Classification

$x \longrightarrow$ function $\longrightarrow$ class n

应用: ① Credit Scoring
② Medical Diagnosis
③ Handwritten character recognition
④ face recognition

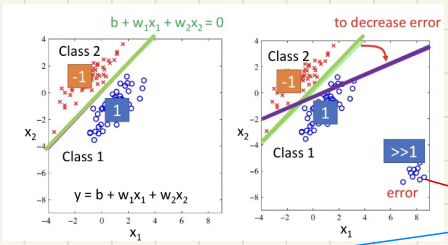
example: class: 属于哪种属性.

How to do classification?

training data for classification

considering Binary classification 二分类 as example

Training: class 1 means the target is 1; class 2 means the target is -1
Testing: closer to 1 $\longrightarrow$ class 1; closer to -1 $\longrightarrow$ class 2



若用 classification, 则分界线是绿色的; 若用 regression, 则是紫色的. 因为 regression 会惩罚这些点

所以要重改!!! 不能用 regression

Ideal Alternatives 理想的方法

- function (model):

$$x \Longrightarrow \begin{cases} g(x) > 0 & \text{output} = \text{class 1} \\ \text{else} & \text{output} = \text{class 2} \end{cases}$$

- Loss function

$$L(f) = \sum_n \delta(f(x^n) \neq \hat{y}^n)$$ 无法行数分

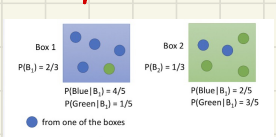
The number of times f gets incorrect results on training data.

- find the best function

example: perceptron 感知机 sum

怎么去寻找 the best function

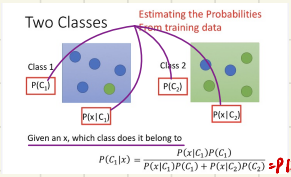
Two Boxes



$$P(B_1|Blue) = \frac{P(Blue|B_1)P(B_1)}{P(Blue)}$$

全概率公式

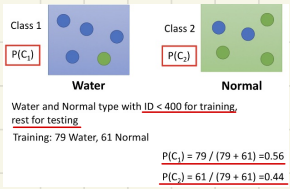
Two classes



Generative model $P(x) = P(x|C_1)P(C_1) + P(x|C_2)P(C_2)$

Then we to compute

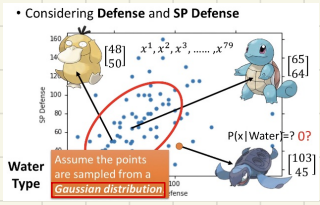
Prior



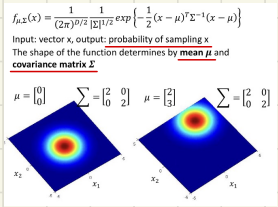
P(x|C) ??

each Pokémon is represented as a vector by its attribute. ⟹ feature

⇓ probability from class - feature



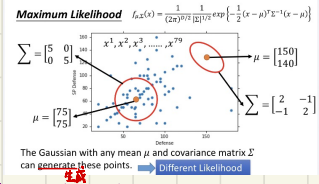
Gaussian Distribution



⇓

根据79个点求得 μ、Σ 则 $f_{\mu,\Sigma}(x)$ 可写出来

怎么计算 μ、Σ = Maximum Likelihood



Likelihood of a Gaussian with mean μ and covariance matrix Σ

= the probability of the Gaussian samples $x^1\ x^2\ x^3 \cdots x^{79}$

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

⇓

We assume $x^1, x^2, \cdots x^{79}$ generate from the Gaussian (μ^*, Σ^*) with the maximum likelihood.

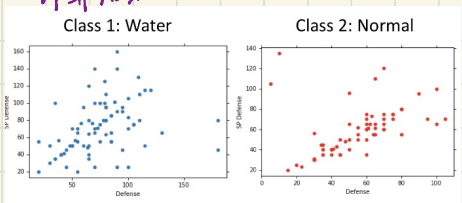
$L(\mu,\Sigma) = f_{\mu,\Sigma}(x^1) f_{\mu,\Sigma}(x^2) \cdots f_{\mu,\Sigma}(x^{79})$

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

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

$\mu^* = \frac{1}{79}\sum x^n \qquad \Sigma^* = \frac{1}{79}\sum (x^n-\mu^*)(x^n-\mu^*)^T$

