

Chapter 5 Classification: Probability Generative Model

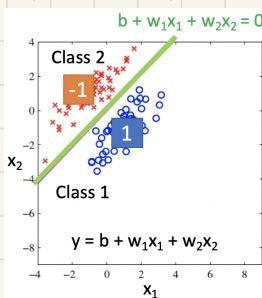
1. 问题描述

输入：- 宝可梦的属性值 (HP / Attack / Defense / SP Atk / SP Def / Speed)

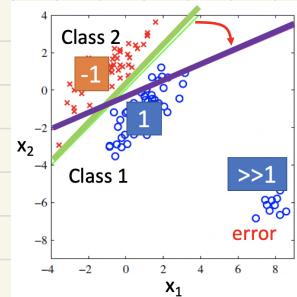
输出：宝可梦的阶层系列 (火 / 水 / 土 / 风)

① Classification 与 Regression 的联系与区别：

从二分类为例，如果用回归思想去解决分类问题，那么对训练集数据集，类别1的标签为1，类别2的标签为-1；对测试数据集，如果预测结果越接近1则判定为类别1，越接近-1，则判定为类别2。

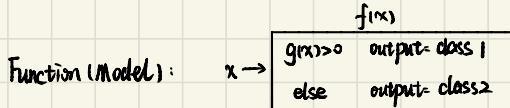


- 对左侧希望红色点的预测结果都接近-1，蓝色点的预测结果都接近1
- 对右侧的分布情况，远离绿色分界线的点的预测值远大于1，变成了error
- 对于分类，绿色分界线更好；对于回归，紫色分界更好



对于Multi-Classification而言，如果将分类问题当作回归处理，则 Class 1 \rightarrow target $y=1$ ，Class 2 \rightarrow target $y=2$ ，Class 3 \rightarrow target $y=3$ 。这时就意味着，某种程度上，类别1与类别2更近，类别2与类别3更近，但这种关系可能并不存在，所以这种方法是存在问题的。

② 理想解决方案



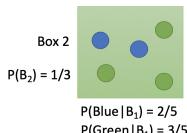
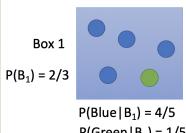
Loss Function: $L(f) = \sum_i \delta(f(x^n) \neq \hat{y}^n)$ 预测错误的数量

Find Best Function: Perceptron, SVM

2. 贝叶斯分类基本原理

① 基本原理：

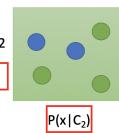
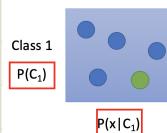
Two Boxes



from one of the boxes

$$P(B_1 \mid \text{Blue}) = \frac{P(\text{Blue}|B_1)P(B_1)}{P(\text{Blue}|B_1)P(B_1) + P(\text{Blue}|B_2)P(B_2)}$$

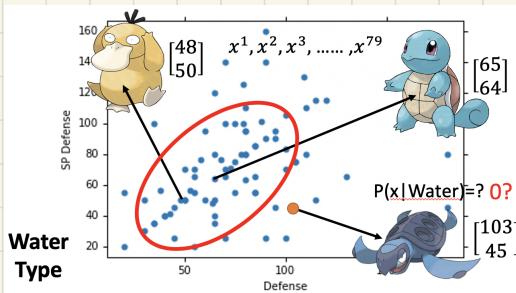
Two Classes



Given an x , which class does it belong to

$$P(C_1|x) = \frac{P(x|C_1)P(C_1)}{P(x|C_1)P(C_1) + P(x|C_2)P(C_2)}$$

- 右图中, $P(C_1)$, $P(x|C_1)$, $P(C_2)$, $P(x|C_2)$ 需要从训练数据中进行估计
- 如模型也称为生成模型 (Generative Model), 因为样本 x 的分布 $P(x) = P(x|C_1)P(C_1) + P(x|C_2)P(C_2)$ 可以计算出, 也就可根据此分布生成样本
- 先验概率 (Prior Probability): 训练样本中, Class 1 有 n 个, Class 2 有 m 个, 则先验概率 $P(C_1) = \frac{n}{n+m}$, $P(C_2) = \frac{m}{n+m}$
- 类条件概率 (Class Conditional Probability):
假设每一次宝可梦由一组特征向量代表, 在水系宝可梦 (79只) 中出现海龟的概率
现只考虑 Defense 和 SP Defense 两个特征, 如下图所示:



