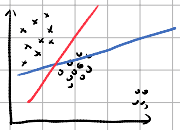
Classification

1. Regression 和 Classification 的区别

eg.



blue: regression
red: classification

2. Ideal Alternatives

Function (Model)

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

Loss function

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

Find the best function
eg. SVM ; Perception

3. Generative Model

※ Bayes Formula

Class 1  Class 2 

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

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

假设样本各 feature 取值服从 Gaussian Distribution

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

※使用 Maximum Likelihood 求解参数

推导: $L(\mu,\Sigma) = \left(\frac{1}{\sqrt{(2\pi)^D \det(\Sigma)}}\right)^n \exp\left\{\sum_{i=1}^n -\frac{1}{2}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)\right\}$

$$\ln L(\mu,\Sigma) = -\frac{n}{2}\ln(2\pi)^D\det(\Sigma) - \frac{1}{2}\sum_{i=1}^n (x_i-\mu)^T\Sigma^{-1}(x_i-\mu)$$

$$\frac{\partial \ln L}{\partial \mu} = \sum_{i=1}^n \Sigma^{-1}(x_i-\mu) = 0 \Rightarrow \mu = \frac{1}{n}\sum_{i=1}^n x_i$$

$\frac{\partial \ln L}{\partial \Sigma} = -\frac{n}{2}\Sigma^{-1} + \frac{1}{2}\Sigma^{-1}\sum_{i=1}^{n}(x_i-\mu)(x_i-\mu)^T\Sigma \Rightarrow \Sigma^* = \frac{1}{n}\sum_{i=1}^{n}(x_i-\hat{\mu})(x_i-\hat{\mu})^T$

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

其中 $x|C_1 \sim N(\mu, \Sigma)$ 用上述 ML 估计求出以、

$P(C_1) = \frac{num(C_1)}{\sum num(C_i)}$

改进：① 减少参数，以避免 overfitting

假定不同 Class 分布 μ 不同，但 Σ 相同

find μ^1, μ^2, Σ

$L(\mu^1, \mu^2, \Sigma) = \left(\frac{1}{(2\pi)^D det(\Sigma)}\right)^{\frac{n_1+n_2}{2}} \exp -\frac{1}{2}\{\sum_{i=1}^{n_1}(x_i^1-\mu^1)^T\Sigma^{-1}(x_i^1-\mu^1) + \sum_{i=1}^{n_2}(x_i^2-\mu^2)^T\Sigma^{-1}(x_i^2-\mu^2)\}$

$L(\mu^1, \mu^2, \Sigma) = -\frac{n_1+n_2}{2}\ln(2\pi)^D det(\Sigma) - \frac{1}{2}[\sum_{i=1}^{n_1}(x_i^1-\mu^1)^T\Sigma^{-1}(x_i^1-\mu^1) + \sum_{i=1}^{n_2}(x_i^2-\mu^2)^T\Sigma^{-1}(x_i^2-\mu^2)]$

$\frac{\partial L}{\partial \mu^1} = \sum_{i=1}^{n_1}\Sigma^{-1}(x_i^1-\mu^1) \Rightarrow \mu^1 = \frac{1}{n_1}\sum_{i=1}^{n_1}x_i^1 \qquad \frac{\partial L}{\partial \mu^2} = \sum_{i=1}^{n_2}\Sigma^{-1}(x_i^2-\mu^2) \Rightarrow \mu^2 = \frac{1}{n_2}\sum_{i=1}^{n_2}x_i^2$

$\frac{\partial L}{\partial \Sigma} = -\frac{n_1+n_2}{2}\Sigma + \frac{1}{2}\sum_{i=1}^{n_1}(x_i^1-\mu^1)(x_i^1-\mu^1)^T + \frac{1}{2}\sum_{i=1}^{n_2}(x_i^2-\mu^2)(x_i^2-\mu^2)^T \Rightarrow \Sigma^* = \frac{1}{n_1+n_2}(\sum_{i=1}^{n_1}(x_i^1-\mu^1)(x_i^1-\mu^1)^T + \sum_{i=1}^{n_2}(x_i^2-\mu^2)(x_i^2-\mu^2)^T)$

② Naive Bayes Classifier

假定不同 feature 间独立

③ 与线性联系

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

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

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

$P(x|C_1) = \frac{1}{\sqrt{(2\pi)^D|\Sigma|}}\exp\{-\frac{1}{2}(x-\mu^1)^T[\Sigma^{-1}](x-\mu^1)\} \qquad P(x|C_2) = \frac{1}{\sqrt{(2\pi)^D|\Sigma|}}\exp\{-\frac{1}{2}(x-\mu^2)^T[\Sigma^{-1}](x-\mu^2)\}$

$\therefore \ln\frac{P(x|C_1)}{P(x|C_2)} = \ln\sqrt{\frac{|\Sigma^2|}{|\Sigma^1|}}\exp\{-\frac{1}{2}[(x-\mu^1)^T(\Sigma^1)^{-1}(x-\mu^1) - (x-\mu^2)^T(\Sigma^2)^{-1}(x-\mu^2)]\}$

