boldsymbol{x}_n, \mathcal{C}_1) = p(\mathcal{C}_1)p(\boldsymbol{x}_n|\mathcal{C}_1) = \pi\mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$$

- 对于一个来自类$\mathcal{C}_2$的数据点$\boldsymbol{x}_n$，其目标值$t_n = 0$，有：

$$p(\boldsymbol{x}_n, \mathcal{C}_2) = p(\mathcal{C}_2)p(\boldsymbol{x}_n|\mathcal{C}_2) = (1 - \pi)\mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$$

最大似然学习

- 训练数据上的似然函数是：

$$p(\boldsymbol{t}|\pi,\boldsymbol{\mu}_1,\boldsymbol{\mu}_2,\boldsymbol{\Sigma}) = \prod_{n=1}^{N}[\pi\mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_1,\boldsymbol{\Sigma})]^{t_n}[(1-\pi)\mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_2,\boldsymbol{\Sigma})]^{1-t_n}$$

where $\mathbf{t} = (t_1,\ldots,t_N)^{\mathrm{T}}$

- 用θ代表所有参数，对似然函数取对数，有：

$$\ell(\boldsymbol{t};\theta) = \sum_{n=1}^{N} \underbrace{t_n \ln\pi + (1-t_n)\ln(1-\pi)}_{\pi} + \underbrace{t_n \ln\mathcal{N}_1 + (1-t_n)\ln\mathcal{N}_2}_{\boldsymbol{\mu}_1,\boldsymbol{\mu}_2,\boldsymbol{\Sigma}}$$

然后对π和$\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$,$\sum$分别求最大化。

最大似然学习 – 类先验

- 关于类先验参数π求最大化很简单，即关于π求导：

$$\frac{\partial}{\partial \pi}\ell(\boldsymbol{t};\theta) = \sum_{n=1}^{N} \frac{t_n}{\pi} - \frac{1-t_n}{1-\pi}$$

$$\Rightarrow \boxed{\pi = \frac{N_1}{N_1 + N_2}}$$

- N_1和N_2是分别是类$\mathcal{C}_1$和类$\mathcal{C}_2$中训练数据点的个数。
- 先验π是类$\mathcal{C}_1$中数据点个数与训练数据点总数之比。

$$\ell(\boldsymbol{t};\theta) = \sum_{n=1}^{N} \underbrace{t_n \ln \pi + (1-t_n)\ln(1-\pi)}_{\pi} + \underbrace{t_n \ln \mathcal{N}_1 + (1-t_n)\ln \mathcal{N}_2}_{\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}}$$

$$\frac{\partial}{\partial \pi}\ell(\boldsymbol{t},\theta) = \sum_{n=1}^{N}\left(\frac{t_n}{\pi} - \frac{1-t_n}{1-\pi}\right)$$

$$= \sum_{n=1}^{N}\left(\frac{t_n(1-\pi) - (1-t_n)\pi}{\pi(1-\pi)}\right)$$

$$= \sum_{n=1}^{N}\left(\frac{t_n - \pi}{\pi(1-\pi)}\right)$$

$$\sum_{n=1}^{N}\left(\frac{t_n - \pi}{\pi(1-\pi)}\right) = 0$$

$$\sum_{n=1}^{N} t_n = \sum_{n=1}^{N} \pi$$

$$\Rightarrow \quad \pi = \frac{N_1}{N_1 + N_2}$$

$$\ell(\boldsymbol{t};\theta) = \sum_{n=1}^{N} \underbrace{t_n \ln \pi + (1-t_n)\ln(1-\pi)}_{\pi} + \underbrace{t_n \ln \mathcal{N}_1 + (1-t_n)\ln \mathcal{N}_2}_{\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}}$$