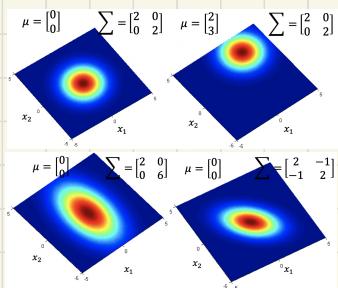
$P(x|C_1) = ?$ $P(x|C_2) = ?$



假设图中的点都是由高斯分布产生的, 其形状由均值 μ 和方差矩阵 Σ 控制

$$f_{\mu, \Sigma}(x) = \frac{1}{(2\pi)^D / |\Sigma|^{1/2}} \exp \left\{ -\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right\}$$

因此, 类条件概率 $P(x|C) = f_{\mu_C, \Sigma_C}(x)$



• 分类概率计算 (贝叶斯概率):

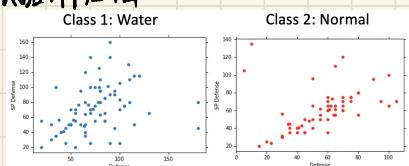
因为问题转化为如何根据训练样本估计出其符合的分布参数, 可用 Maximum Likelihood 解决. 水系宝可梦的 79 个样本点可以被任意一个高斯分布产生, 只是产它的机率不同, 因此只需要找到可能性最大的高斯分布 $N(\mu^*, \Sigma^*)$. 高斯分布 $N(\mu, \Sigma)$ 的 likelihood = 由该高斯条样出 $x^1, x^2, x^3, \dots, x^7$ 的概率 (每条的采样是独立的)

$$L(\mu, \Sigma) = f_{\mu, \Sigma}(x^1) f_{\mu, \Sigma}(x^2) f_{\mu, \Sigma}(x^3) \cdots f_{\mu, \Sigma}(x^n)$$

$$\mu^*, \Sigma^* = \arg \max_{\mu, \Sigma} L(\mu, \Sigma)$$

可以求解 $L(\mu, \Sigma)$ 的微分方程解出 μ^*, Σ^* , 同时也可以根据样本估计出:

$$\mu^* = \frac{1}{n} \sum_{i=1}^n x^i \quad \Sigma^* = \frac{1}{n} \sum_{i=1}^n (x^i - \mu^*) (x^i - \mu^*)^T$$



估计出 μ^* 和 Σ^* 之后, 就可以进行贝叶斯分类:

$$f_{\mu^1, \Sigma^1}(x) = \frac{1}{(2\pi)^{D/2} |\Sigma^1|^{1/2}} \exp \left\{ -\frac{1}{2} (x - \mu^1)^T (\Sigma^1)^{-1} (x - \mu^1) \right\}$$

$$\mu^1 = \begin{bmatrix} 75.0 \\ 71.3 \end{bmatrix} \quad \Sigma^1 = \begin{bmatrix} 874 & 327 \\ 327 & 929 \end{bmatrix}$$

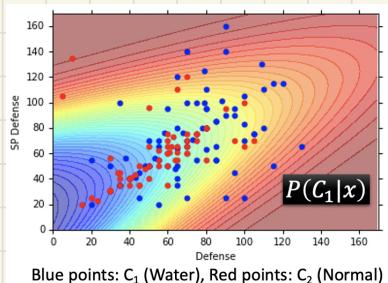
$$P(C_1|x) = \frac{P(x|C_1)P(C_1)}{P(x|C_1)P(C_1) + P(x|C_2)P(C_2)}$$

$$f_{\mu^2, \Sigma^2}(x) = \frac{1}{(2\pi)^{D/2} |\Sigma^2|^{1/2}} \exp \left\{ -\frac{1}{2} (x - \mu^2)^T (\Sigma^2)^{-1} (x - \mu^2) \right\}$$

$$\mu^2 = \begin{bmatrix} 55.6 \\ 59.8 \end{bmatrix} \quad \Sigma^2 = \begin{bmatrix} 847 & 422 \\ 422 & 685 \end{bmatrix}$$