计算结果



$\mu^1 = \begin{bmatrix}75.0\\71.3\end{bmatrix}$ $\Sigma^1 = \begin{bmatrix}874 & 327\\327 & 929\end{bmatrix}$ $\mu^2 = \begin{bmatrix}55.6\\59.8\end{bmatrix}$ $\Sigma^2 = \begin{bmatrix}847 & 422\\422 & 685\end{bmatrix}$

Now we can do classification

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

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

P(C1) = 79 / (79 + 61) =0.56

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

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

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

P(C2) = 61 / (79 + 61) =0.44

If $P(C_1|x) > 0.5 \Rightarrow x$ belongs to class 1 (water)

How's the results?

Testing data: 47% accuracy

All: total, hp, att, sp att, de, sp de, speed (7 features)

μ^1, μ^2: 7-dim vector

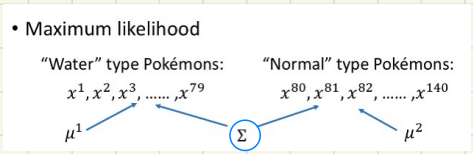
Σ^1, Σ^2: 7 x 7 matrices

54% accuracy ... ☹

事实上每个宝可梦有7个特征，可能在7维空间内从正常识别，但结果仍然是54%的正确率

Modifying Model

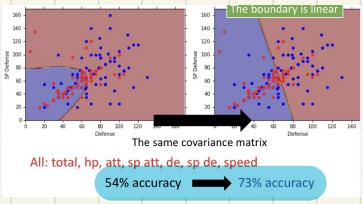
将不同类别的协方差设为相同的 Σ，原因：由于一个协方差矩阵的参数是 x^2 个，如果设置不同，就会造成参数过多。



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

μ^1 μ^2 与之前算法一致 $\Sigma = \frac{79}{140}\Sigma^1 + \frac{61}{140}\Sigma^2$

result:



conclusion:

① Three step

- Function Set (Model)

$x \longrightarrow$

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

if $P(C_1|x) > 0.5$ output: class 1

otherwise: output: class 2.

- Goodness of a function

The mean μ and covariance Σ that maximizing the likelihood (the probability of generating data)

- Find the best function: easy

question? why choose this probability distribution

You can always use the distribution you like

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

$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_k \\ \vdots \\ x_K \end{bmatrix}$

1-D Gaussian

For binary features, you may assume they are from Bernoulli distributions.

二分类

伯努利分布

If you assume all the dimensions are independent, then you are using Naive Bayes Classifier.

朴素贝叶斯

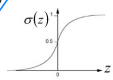
Posterior Probability 后验概率

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

$$= \frac{1}{1 + \frac{P(x|C_2)P(C_2)}{P(x|C_1)P(C_1)}} = \frac{1}{1+exp(-z)} = \sigma(z)$$

Sigmoid function

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



$P(C_1|x) = \sigma(z)$ $\quad z = ln\frac{P(x|C_1)P(C_1)}{P(x|C_2)P(C_2)}$

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

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

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

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

$$= ln\frac{|\Sigma^2|^{1/2}}{|\Sigma^1|^{1/2}} exp\left\{-\frac{1}{2}\left[(x-\mu^1)^T(\Sigma^1)^{-1}(x-\mu^1) - (x-\mu^2)^T(\Sigma^2)^{-1}(x-\mu^2)\right]\right\}$$

$$= ln\frac{|\Sigma^2|^{1/2}}{|\Sigma^1|^{1/2}} - \frac{1}{2}\left[(x-\mu^1)^T(\Sigma^1)^{-1}(x-\mu^1) - (x-\mu^2)^T(\Sigma^2)^{-1}(x-\mu^2)\right]$$

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

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

$\Sigma^1 = \Sigma^2 = \Sigma$

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

$P(C_1|x) = \sigma(w \cdot x + b)$ 为什么 boundary 是 linear

$P(C_1|x) = \sigma(w \cdot x + b)$ How about directly find **w** and b?

In generative model, we estimate $N_1, N_2, \mu^1, \mu^2, \Sigma$

Then we have **w** and b

生成模型

模型表示了给定输入 X 产生输出 Y 的生成关系，基于统计学与 Bayes 理论