P13 Background

- Frequentists : Statistical Machine Learning $\xrightarrow{e.g.}$ Linear Regression ① Linearity ② Global ③ Data isn't processed (↗ The results are output directly)
- Bayesians : Probabilistic Graphical Model

Create a new model by changing ①②③

(1.1) Attribute Nonlinearity : Feature Conversion (Polynominal Regression e.g. $w_i x_i^2$) (1.2) Global Nonlinearity : Linear Classification (The results are output to a nonlinear activation function)

(1.3) Coefficient nonlinearity : Neural Networks (Coefficient are not fixed) ⇒ ∵ w initial is different, the result may be different

② Segmenting the data : Regression Splines (样条回归: 数据分段拟合) , Decision Tree

③ Use after data processing : PCA

Linear Regression $\xrightarrow[\text{Dimensionality Reduction}]{\text{Activation Functions}}$ Linear Classfication ⇒ $\begin{cases} y = f(w^T x + b) & y \in \begin{cases}\{0,1\} \\ [0,1]\end{cases} \\ f : w^T x + b \to y & f^{-1} : y \to w^T x + b \end{cases}$

Drop to one dimesion, set the threshold. Greater than the threshold is 1, less than the threshold is 0.

P14 Preceptron

(Only binary classification is passible for one preceptron)

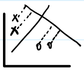
↳ Error Driver D: {Misclassified samples} Model : $y = \text{sign}(w^T x)$ $\text{sign}(a) = \begin{cases} 1 & a \geq 0 \\ -1 & a < 0 \end{cases}$

Loss Function : ① $L(w) = \sum_{i=1}^{N} I\{y_i w^T x_i < 0\}$ $I(a) = \begin{cases} 1 & a = \text{True} \\ 0 & a = \text{False} \end{cases}$ $y_i w^T x_i < 0$ indicates that the sample was misclassified

↳ Not Dirivable ② $L(w) = \sum_{x_i \in D} -y_i w^T x_i$

P15 Linear Discriminat Analysis (LDA) (Muticlass Classification)

$x_{c_1} = \{x_i \mid y_i = +1\}$, $x_{c_2} = \{x_i \mid y_i = -1\}$



Coordinate after projection : $z = w^T x_i$ $\bar{z} = \frac{1}{N}\sum_{i=1}^{N} w^T x_i$ $S_z = \frac{1}{N}\sum_{i=1}^{N}(z_i - \bar{z})(z_i - \bar{z})^T = \frac{1}{N}\sum_{i=1}^{N}(w^T x_i - \bar{z})(w^T x_i - \bar{z})^T$

Let category C_1 C_2 Algorithm idea : Small sample distance within the same categroy, large sample distance between different categories

Within the same categroy : $S_1 + S_2$ Between different categroies : $(\bar{z}_1 - \bar{z}_2)^2$ Loss Fuction : $J(w) = \frac{(\bar{z}_1 - \bar{z}_2)^2}{S_1 + S_2}$ $\hat{w} = \arg\max_w J(w)$

$$(\bar{z}_1 - \bar{z}_2)^2 = \left[w^T\left(\frac{1}{N_1}\sum_{i=1}^{N_1} x_i - \frac{1}{N_2}\sum_{j=1}^{N} x_j\right)\right]^2 = w^T(\bar{x}_{c_1} - \bar{x}_{c_2})(\bar{x}_{c_1} - \bar{x}_{c_2})^T w$$

$$S_1 = \frac{1}{N_1}\sum_{i=1}^{N_1}\left(w^T x_i - \frac{1}{N_1}\sum_{j=1}^{N_1} w^T x_j\right)\left(w^T x_i - \frac{1}{N_1}\sum_{j=1}^{N_1} w^T x_j\right)^T = w^T \sum_{i=1}^{N_1} \frac{1}{N_1}(x_i - \bar{x}_{c_1})(x_i - \bar{x}_{c_1})^T w = w^T S_{c_1} w$$

