1. 背景.

linear regression $y=w^Tx+b$

- 线性 (特征.) —✗—
  - 属性非线性 (特征转换, 多项式回归)
  - 全局非线性 (线性分类, 激活函数非线性)
  - 系数非线性 (神经网络)
- 全局性 —✗— 线性样条回归, 决策树.
- 数据未加工 —✗— PCA, 流形.

线性回归 $\xrightarrow[\text{降维}]{\text{激活函数}}$ 线性分类 $\Rightarrow$

- $y=f(w^Tx+b) \in$
  - $\{0,1\}$ 硬分类 (线性判别分析, 感知机)
  - $[0,1]$ 软分类
    - 生成式 (GDA, Naive Bayes)
    - 判别式 (Logestic Regression)
- f: activation function, f^{-1}: link function

2. 感知机算法. (假设线性可分).

* 思想: 错误驱动.

* 模型: $f(x)=\text{sign}(w^Tx)$. $x\in\mathbb{R}^p$, $w\in\mathbb{R}^p$

$$\text{sign}(a)=\begin{cases}1 & a\geq 0\\ -1 & a<0\end{cases}$$ 符号函数.

* Loss function: 定义: D: {被错误分类的样本}. 样本集: $\{(x_i,y_i)\}_{i=1}^N$

① $L(w)=\sum_{i=1}^{N} I\{y_iw^Tx_i<0\}$. (错误分类点的个数).

不连续, 不可导.

② $L(w)=\sum_{x_i\in D} -y_iw^Tx_i$ (可看为错误分类点到超平面距离)

$\nabla_w L = -y_ix_i$

* 算法: SGD: $w^{(t+1)} \leftarrow w^{(t)}-\lambda\nabla_w L = w^{(t)}+\lambda y_ix_i$, λ: learning rate.

pocket algorithm (线性不可分情况): 考虑权重更新前后, 错分类点个数.

3. 线性 (Fisher) 判别分析.

* 符号定义: $X=(x_1,\cdots,x_N)^T=\begin{bmatrix}x_1^T\\ \vdots\\ x_N^T\end{bmatrix}_{N\times p}$ 样本阵. $Y=\begin{bmatrix}y_1\\ \vdots\\ y_N\end{bmatrix}_{N\times 1}$.

$\{(x_i,y_i)\}_{i=1}^N$: $x_i\in\mathbb{R}^p$, $y_i\in\{-1,1\}$ 样本集合.

$x_{C_1}=\{x_i\mid y_i=1\}$. $x_{C_2}=\{x_i\mid y_i=-1\}$.

$|x_{C_1}|=N_1$, $|x_{C_2}|=N_2$. $N_1+N_2=N$.

* 思想: 投影后, 样本的类内距离小, 类间距离大.

C_1: $\bar{z}_1=\frac{1}{N_1}\sum_{i=1}^{N_1} w^Tx_i$

$S_1=\frac{1}{N_1}\sum_{i=1}^{N_1}(w^Tx_i-\bar{z}_1)(w^Tx_i-\bar{z}_1)^T$

C_2: $\bar{z}_2=\frac{1}{N}\sum_{i=1}^{N_2} w^Tx_i$

$$S_2 = \frac{1}{N_2}\sum_{i=1}^{N_2}(w^Tx_i - \bar{z}_2)(w^Tx_i - \bar{z}_2)^T$$

投影后类间距离： $(\bar{z}_1-\bar{z}_2)^2$

投影后类内距离： S_1+S_2

* 目标函数： $J(w)=\frac{(\bar{z}_1-\bar{z}_2)^2}{S_1+S_2} \Longrightarrow \hat{w}=\arg\max_w J(w)$

* 求解： 分子： $(\bar{z}_1-\bar{z}_2)^2=[\frac{1}{N_1}\sum_{i=1}^{N_1}w^Tx_i - \frac{1}{N_2}\sum_{i=1}^{N_2}w^Tx_i]^2$

$$= [w^T(\frac{1}{N_1}\sum_{i=1}^{N_1}x_i - \frac{1}{N_2}\sum_{i=1}^{N_2}x_i)]^2$$

$$= [w^T(\bar{x}_{c_1}-\bar{x}_{c_2})]^2 = w^T(\bar{x}_{c_1}-\bar{x}_{c_2})(\bar{x}_{c_1}-\bar{x}_{c_2})^T w$$