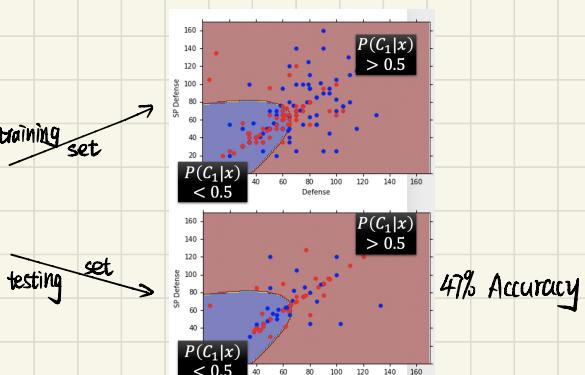
If $P(C_1|x) > 0.5 \rightarrow x \text{ belongs to class 1 (Water)}$

• 分类结果



Blue points: C_1 (Water), Red points: C_2 (Normal)

蓝变红 (水瓶神奇宝物的可能性能逐渐增大)



将训练数据扩展到 7 维, 则准确率可以提升至 64%

3. 概率生成模型的改进

① Parameters Sharing:

当训练数据的维度过大时, 为每一个维度分配一个均值和方差, 可能导致参数过多, 进而 Overfitting。因此可以在不同维度的特征之间进行 Covariance Sharing, 减少模型的参数; 其中五类类别样本的方差按样本数量加权平均

"Water" type Pokémons:

$$x^1, x^2, x^3, \dots, x^{79}$$

$$\mu^1$$

"Normal" type Pokémons:

$$x^{80}, x^{81}, x^{82}, \dots, x^{140}$$

$$\Sigma$$

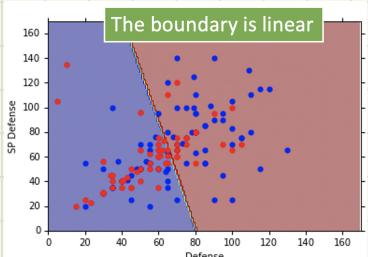
$$\mu^2$$

Find μ^1, μ^2, Σ maximizing the likelihood $L(\mu^1, \mu^2, \Sigma)$

$$L(\mu^1, \mu^2, \Sigma) = f_{\mu^1, \Sigma}(x^1) f_{\mu^1, \Sigma}(x^2) \cdots f_{\mu^1, \Sigma}(x^{79}) \\ \times f_{\mu^2, \Sigma}(x^{80}) f_{\mu^2, \Sigma}(x^{81}) \cdots f_{\mu^2, \Sigma}(x^{140})$$

$$\mu^1 \text{ and } \mu^2 \text{ is the same} \quad \Sigma = \frac{79}{140} \Sigma^1 + \frac{61}{140} \Sigma^2$$

(计算方法与之相同)



4. 数学模型与推导

① 概率生成模型总结:

- Function Set (Model):

$$P(C_1 | x) = \frac{P(x|C_1) P(C_1)}{P(x|C_1) P(C_1) + P(x|C_2) P(C_2)}, \quad \begin{cases} \text{if } P(C_1|x) > 0.5, & x \in \text{class 1} \\ \text{if } P(C_1|x) \leq 0.5, & x \in \text{class 2} \end{cases}$$

其中 $P(C_1), P(C_2), P(x|C_1), P(x|C_2)$ 均为模型的参数, 选择不同的 probability distribution 就会得到不同的 Function Set.

- Goodness of Function (Evaluating Model):

对于类条件概率的分布, 不同的 μ 和 Σ 可以组合成不同的模型。在这些模型中, 需要使用最大似然估计的方法寻找使得类条件概率 $P(x|C_1)$ 最大的 μ 和 Σ , 这样可以得到分类效果较好的模型。

② 概率分布的选择:

对于 $P(x|C_1), P(x|C_2)$, 上述过程均使用的是高斯分布, 但实际上可以选择任何一个概率分布。

假设 x 由 k 个维度的特征组成, 即 $x = [x_1, x_2, \dots, x_k]$, 如果类别 C 产生样本 x 的事件是相互独立的, 则可以使用

Naive Bayes

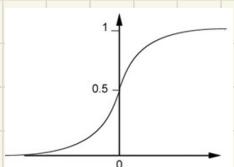
$$P(x|C_1) = P(x_1|C_1) \cdot P(x_2|C_1) \cdot P(x_3|C_1) \cdots \cdots P(x_k|C_1)$$