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

P16 LDA Model Solution

$J(w) = \frac{(\bar{z}_1 - \bar{z}_2)^2}{S_1 + S_2} = \frac{w^T(\bar{x}_{c_1} - \bar{x}_{c_2})(\bar{x}_{c_1} - \bar{x}_{c_2})^T w}{w^T(S_{c_1} + S_{c_2})w}$ 设 $\underset{p\times p}{S_b} = (\bar{x}_{c_1} - \bar{x}_{c_2})(\bar{x}_{c_1} - \bar{x}_{c_2})^T$ between-class variance

$\underset{p\times p}{S_w} = S_{c_1} + S_{c_2}$ with-class variance

$J(w) = \frac{w^T S_b w}{w^T S_w w}$ $\frac{\partial(J(w))}{\partial w} = \frac{S_b w w^T S_w w - w^T S_b w S_w w}{(w^T S_w w)(w^T S_w w)^T} = 0$ ∴ $S_b w \underbrace{w^T S_w w}_{1\times p \times p \times p \times p \times 1 \in \mathbb{R}} = \underbrace{w^T S_b w}_{} S_w w$

Focus mainly on the direction of w, not the size

$w = \frac{w^T S_w w}{w^T S_b w} S_w^{-1} S_b \cdot w = \frac{w^T S_w w}{w^T S_b w} S_w^{-1}(\bar{x}_{c_1} - \bar{x}_{c_2})\underbrace{(\bar{x}_{c_1} - \bar{x}_{c_2})^T w}_{1\times p \times p \times 1 \in \mathbb{R}}$

$w \propto S_w^{-1}(\bar{x}_{c_1} - \bar{x}_{c_2})$ if S_w^{-1} is diagonal matrix, isotropic (各向同性) $S_w^{-1} \propto I \Rightarrow w \propto (\bar{x}_{c_1} - \bar{x}_{c_2})$

Multiclass LDA : $S_w = \sum_{i=1}^{C} S_i$, $S_b = \sum_{i=1}^{C} N_i(\bar{x}_i - \bar{x})(\bar{x}_i - \bar{x})^T$

$S_w^{-1} S_b w = \frac{w^T S_b w}{w^T S_w w} w$, ∴ w is eigenvetor of $S_w^{-1} S_b$

P17 Logistic Regression

Discriminative models find $P(y|x)$ directly ⇒ $\hat{y} = \arg\max_{y \in \{0,1\}} P(y|x)$ which can maximum $P(y|x)$

sigmoid function: $\sigma(z) = \frac{1}{1+e^{-z}}$ ⇒ $\ln\frac{\sigma(z)}{1-\sigma(z)} = z$, $\frac{\sigma(z)}{1-\sigma(z)}$ is odds (几率)

Let $P(y=1|x) = \sigma(z)$ $\begin{cases} P_1 = P(y=1|x) = \sigma(w^T x) = \frac{1}{1+e^{-w^T x}} , & y = 1 \\ P_0 = P(y=0|x) = 1 - \sigma(w^T x) = \frac{e^{-w^T x}}{1+e^{-w^T x}} , & y = 0 \end{cases}$ ⟶ $P(y|x) = P_1^{y} P_0^{(1-y)}$

MLE : $\hat{w} = \arg\max_w P(Y|X) = \arg\max_w \log\prod_{i=1}^{N} P(y_i|x_i) = \arg\max_w \sum_{i=1}^{N} \log P(y_i|x_i)$

$= \arg\max_w \sum_{i=1}^{N}\underbrace{[y_i(\log P_1) + (1-y)(\log P_0)]}_{}$ → $-$Cross Entropy $\underset{\text{MLE}}{\max}$ ⇒ $\underset{\text{Loss Function}}{\min}$ (min Cross Entropy)