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

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

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

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

一般会 $\Sigma^1 = \Sigma^2 = \Sigma$

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

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

$P(x|C_1) = \sigma(w\cdot x + b)$

4. Logistic Regression

(1) Function Set

$f_{w,b}(x) = P(C_1|x) = \sigma(w\cdot x + b) \quad \in (0,1)$

(2) Goodness of a Function

Training Data : x^1 x^2 x^3 $\cdots$ x^N

C_1 C_1 C_2 $\cdots$ C_1

$L(w,b) = f_{w,b}(x^1)\, f_{w,b}(x^2)\,(1-f_{w,b}(x^3))\cdots f_{w,b}(x^N)$

$w^*, b^* = \arg\max_{w,b} L(w,b)$

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

$= -\ln f_{w,b}(x^1) - \ln f_{w,b}(x^2) - \ln(1-f_{w,b}(x^3)) - \cdots - \ln f_{w,b}(x^N)$

代入 $\hat{y}$ 令 $x^1 : \hat{y}^1 = 1$ ，$x^2 : \hat{y}^2 = 1$ ，$x^3 : \hat{y}^3 = 0$ $\cdots$

每一项变为 $-[\hat{y}^i \ln f(x^i) + (1-\hat{y}^i)\ln(1-f(x^i))]$

$= -\sum_{i=1}^{N}[\hat{y}^i \ln f(x^i) + (1-\hat{y}^i)\ln(1-f(x^i))]$ Cross Entropy

P:

$P(x=1) = \hat{y}^n$

$P(x=0) = 1-\hat{y}^n$

Cross Entropy $= -\sum_x p(x)\log q(x)$

q:

$q(x=1) = f(x^n)$

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

13) Optimize

$$-\frac{\partial \ln L(w,b)}{\partial w} = \sum_n -\left[\hat{y}^n \frac{\partial \ln f(w,b)}{\partial w} + (1-\hat{y}^n)\frac{\partial \ln(1-f(w,b))}{\partial w}\right]$$

其中 $\frac{\partial \ln f(w,b)}{\partial w} = \frac{\partial \ln \sigma(w\cdot x+b)}{\partial w} = \frac{1}{\sigma(w\cdot x+b)}\frac{\partial \sigma(w\cdot x+b)}{\partial w} = \frac{1}{\sigma(w\cdot x+b)}\cdot\sigma(w\cdot x+b)(1-\sigma(w\cdot x+b))\cdot x = [1-\sigma(w\cdot x+b)]x$

$$\frac{\partial \ln(1-f(w,b))}{\partial w} = -\frac{1}{1-f(w,b)}\cdot f(w,b)[1-f(w,b)]\cdot x = -f(w,b)x$$

$$= \sum_n -\left[\hat{y}^n(1-f_{w,b}(x^n))x^n - (1-\hat{y}^n)f_{w,b}(x^n)x^n\right]$$

$$= \sum_n -(\hat{y}^n - f_{w,b}(x^n))x^n$$

$$\therefore w^{i+1} \leftarrow w^i + \eta\sum_n(\hat{y}^n - f_{w,b}(x^n))x^n$$

hint: ① 不同 square loss 原因

$$L(f) = \frac{1}{2}\sum_n (f_{w,b}(x^n) - \hat{y}^n)^2$$

$$\frac{\partial (f_{w,b}(x^n)-\hat{y}^n)^2}{\partial w} = 2(f_{w,b}(x^n)-\hat{y}^n)f_{w,b}(x^n)(1-f_{w,b}(x^n))x^n$$

if $\hat{y}^n = 0$ 或 $f_{w,b}(x^n) = 1$ update $\frac{\partial L}{\partial w} = 0$

② 与 Generative model 关系.

都为 $P(C_1|x) = \sigma(w\cdot x+b)$ (shared Σ)

discriminative model: directly find w, b

generative model: find $\mu^1, \mu^2, \Sigma^{-1}$, then calculate w, b

w, b 不同: generative model: 带有一些假设

discriminative model: 可能局部最优

一般 discriminative model 效果好一些

③ multi-class classification

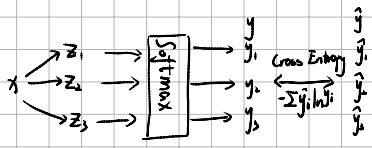
原理同上

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

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

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

Softmax $\quad y_i = \frac{e^{z_i}}{\sum_j e^{z_j}}$



class 1 $\hat{y} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$ class 2 $\begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$ class 3 $\begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$

· 这 boundary 为一条直线

e.g. 无法分类!!!

Pro: x_1 x_2 z_1 z_2 a_1' a_2' z y

Feature Transformation　　Classification

∴每个 Neuron 是一个 Logistic Regression