需要根据特征的具体情况选用分布, 例如 feature 是 binary 的, 则应该选用 Bernoulli Distribution

③ Posterior Probability

$$P(C_1 | x) = \frac{P(x|C_1) P(C_1)}{P(x|C_1) P(C_1) + P(x|C_2) P(C_2)} = \frac{1}{1 + \frac{P(x|C_2) P(C_2)}{P(x|C_1) P(C_1)}}$$

$$\text{令 } z = \frac{P(x|C_2) P(C_2)}{P(x|C_1) P(C_1)}, \text{ 则 } P(C_1 | x) = \frac{1}{1 + \exp(-z)} = \sigma(z) \text{ sigmoid function}$$



$$z = \ln \frac{P(x|C_1)}{P(x|C_2)} + \ln \frac{\frac{N_1}{N_1 + N_2}}{\frac{N_2}{N_1 + N_2}} \Rightarrow \frac{N_1}{N_1 + N_2} = \frac{N_1}{N_2}$$

$$P(x|C_1) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma^1|^{1/2}} \exp \left\{ -\frac{1}{2}(x - \mu^1)^T (\Sigma^1)^{-1} (x - \mu^1) \right\}$$

$$P(x|C_2) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma^2|^{1/2}} \exp \left\{ -\frac{1}{2}(x - \mu^2)^T (\Sigma^2)^{-1} (x - \mu^2) \right\}$$

其中, $\ln \frac{P(x|C_1)}{P(x|C_2)}$ 如下:

$$\begin{aligned} & \ln \frac{\frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma^1|^{1/2}} \exp \left\{ -\frac{1}{2}(x - \mu^1)^T (\Sigma^1)^{-1} (x - \mu^1) \right\}}{\frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma^2|^{1/2}} \exp \left\{ -\frac{1}{2}(x - \mu^2)^T (\Sigma^2)^{-1} (x - \mu^2) \right\}} \\ &= \ln \frac{|\Sigma^2|^{1/2}}{|\Sigma^1|^{1/2}} \exp \left\{ -\frac{1}{2} [(x - \mu^1)^T (\Sigma^1)^{-1} (x - \mu^1) - (x - \mu^2)^T (\Sigma^2)^{-1} (x - \mu^2)] \right. \\ &\quad \left. - (x - \mu^2)^T (\Sigma^2)^{-1} (x - \mu^2) \right\} \\ &= \ln \frac{|\Sigma^2|^{1/2}}{|\Sigma^1|^{1/2}} - \frac{1}{2} [(x - \mu^1)^T (\Sigma^1)^{-1} (x - \mu^1) - (x - \mu^2)^T (\Sigma^2)^{-1} (x - \mu^2)] \\ &= (x - \mu^1)^T (\Sigma^1)^{-1} (x - \mu^1) \\ &= x^T (\Sigma^1)^{-1} x - x^T (\Sigma^1)^{-1} \mu^1 - (\mu^1)^T (\Sigma^1)^{-1} x + (\mu^1)^T (\Sigma^1)^{-1} \mu^1 \\ &= x^T (\Sigma^1)^{-1} x - 2(\mu^1)^T (\Sigma^1)^{-1} x + (\mu^1)^T (\Sigma^1)^{-1} \mu^1 \end{aligned}$$

$$(x - \mu^2)^T (\Sigma^2)^{-1} (x - \mu^2) \\ = x^T (\Sigma^2)^{-1} x - 2(\mu^2)^T (\Sigma^2)^{-1} x + (\mu^2)^T (\Sigma^2)^{-1} \mu^2$$

故 $P(C_1|x) = \sigma(z)$, 其中 z 表示为:

$$\begin{aligned} z &= \ln \frac{|\Sigma^2|^{1/2}}{|\Sigma^1|^{1/2}} - \frac{1}{2} x^T (\Sigma^1)^{-1} x + (\mu^1)^T (\Sigma^1)^{-1} x - \frac{1}{2} (\mu^1)^T (\Sigma^1)^{-1} \mu^1 \\ &\quad + \frac{1}{2} x^T (\Sigma^2)^{-1} x - (\mu^2)^T (\Sigma^2)^{-1} x + \frac{1}{2} (\mu^2)^T (\Sigma^2)^{-1} \mu^2 + \ln \frac{N_1}{N_2} \end{aligned}$$

因为不同维度的特征之间共享协方差矩阵之参数, 即 $\Sigma_1 = \Sigma_2 = \Sigma$, 则 z 可化简为:

$$z = \underbrace{(\mu^1 - \mu^2)^T \Sigma^{-1} x}_{w^T} - \underbrace{\frac{1}{2} (\mu^1)^T \Sigma^{-1} \mu^1 + \frac{1}{2} (\mu^2)^T \Sigma^{-1} \mu^2}_{b} + \ln \frac{N_1}{N_2}$$