分母： $S_1 = \frac{1}{N_1}\sum_{i=1}^{N_1}(w^Tx_i - \frac{1}{N_1}\sum_{j=1}^{N_1}w^Tx_j)(w^Tx_i-\frac{1}{N_1}\sum_{j=1}^{N_1}w^Tx_j)^T$

$$= \frac{1}{N_1}\sum_{i=1}^{N_1} w^T(x_i-\bar{x}_{c_1})(x_i-\bar{x}_{c_1})^T w$$

$$= w^T[\frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{x}_{c_1})(x_i-\bar{x}_{c_1})^T]w = w^TS_{c_1}w$$

$$S_1+S_2 = w^T(S_{c_1}+S_{c_2})w$$

$$J(w)=\frac{w^T(\bar{x}_{c_1}-\bar{x}_{c_2})(\bar{x}_{c_1}-\bar{x}_{c_2})^Tw}{w^T(S_{c_1}+S_{c_2})w} = (w^TS_bw)(w^TS_ww)^{-1}$$

定义： $S_b = (\bar{x}_{c_1}-\bar{x}_{c_2})(\bar{x}_{c_1}-\bar{x}_{c_2})^T$ 类间方差

$S_w = S_{c_1}+S_{c_2}$ ： 类内方差

$$\frac{\partial J(w)}{\partial w} = 2S_bw(w^TS_ww)^{-1} + (w^TS_bw)\cdot(-1)(w^TS_ww)^{-2}\cdot 2S_w\cdot w = 0$$

$$\Rightarrow S_bw(w^TS_ww) - w^TS_bw\,S_w\cdot w = 0$$

w: $p\times1$
w^T: $1\times p$
S_w: $p\times p$
S_b: $p\times p$.

$$\Rightarrow \underbrace{w^TS_bw}_{\in\mathbb{R}}\,S_w\cdot w = S_bw\underbrace{(w^TS_ww)}_{\in\mathbb{R}}$$

$$\Rightarrow S_w\cdot w = \frac{w^TS_ww}{w^TS_bw}S_bw.$$

由于我们只关心 w 的方向，而不关心其大小.

$$\Rightarrow w = \frac{w^TS_ww}{w^TS_bw}S_w^{-1}S_bw. \propto S_w^{-1}S_b\cdot w = S_w^{-1}(\bar{x}_{c_1}-\bar{x}_{c_2})(\bar{x}_{c_1}-\bar{x}_{c_2})^Tw$$

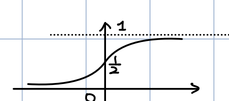
$$\propto S_w^{-1}(\bar{x}_{c_1}-\bar{x}_{c_2})$$

若 S_w：对角，各项同性， $S_w^{-1}\propto I$ ，则 $w\propto(\bar{x}_{c_1}-\bar{x}_{c_2})$

4. Logestic 回归.

* 符号： 样本集 $\{(x_i,y_i)\}_{i=1}^N$ ， $x_i\in\mathbb{R}^p$. $y_i\in\{0,1\}$.

* Sigmoid Function ： $\sigma(z)=\frac{1}{1+e^{-z}}$



$\mathbb{R}\longmapsto(0,1)$
$w^Tx\longmapsto p$

* 模型： $p_1\triangleq p(y=1|x)=\sigma(w^Tx)=\frac{1}{1+\exp(-w^Tx)}$

$$p_0\triangleq p(y=0|x)=1-\sigma(w^Tx)=1-\frac{1}{1+\exp(-w^Tx)}=\frac{\exp(-w^Tx)}{1+\exp(-w^Tx)}$$

$\Rightarrow p(y|x) = p_1^y p_0^{1-y}$

* MLE : $\hat{w} = \underset{w}{\operatorname{argmax}} \log P(Y|X)$

$= \underset{w}{\operatorname{argmax}} \sum_{i=1}^{N} \log P(y_i|x_i)$

$= \underset{w}{\operatorname{argmax}} \underbrace{\sum_{i=1}^{N} y_i \log \sigma(w^T x_i) + (1-y_i) \log (1-\sigma(w^T x_i))}_{-\text{cross Entropy}} \triangleq -J(w)$