最大似然学习 – 高斯参数

- 其它参数$\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$, $\sum$也可以用相同方式获得。
- 类均值：

$$\boldsymbol{\mu}_1 = \frac{1}{N_1}\sum_{n=1}^{N} t_n \boldsymbol{x}_n$$

$$\boldsymbol{\mu}_2 = \frac{1}{N_2}\sum_{n=1}^{N} (1 - t_n)\boldsymbol{x}_n$$

  - 每类中训练样本的均值。

$$\ell(\boldsymbol{t}; \theta) = \sum_{n=1}^{N} \underbrace{t_n \ln \pi + (1 - t_n) \ln(1 - \pi)}_{\pi} + \underbrace{t_n \ln \mathcal{N}_1 + (1 - t_n) \ln \mathcal{N}_2}_{\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}}$$

$$\frac{\partial}{\partial \boldsymbol{\mu}_1}\ell(\boldsymbol{t},\theta) = \sum_{n=1}^{N} t_n \frac{\partial}{\partial \boldsymbol{\mu}_1}\ln \mathcal{N}_1$$

$$= \sum_{n=1}^{N} t_n \frac{\partial}{\partial \boldsymbol{\mu}_1}\ln\left(\frac{1}{2\pi|\sum|^{1/2}}\exp\left\{-\frac{1}{2}(\boldsymbol{x_n}-\boldsymbol{\mu}_1)^T\sum^{-1}(\boldsymbol{x_n}-\boldsymbol{\mu}_1)\right\}\right)$$

$$= \sum_{n=1}^{N} t_n \frac{\partial}{\partial \boldsymbol{\mu}_1}\left(-\ln 2\pi|\sum|^{1/2} - \frac{1}{2}(\boldsymbol{x_n}-\boldsymbol{\mu}_1)^T\sum^{-1}(\boldsymbol{x_n}-\boldsymbol{\mu}_1)\right)$$

$$= -\sum_{n=1}^{N} t_n \sum^{-1}(\boldsymbol{x_n}-\boldsymbol{\mu}_1)$$

$$p(\boldsymbol{x}|\mathcal{C}_k) = \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}}\exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k)\right\}$$

$$\ell(\boldsymbol{t};\theta) = \sum_{n=1}^{N} \underbrace{t_n \ln \pi + (1-t_n)\ln(1-\pi)}_{\pi} + \underbrace{t_n \ln \mathcal{N}_1 + (1-t_n)\ln \mathcal{N}_2}_{\boldsymbol{\mu}_1,\boldsymbol{\mu}_2,\boldsymbol{\Sigma}}$$

$$\sum_{n=1}^{N} t_n \Sigma^{-1}(\boldsymbol{x_n} - \boldsymbol{\mu}_1) = \mathbf{0}$$

$$\sum_{n=1}^{N} t_n \boldsymbol{x_n} = \sum_{n=1}^{N} t_n \boldsymbol{\mu}_1$$

$$\boldsymbol{\mu}_1 = \frac{\sum_{n=1}^{N} t_n \boldsymbol{x_n}}{\sum_{n=1}^{N} t_n}$$

$$\boldsymbol{\mu}_1 = \frac{1}{N_1} \sum_{n=1}^{N} t_n \boldsymbol{x_n}$$

$$p(\boldsymbol{x}|\mathcal{C}_k) = \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_k)\right\}$$

$$\ell(\boldsymbol{t}; \theta) = \sum_{n=1}^{N} \underbrace{t_n \ln \pi + (1 - t_n) \ln(1 - \pi)}_{\pi} + \underbrace{t_n \ln \mathcal{N}_1 + (1 - t_n) \ln \mathcal{N}_2}_{\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}}$$

$$\frac{\partial}{\partial \boldsymbol{\mu}_2}\ell(\boldsymbol{t},\theta)=\sum_{n=1}^{N}(1-t_n)\frac{\partial}{\partial \boldsymbol{\mu}_2}\ln\mathcal{N}_2$$

$$=\sum_{n=1}^{N}(1-t_n)\frac{\partial}{\partial \boldsymbol{\mu}_2}\ln\left(\frac{1}{2\pi|\sum|^{1/2}}\exp\left\{-\frac{1}{2}(\boldsymbol{x}_n-\boldsymbol{\mu}_2)^T\Sigma^{-1}(\boldsymbol{x}_n-\boldsymbol{\mu}_2)\right\}\right)$$