$(\text{Vector})^T \cdot (\text{Matrix}) \cdot (\text{Vector}) = \text{标量}$

则后验概率可写为: $P(C_1|x) = \sigma(w \cdot x + b)$, 因此当特征之间是线性的

在生成模型中, 先对 N_1, N_2, μ^1, μ^2 进行估计, 然后计算 w 与 b

然而可以直接受寻参数 w 和 b , 不对分布参数进行估计, 该过程称为 Logistic Regression

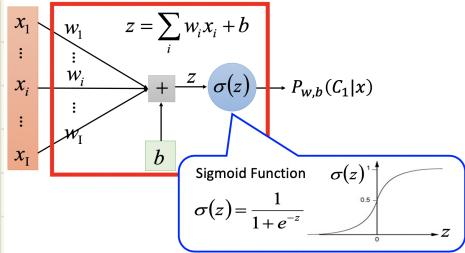
Chapter 5 Classification: Logistic Regression

1. Function Set

①从 Bayes 到 Logistic

在贝叶斯公式中，后验概率 $P(C_i|x) = \pi_i(z)$ ，其中 $\pi_i(z) = \frac{1}{1 + \exp(-z)}$ ， $z = w \cdot x + b = \sum_i w_i x_i + b$ ，不同的 w 和 b 对应不同的方程

将其提取为 Logistic 表达式： $f_{w,b}(x) = P_{w,b}(C_1|x)$



2. Goodness of Function

①最大似然估计

有一组训练数据为： $(x^1, C_1), (x^2, C_1), (x^3, C_2), \dots, (x^n, C_1)$ ，假定这组数据由 $f_{w,b}(x) = P_{w,b}(C_1|x)$ 产生

给定一组 w 和 b ，则其产生训练数据的可靠性为： $L(w,b) = f_{w,b}(x^1) f_{w,b}(x^2) (1 - f_{w,b}(x^3)) \dots f_{w,b}(x^n)$

目标为寻找 w^* 和 b^* 使得 $L(w,b)$ 取最大值

$$w^*, b^* = \arg \max_{w,b} L(w,b) \Leftrightarrow w^*, b^* = \arg \min_{w,b} -\ln L(w,b)$$

当 $x^i \in C_1$ 时， $\hat{y}^i = 1$ ；当 $x^i \in C_2$ 时， $\hat{y}^i = 0$ ，则有

$$\begin{aligned} & -\ln L(w,b) \\ &= -\ln f_{w,b}(x^1) \rightarrow -[\hat{y}^1 \ln f(x^1) + (1 - \hat{y}^1) \ln(1 - f(x^1))] \rightarrow -[1 \ln f(x^1) + 0 \ln(1 - f(x^1))] \\ & -\ln f_{w,b}(x^2) \rightarrow -[\hat{y}^2 \ln f(x^2) + (1 - \hat{y}^2) \ln(1 - f(x^2))] \rightarrow -[0 \ln f(x^2) + 0 \ln(1 - f(x^2))] \\ & -\ln(1 - f_{w,b}(x^3)) \rightarrow -[\hat{y}^3 \ln f(x^3) + (1 - \hat{y}^3) \ln(1 - f(x^3))] \rightarrow -[0 \ln f(x^3) + 1 \ln(1 - f(x^3))] \\ & \vdots \end{aligned}$$

$$\text{故 } w^*, b^* = \arg \min_{w,b} -\ln L(w,b) = \arg \min_{w,b} \sum_n -[\hat{y}^n \ln f_{w,b}(x^n) + (1 - \hat{y}^n) \ln(1 - f_{w,b}(x^n))]$$