(max MLE ⇔ min loss function (min cross Entropy))

$\frac{\partial J(w)}{\partial w} = -\sum_{i=1}^{N} [y_i \frac{1}{\sigma(w^T x_i)} \cdot \sigma'(w^T x_i) - (1-y_i) \frac{1}{1-\sigma(w^T x_i)} \cdot \sigma'(w^T x_i)]$

$= -\sum_{i=1}^{N} [y_i \cdot \frac{1}{\sigma(w^T x_i)} - (1-y_i) \frac{1}{1-\sigma(w^T x_i)}] \cdot \sigma'(w^T x_i)$ $\quad [\sigma'(w^T x) = \sigma(w^T x)(1-\sigma(w^T x)) \cdot x]$

$= -\sum_{i=1}^{N} [y_i (1-\sigma(w^T x_i)) - (1-y_i) \sigma(w^T x_i)] \cdot x_i$

$= -\sum_{i=1}^{N} [y_i - \sigma(w^T x_i)] \cdot x_i = 0$

可用 SGD 更新权重 (Loss function = $J(w)$).

$w^{(t+1)} \leftarrow w^{(t)} - \lambda \nabla_w J(w) = w^{(t)} + \eta \sum_{i=1}^{N} [y_i - \sigma(w^T x_i)] x_i$ $\quad \lambda$: learning rate.

5. Gaussian Discriminant Analysis

* 符号: $\{(x_i, y_i)\}_{i=1}^{N}$, $x_i \in \mathbb{R}^p$, $y_i \in \{0,1\}$, $|\{x_i | y_i = 1\}| = N_1$, $|\{x_i | y_i = 0\}| = N_2$, $N = N_1 + N_2$

* 假设: $y \sim \text{Bernoulli}(\phi) \Rightarrow$

y	1	0
p	ϕ	$1-\phi$

$p(y) = \phi^y (1-\phi)^{1-y}$

$x|y=1 \sim N(\mu_1, \Sigma)$. $\quad x|y=0 \sim N(\mu_2, \Sigma)$. $\quad p(x|y) = N(\mu_1, \Sigma)^y N(\mu_2, \Sigma)^{1-y}$.

* 模型: log-likehood: $\ell(\theta) = \log \prod_{i=1}^{N} p(x_i, y_i)$

$\theta = (\mu_1, \mu_2, \Sigma, \phi)$. $\quad = \sum_{i=1}^{N} \log(p(x_i|y_i) \cdot p(y_i))$

$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \ \ell(\theta)$ $\quad = \sum_{i=1}^{N} (\log p(x_i|y_i) + \log p(y_i))$

$= \sum_{i=1}^{N} [\log (N(\mu_1, \Sigma)^{y_i} N(\mu_2, \Sigma)^{1-y_i}) + y_i \log \phi + (1-y_i) \log (1-\phi)]$

$= \sum_{i=1}^{N} [y_i \log (N(\mu_1, \Sigma)) + (1-y_i) \log (N(\mu_2, \Sigma)) + y_i \log \phi + (1-y_i) \log (1-\phi)]$

求 ϕ: $\frac{\partial \ell(\theta)}{\partial \phi} = \sum_{i=1}^{N} \frac{y_i}{\phi} - \frac{1-y_i}{1-\phi} = 0$

$\Rightarrow \sum_{i=1}^{N} y_i (1-\phi) - \phi (1-y_i) = 0 \Rightarrow \sum_{i=1}^{N} y_i - \phi = 0 \Rightarrow \phi = \frac{1}{N} \sum_{i=1}^{N} y_i = \frac{N_1}{N}$ ($y_i=1$ 频率)

求 μ_1: 与μ_1相关的项只有 $\sum_{i=1}^{N} y_i \log (N(\mu_1, \Sigma))$

$= \sum_{i=1}^{N} y_i \log \frac{1}{(2\pi)^{\frac{p}{2}} |\Sigma|^{\frac{1}{2}}} \exp\{-\frac{1}{2}(x_i - \mu_1)^T \Sigma^{-1} (x_i - \mu_1)\}$

$\propto \sum_{i=1}^{N} -\frac{1}{2} y_i (x_i - \mu_1)^T \Sigma^{-1} (x_i - \mu_1)$ (去掉不相关常数项).