$$=\sum_{n=1}^{N}(1-t_n)\frac{\partial}{\partial \boldsymbol{\mu}_2}\left(-\ln 2\pi|\sum|^{1/2}-\frac{1}{2}(\boldsymbol{x}_n-\boldsymbol{\mu}_2)^T\Sigma^{-1}(\boldsymbol{x}_n-\boldsymbol{\mu}_2)\right)$$

$$=-\sum_{n=1}^{N}(1-t_n)\Sigma^{-1}(\boldsymbol{x}_n-\boldsymbol{\mu}_2)$$

$$p(\boldsymbol{x}|\mathcal{C}_k)=\frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}}\exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k)\right\}$$

$$\ell(\boldsymbol{t};\theta)=\sum_{n=1}^{N}\underbrace{t_n\ln\pi+(1-t_n)\ln(1-\pi)}_{\pi}+\underbrace{t_n\ln\mathcal{N}_1+(1-t_n)\ln\mathcal{N}_2}_{\boldsymbol{\mu}_1,\boldsymbol{\mu}_2,\boldsymbol{\Sigma}}$$

$$\sum_{n=1}^{N}(1-t_n)\Sigma^{-1}(\boldsymbol{x}_n-\boldsymbol{\mu}_2)=\mathbf{0}$$

$$\sum_{n=1}^{N}(1-t_n)\,\boldsymbol{x}_n=\sum_{n=1}^{N}(1-t_n)\boldsymbol{\mu}_2$$

$$\boldsymbol{\mu}_2=\frac{\sum_{n=1}^{N}(1-t_n)\,\boldsymbol{x}_n}{\sum_{n=1}^{N}(1-t_n)}$$

$$\boldsymbol{\mu}_2=\frac{1}{N-N_1}\sum_{n=1}^{N}(1-t_n)\,\boldsymbol{x}_n$$

$$\boldsymbol{\mu}_2=\frac{1}{N_2}\sum_{n=1}^{N}(1-t_n)\,\boldsymbol{x}_n$$

$$p(\boldsymbol{x}|\mathcal{C}_k)=\frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}}\exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k)\right\}$$

$$\ell(\boldsymbol{t};\theta)=\sum_{n=1}^{N}\underbrace{t_n\ln\pi+(1-t_n)\ln(1-\pi)}_{\pi}+\underbrace{t_n\ln\mathcal{N}_1+(1-t_n)\ln\mathcal{N}_2}_{\boldsymbol{\mu}_1,\boldsymbol{\mu}_2,\boldsymbol{\Sigma}}$$

最大似然学习 – 高斯参数

- 共享的协方差矩阵：

$$\boldsymbol{\Sigma} = \frac{N_1}{N}\frac{1}{N_1}\sum_{n\in\mathcal{C}_1}(\boldsymbol{x}_n-\boldsymbol{\mu}_1)(\boldsymbol{x}_n-\boldsymbol{\mu}_1)^T+\frac{N_2}{N}\frac{1}{N_2}\sum_{n\in\mathcal{C}_2}(\boldsymbol{x}_n-\boldsymbol{\mu}_2)(\boldsymbol{x}_n-\boldsymbol{\mu}_2)^T$$

  - 类协方差的加权平均值。

$$\ell(\boldsymbol{t};\theta) = \sum_{n=1}^{N} \underbrace{t_n \ln \pi + (1-t_n)\ln(1-\pi)}_{\pi} + \underbrace{t_n \ln \mathcal{N}_1 + (1-t_n)\ln \mathcal{N}_2}_{\boldsymbol{\mu}_1,\boldsymbol{\mu}_2,\boldsymbol{\Sigma}}$$

$$\frac{\partial}{\partial \sum}\ell(\boldsymbol{t},\theta)=\sum_{n=1}^{N}t_n\frac{\partial}{\partial \sum}\ln\mathcal{N}_1+\sum_{n=1}^{N}(1-t_n)\frac{\partial}{\partial \sum}\ln\mathcal{N}_2$$