Cross Entropy between two Bernoulli Distribution
交叉熵表示两个分布有多相似

Distribution p:
 $p(x=1) = \hat{y}^n$
 $p(x=0) = 1 - \hat{y}^n$

cross entropy

Distribution q:
 $q(x=1) = f(x^n)$
 $q(x=0) = 1 - f(x^n)$

$$H(p, q) = - \sum_x p(x) \ln(q(x))$$

② 寻找最优解

计算 $-\ln L(w, b)$ 关于 w_i 的偏导数， $-\ln L(w, b) = \sum_n -[\hat{y}^n \ln f_{w, b}(x^n) + (1 - \hat{y}^n) \ln (1 - f_{w, b}(x^n))]$, $f_{w, b}(x) = \sigma(z) = \frac{1}{1 + \exp(-z)}$

$$\frac{\partial \ln L(w, b)}{\partial w_i} = \sum_n -\left[\hat{y}^n \cdot \frac{\partial \ln f_{w, b}(x^n)}{\partial w_i} + (1 - \hat{y}^n) \cdot \frac{\partial \ln (1 - f_{w, b}(x^n))}{\partial w_i} \right] \quad z = w \cdot x + b = \sum_i w_i x_i + b$$

$$\begin{aligned} \frac{\partial \ln f_{w, b}(x^n)}{\partial w_i} &= \frac{\partial \ln f_{w, b}(x)}{\partial z} \cdot \frac{\partial z}{\partial x} \\ &= \frac{\partial \ln (\sigma(z))}{\partial z} \cdot x_i \\ &= \frac{1}{\sigma(z)} \cdot \sigma'(z) (1 - \sigma(z)) \cdot x_i \\ &= (1 - \sigma(z)) \cdot x_i \\ &= (1 - f_{w, b}(x^n)) \cdot x_i^n \end{aligned}$$

$$\begin{aligned} \frac{\partial \ln (1 - f_{w, b}(x^n))}{\partial w_i} &= \frac{\partial \ln (1 - \sigma(z))}{\partial z} \cdot \frac{\partial z}{\partial x} \\ &= \frac{\partial \ln (1 - \sigma(z))}{\partial z} \cdot x_i \\ &= \frac{1}{1 - \sigma(z)} \cdot \sigma'(z) (1 - \sigma(z)) \cdot x_i \\ &= \sigma(z) \cdot x_i \\ &= f_{w, b}(x^n) \cdot x_i^n \end{aligned}$$

$$\begin{aligned} \frac{\partial \ln L(w, b)}{\partial w_i} &= \sum_n -\left[\hat{y}^n \cdot (1 - f_{w, b}(x^n)) \cdot x_i^n + (1 - \hat{y}^n) \cdot f_{w, b}(x^n) \cdot x_i^n \right] \\ &= \sum_n -\left[\hat{y}^n - \hat{y}^n \cdot f_{w, b}(x^n) + f_{w, b}(x^n) - \hat{y}^n \cdot f_{w, b}(x^n) \right] \\ &= \sum_n -(\hat{y}^n - f_{w, b}(x^n)) x_i^n \end{aligned}$$

将梯度应用于梯度下降，则 $w_i \leftarrow w_i - \eta \sum_n -(\hat{y}^n - f_{w, b}(x^n)) x_i^n$

当前输出与目标值差距越大，当前的更新量越大

3. Logistic & Linear Regression

① 区别与联系

	Logistic Regression (Discriminative)	Linear Regression (Generative)
Function Set	$f_{w, b}(x) = \sigma(\sum_i w_i x_i + b)$ 输出在 $(0, 1)$ 之间	$f_{w, b}(x) = \sum_i w_i x_i + b$ 输出值无具体限制
Loss Function	$\hat{y}^n : 1 \text{ for } C_1, 0 \text{ for } C_2$ $L(f) = \sum_n C(f(x^n), \hat{y}^n)$ 其中 $C(f(x^n), \hat{y}^n) = -[\hat{y}^n \ln f(x^n) + (1 - \hat{y}^n) \ln (1 - f(x^n))]$	$\hat{y}^n : \text{一个真实数字}$ $L(f) = \frac{1}{2} \sum_n (f(x^n) - \hat{y}^n)^2$
Gradient Descent	$w_i \leftarrow w_i - \eta \cdot \sum_n -(\hat{y}^n - f_{w, b}(x^n)) x_i^n$	$w_i \leftarrow w_i - \eta \cdot \sum_n -(\hat{y}^n - f_{w, b}(x^n)) x_i^n$