$= -\frac{1}{2} \sum_{i=1}^{N} y_i (\underbrace{x_i^T \Sigma^{-1} x_i}_{常数} + \mu_1^T \Sigma^{-1} \mu_1 - 2\mu_1^T \Sigma^{-1} x_i)$

注意到括号中是关于μ_1的二次项. 因此 $\hat{\mu}_1 = \frac{\sum_{i=1}^{N} y_i x_i}{\sum_{i=1}^{N} y_i} = \frac{\sum_{i=1}^{N} y_i x_i}{N_1}$

求Σ， 与Σ相关的项只有 $\sum_{i=1}^{N} y_i \log(\mathcal{N}(\mu_1,\Sigma)) + \sum_{i=1}^{N}(1-y_i)\log(\mathcal{N}(\mu_2,\Sigma))$

而 $\sum_{i=1}^{N} \log(\mathcal{N}(\mu,\Sigma)) = \sum_{i=1}^{N} \log (2\pi)^{-\frac{p}{2}}|\Sigma|^{-\frac{1}{2}}\exp\{-\frac{1}{2}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)\}$.

$$= \sum_{i=1}^{N} -\frac{1}{2}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu) - \frac{p}{2}\log 2\pi - \frac{1}{2}\log|\Sigma|$$

$$= C - \frac{N}{2}\log|\Sigma| - \frac{1}{2}\sum_{i=1}^{N}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)$$

$$= C - \frac{N}{2}\log|\Sigma| - \frac{1}{2}tr(\sum_{i=1}^{N}(x_i-\mu)(x_i-\mu)^T\Sigma^{-1})$$

$$= C - \frac{N}{2}\log|\Sigma| - \frac{N}{2}tr(S\Sigma^{-1}). \quad [S = \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)(x_i-\mu)^T]$$

故 $\sum_{i=1}^{N} y_i \log(\mathcal{N}(\mu_1,\Sigma)) + \sum_{i=1}^{N}(1-y_i)\log(\mathcal{N}(\mu_2,\Sigma))$

$$= -\frac{N_1}{2}\log|\Sigma| - \frac{N_1}{2}tr(S_1\Sigma^{-1}) - \frac{N_2}{2}\log|\Sigma| - \frac{N_2}{2}tr(S_2\Sigma^{-1}) + C$$

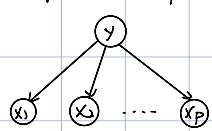
$$= -\frac{N}{2}\log|\Sigma| - \frac{N_1}{2}tr(S_1\Sigma^{-1}) - \frac{N_2}{2}tr(S_2\Sigma^{-1}) + C$$

$$\Rightarrow \frac{\partial L(\theta)}{\partial \Sigma} = -\frac{1}{2}\left(\frac{N}{|\Sigma|}\cdot|\Sigma|\cdot\Sigma^{-1} - N_1S_1\Sigma^{-2} - N_2S_2\Sigma^{-2}\right) = 0 \qquad \left[\frac{\partial tr(AB)}{\partial A} = B^{-1},\ \frac{\partial|A|}{\partial A} = |A|\cdot A^{-1}\right]$$

$$\Rightarrow \quad N\Sigma + N_1S_1 + N_2S_2 = 0 \quad \Rightarrow \hat{\Sigma} = \frac{N_1S_1+N_2S_2}{N}$$

6. Naive Bayes Classifer.

* Naive Bayes Assumption (条件独立性假设).



$x_i \perp x_j \mid Y$ (概率图角度).

$$\Rightarrow p(x|y) = \prod_{j=1}^{p} P(x_j|y).$$

* 模型： $\hat{y} = \arg\max_y P(y|x) = \arg\max_y p(x|y)\cdot p(y). = \arg\max_y p(y)\prod_{j=1}^{p} P(x_i|y).$ 类别分布.

* 参数学习： MLE