$$=\sum_{n=1}^{N}t_n\frac{\partial}{\partial \sum}\ln\left(\frac{1}{2\pi|\sum|^{1/2}}\exp\left\{-\frac{1}{2}(\boldsymbol{x}_n-\boldsymbol{\mu}_1)^T\sum^{-1}(\boldsymbol{x}_n-\boldsymbol{\mu}_1)\right\}\right)$$

$$+\sum_{n=1}^{N}(1-t_n)\frac{\partial}{\partial \sum}\ln\left(\frac{1}{2\pi|\sum|^{1/2}}\exp\left\{-\frac{1}{2}(\boldsymbol{x}_n-\boldsymbol{\mu}_2)^T\sum^{-1}(\boldsymbol{x}_n-\boldsymbol{\mu}_2)\right\}\right)$$

$$=\sum_{n=1}^{N}t_n\frac{\partial}{\partial \sum}\left(-\ln 2\pi-\frac{1}{2}\ln|\sum|-\frac{1}{2}(\boldsymbol{x}_n-\boldsymbol{\mu}_1)^T\sum^{-1}(\boldsymbol{x}_n-\boldsymbol{\mu}_1)\right)$$

$$+\sum_{n=1}^{N}(1-t_n)\frac{\partial}{\partial \sum}\left(-\ln 2\pi-\frac{1}{2}\ln|\sum|-\frac{1}{2}(\boldsymbol{x}_n-\boldsymbol{\mu}_2)^T\sum^{-1}(\boldsymbol{x}_n-\boldsymbol{\mu}_2)\right)$$

$$=\sum_{n=1}^{N}t_n\left(-\frac{1}{2}|\sum|^{-1}+\frac{1}{2}\sum^{-2}(\boldsymbol{x}_n-\boldsymbol{\mu}_1)(\boldsymbol{x}_n-\boldsymbol{\mu}_1)^T\right)$$

$$+\sum_{n=1}^{N}(1-t_n)\left(-\frac{1}{2}|\sum|^{-1}+\frac{1}{2}\sum^{-2}(\boldsymbol{x}_n-\boldsymbol{\mu}_2)(\boldsymbol{x}_n-\boldsymbol{\mu}_2)^T\right)$$

$$\sum_{n=1}^{N} t_n(-|\sum|^{-1} + \sum^{-2}(\boldsymbol{x}_n - \boldsymbol{\mu}_1)(\boldsymbol{x}_n - \boldsymbol{\mu}_1)^T)$$

$$+ \sum_{n=1}^{N} (1 - t_n)\,(-|\sum|^{-1} + \sum^{-2}(\boldsymbol{x}_n - \boldsymbol{\mu}_2)(\boldsymbol{x}_n - \boldsymbol{\mu}_2)^T) = \mathbf{0}$$

$$\sum_{n=1}^{N} t_n|\sum|^{-1} + \sum_{n=1}^{N} (1 - t_n)|\sum|^{-1}$$

$$= \sum_{n=1}^{N} t_n\sum^{-2}(\boldsymbol{x}_n - \boldsymbol{\mu}_1)(\boldsymbol{x}_n - \boldsymbol{\mu}_1)^T + \sum_{n=1}^{N} (1 - t_n)\sum^{-2}(\boldsymbol{x}_n - \boldsymbol{\mu}_2)(\boldsymbol{x}_n - \boldsymbol{\mu}_2)^T$$

$$\sum_{n=1}^{N} t_n + \sum_{n=1}^{N} (1 - t_n)$$

$$= \sum^{-1} \sum_{n=1}^{N} t_n(\boldsymbol{x}_n - \boldsymbol{\mu}_1)(\boldsymbol{x}_n - \boldsymbol{\mu}_1)^T + \sum^{-1} \sum_{n=1}^{N} (1 - t_n)(\boldsymbol{x}_n - \boldsymbol{\mu}_2)(\boldsymbol{x}_n - \boldsymbol{\mu}_2)^T$$