② Logistic 为什么不用 Square Error

$$f_{w,b}(x) = \sigma\left(\sum_i w_i x_i + b\right)$$

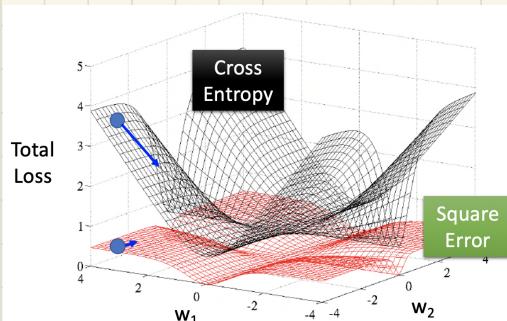
损失函数梯度更新时梯度为 $L(f) = \frac{1}{n} \sum_i (f_{w,b}(x^i) - \hat{y}^i)^2$

$$\frac{\partial (f_{w,b}(x^i) - \hat{y}^i)^2}{\partial w_i} = 2(f_{w,b}(x^i) - \hat{y}^i) \cdot \frac{\partial f_{w,b}(x^i)}{\partial w_i}$$

$$= 2(f_{w,b}(x^i) - \hat{y}^i) \cdot f_{w,b}(x^i) \cdot (1 - f_{w,b}(x^i)) x_i$$

当 $\hat{y}^i = 1$ 时，
 {
 如果某一步迭代时， $f_{w,b}(x^i) = 1$ (离目标点很近) $\rightarrow \frac{\partial L}{\partial w_i} = 0$
 如果某一步迭代时， $f_{w,b}(x^i) = 0$ (离目标点很远) $\rightarrow \frac{\partial L}{\partial w_i} = 0$
 } 梯度变化不够明显

当 $\hat{y}^i = 0$ 时，同上



• 对于 Cross Entropy：最优点和其邻点的梯度差异很大，有利于梯度下降的更新

• 对于 Square Error：所有点之间的梯度差异很小，任意取一点作为初始点都可能导致梯度更新过慢或停止更新

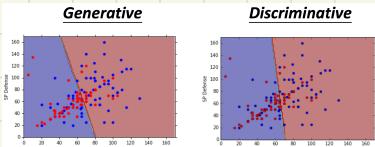
③ 判别模型 (Discriminative) VS. 生成模型 (Generative)

$$P(C_i|x) = \sigma(w \cdot x + b)$$

判别 \rightarrow 直接找出 w 和 b

生成 \rightarrow 估计 $\mu^1, \mu^2, \Sigma^{-1}$ \rightarrow 在高斯/博努尔分布/Naive Bayes 假设下计算 w 和 b

两组 w 和 b 是不相同的



一般情况下，判别模型比生成模型的效果更好一些

对于该组训练数据 (共 13 个样本，每样本两个特征；第一个样本属于 Class 1，其余属于 Class 2)

基测试集为 $\begin{pmatrix} 1 \\ 1 \end{pmatrix}$ ，使用 Naive Bayes 得出的结论是属于 Class 2

$$P(C_1) = \frac{1}{13} \quad P(x_1 = 1|C_1) = 1 \quad P(x_2 = 1|C_1) = 1$$

$$P(C_2) = \frac{12}{13} \quad P(x_1 = 1|C_2) = \frac{1}{3} \quad P(x_2 = 1|C_2) = \frac{1}{3}$$

Training Data	Class 1	Class 2	Class 2	Class 2
1 1	1 1	0 0	0 1	0 0
	X 4	X 4	X 4	X 4

$$P(C_1|x) = \frac{P(x|C_1)P(C_1)}{P(x|C_1)P(C_1) + P(x|C_2)P(C_2)}$$

$$= \frac{1 \times 1}{1 \times 1 + \frac{1}{3} \times \frac{1}{3}} = \frac{1}{\frac{1}{3} + \frac{1}{9}} = \frac{1}{\frac{4}{9}} = \frac{9}{4} < 0.5$$

生成模型的优点：

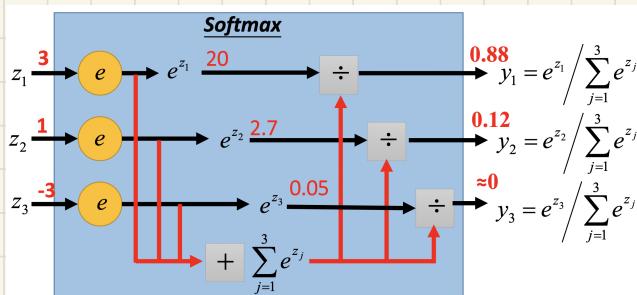
- 由于概率分布假设的存在，生成模型可能在小数据量的效果会超过判别模型。判别模型不存在假设
- 只能从数据中学习，故需要更大量的数据
- 由于概率分布假设的存在，生成模型对噪声的鲁棒性更好
- 先验概率和类条件概率可以以不同的数据源进行估计