$$\Sigma = \frac{\sum_{n=1}^{N} t_n(\boldsymbol{x}_n - \boldsymbol{\mu}_1)(\boldsymbol{x}_n - \boldsymbol{\mu}_1)^T + \sum_{n=1}^{N}(1-t_n)(\boldsymbol{x}_n - \boldsymbol{\mu}_2)(\boldsymbol{x}_n - \boldsymbol{\mu}_2)^T}{\sum_{n=1}^{N} t_n + \sum_{n=1}^{N}(1-t_n)}$$

$$= \frac{\sum_{n=1}^{N} t_n(\boldsymbol{x}_n - \boldsymbol{\mu}_1)(\boldsymbol{x}_n - \boldsymbol{\mu}_1)^T + \sum_{n=1}^{N}(1-t_n)(\boldsymbol{x}_n - \boldsymbol{\mu}_2)(\boldsymbol{x}_n - \boldsymbol{\mu}_2)^T}{N_1 + N - N_1}$$

$$= \frac{1}{N}\sum_{n\in\mathcal{C}_1}(\boldsymbol{x}_n - \boldsymbol{\mu}_1)(\boldsymbol{x}_n - \boldsymbol{\mu}_1)^T + \frac{1}{N}(1-t_n)\sum_{n\in\mathcal{C}_2}(\boldsymbol{x}_n - \boldsymbol{\mu}_2)(\boldsymbol{x}_n - \boldsymbol{\mu}_2)^T$$

$$= \frac{N_1}{N}\frac{1}{N_1}\sum_{n\in\mathcal{C}_1}(\boldsymbol{x}_n - \boldsymbol{\mu}_1)(\boldsymbol{x}_n - \boldsymbol{\mu}_1)^T + \frac{N_2}{N}\frac{1}{N_2}\sum_{n\in\mathcal{C}_2}(\boldsymbol{x}_n - \boldsymbol{\mu}_2)(\boldsymbol{x}_n - \boldsymbol{\mu}_2)^T$$

主要内容

1. 分类举例：图像分类
2. 广义线性模型
3. 判别函数（Discriminant Functions）
4. 生成模型（Generative Models）
5. 判别模型（Discriminative Models）

概率判别模型

- 判别模型直接学习的是条件概率$p(\mathcal{C}_k|\boldsymbol{x})$或评分函数$f(\cdot)$，因而可以对数据进行各种程度的抽象、定义特征并使用特征，从而简化学习问题。生成模型是从数据学习联合概率分布$p(\mathcal{C}_k, \boldsymbol{x})$，然后求出条件概率分布$p(\mathcal{C}_k|\boldsymbol{x})$。

- 判别模型——明确使用logistic Sigmoid函数形式：

$$p(\mathcal{C}_1|\boldsymbol{x}) = \sigma\big(y(\boldsymbol{x})\big) = \frac{1}{1+\exp(-y(\boldsymbol{x}))} = \frac{1}{1+\exp(-(\boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x})+w_0))}$$

直接求解$\boldsymbol{w}$。

最大似然学习 – 判别模型

- 似然函数写做：

$$p(\mathbf{t}|\mathbf{w}) = \prod_{n=1}^{N} y_n^{t_n} \{1 - y_n\}^{1-t_n}$$

$$\text{where } y_n = p(\mathcal{C}_1|\boldsymbol{x}_n)$$

- 损失函数定义为对似然函数取负对数：

$$E(\mathbf{w}) = -\ln p(\mathbf{t}|\mathbf{w}) = -\sum_{n=1}^{N} \{t_n \ln y_n + (1 - t_n) \ln(1 - y_n)\}$$

可以使用（随机）梯度下降求解$\boldsymbol{w}$。

总结

- 广义线性模型 $y(\boldsymbol{x}) = f(\boldsymbol{w}^T\boldsymbol{x} + w_0)$
- $f(\cdot)$ 的阈值/最大值函数
  - 使用最小二乘最小化
  - Fisher准则——类分离
  - 感知机准则——错误分类的样本
- 概率模型：$f(\cdot)$ 的Logistic Sigmoid函数
  - 生成模型——先求联合分布，再求后验分布。
  - 判别模型——使用Sigmoid直接对后验进行建模。

- 所有这些模型都限于特征空间中的线性决策边界。