④ Multi-class Classification (以三分类为例)

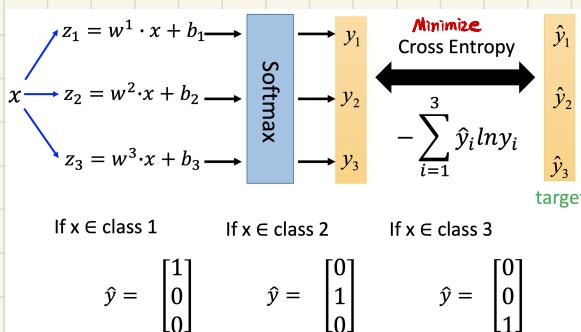
$$C_1: w^1, b_1 \quad z_1 = w^1 \cdot x + b_1$$

$$C_2: w^2, b_2 \quad z_2 = w^2 \cdot x + b_2$$

$$C_3: w^3, b_3 \quad z_3 = w^3 \cdot x + b_3$$



- $0 < y_i < 1$
- $\sum_i y_i = 1$
- softmax对大的值进行线性化，通过取自然指数拉开不同值之间的差距

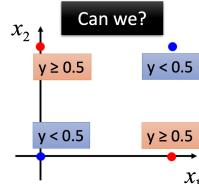


⑤ Limitation of Logistic Regression

$$z = w_1x_1 + w_2x_2 + b$$

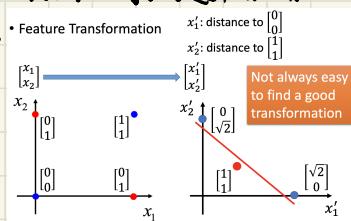
$$\begin{cases} Class 1 & y \geq 0.5 \\ Class 2 & y < 0.5 \end{cases}$$

Input Feature		Label
x ₁	x ₂	Class 2
0	0	Class 2
0	1	Class 1
1	0	Class 1
1	1	Class 2

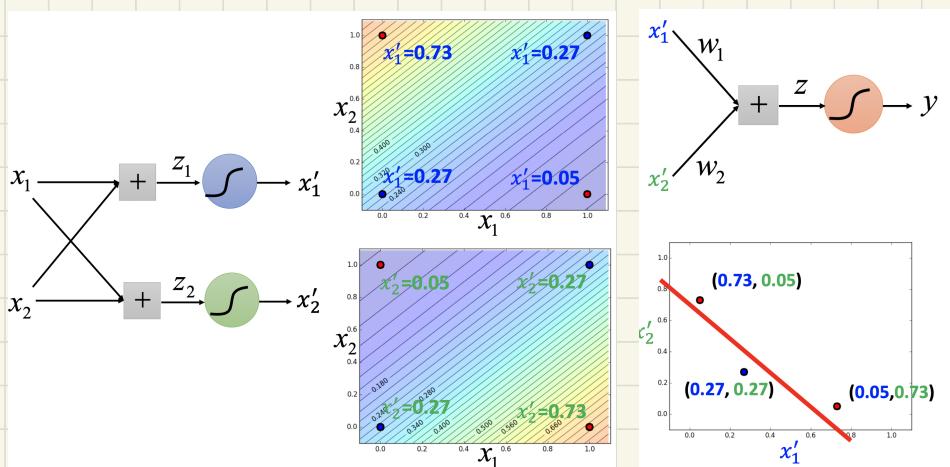
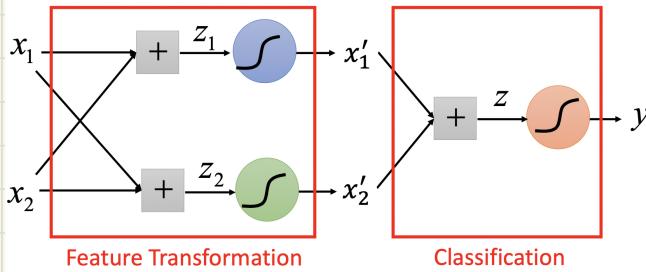


只能解决线性可分问题，对于线性不可分问题需要使用
Feature Transformation

Feature Transformation



Feature Transformation 可以人工进行定义，也可以用 Logistic Regression 叠加进行实现



通过叠加 Logistic Regression 实现的分类方法就是最基本的 Neural Network