P18 Gaussian Discriminant Analysis

Generative models ⇒ Compare $P(y=0|x)$ and $P(y=1|x)$ by $p(x|y)p(y)$ without finding $P(y|x)$ ⇒ $\hat{y} = \arg\max_{y\in\{0,1\}} P(x|y)P(y)$

$y \sim \text{Bernoulli}(\phi)$ ⇒ $y = \begin{cases} \phi & , y = 1 \\ 1-\phi & , y = 0 \end{cases}$ ⇒合并⇒ $y = \phi^{y}(1-\phi)^{(1-y)}$

Let $\left.\begin{array}{l} x|y=1 \sim N(\mu_1, \Sigma) \\ x|y=0 \sim N(\mu_2, \Sigma) \end{array}\right\}$ ⇒ $P(x|y) = N(\mu_1, \Sigma)^{y} \cdot N(\mu_2, \Sigma)^{(1-y)}$

log - likelihood : $l(\theta) = \log\prod_{i=1}^{N} P(y_i|x_i) = \sum_{i=1}^{N}[\log P(x_i|y_i) + \log P(y_i)]$

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

$\theta = (\mu_1, \mu_2, \Sigma, \phi)$ $\hat{\theta} = \arg\max_\theta l(\theta)$

P19 GDA Model Solution

① Find ϕ : $\frac{\partial l(\theta)}{\partial \phi} = \frac{\partial}{\partial \phi}\sum_{i=1}^{N}[y_i \log\phi + (1-y_i)\log(1-\phi)] = \sum_{i=1}^{N}\left[y_i \frac{1}{\phi} - (1-y_i)\frac{1}{1-\phi}\right] = 0$

∴ $\sum_{i=1}^{N}[y_i(1-\phi) - (1-y_i)\phi] = 0$ ⇒ $\sum_{i=1}^{N}(y_i - \phi) = 0$ ⇒ $\phi = \frac{1}{N}\sum_{i=1}^{N} y_i$

∵ $y = 1 : N_1$ $y = 0 : N_2$ $N_1 + N_2 = N$ ∴ $\phi = \frac{N_1}{N}$

② Find μ_1 : $\sum_{i=1}^{N} \log N(\mu_1, \Sigma)^{y_i} = \sum_{i=1}^{N} y_i \log\frac{1}{(2\pi)^{\frac{p}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(x_i - \mu_1)^T \Sigma^{-1}(x_i - \mu_1)\right\}$

$\therefore \sum_{i=1}^{N}[\ldots] = 0 \quad \sum_{i=1}^{N}(y_i - \phi) = 0 \Rightarrow \phi = \frac{1}{N}\sum_{i=1}^{N} y_i$

$\because y=1: N_1 \quad y=0: N_2 \quad N_1+N_2=N \quad \therefore \phi = \frac{N_1}{N}$

② Find μ_1: $\sum_{i=1}^{N} \log N(\mu_1, \Sigma)^{y_i} = \sum_{i=1}^{N} y_i \log \frac{1}{(2\pi)^{\frac{p}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(x_i-\mu_1)^T \Sigma^{-1}(x_i-\mu_1)\right\}$

$\hat{\mu}_1 = \arg\max_{\mu_1} \sum_{i=1}^{N} y_i\left[-\frac{1}{2}(x_i-\mu_1)^T\Sigma^{-1}(x_i-\mu_1)\right] \quad \therefore \frac{\partial f(\mu_1)}{\partial \mu_1} = 0$

$\therefore \sum_{i=1}^{N} y_i \Sigma^{-1}(x_i-\mu_1) = 0 \Rightarrow \sum_{i=1}^{N} y_i \mu_1 = \sum_{i=1}^{N} y_i x_i \Rightarrow \mu_1 = \frac{\sum_{i=1}^{N} y_i x_i}{\sum_{i=1}^{N} y_i} = \frac{\sum_{i=1}^{N} y_i x_i}{N_1}$

P20 GDA Model Solution ②

Find Σ: $\sum_{i=1}^{N} y_i \log N(\mu_1, \Sigma) = \sum_{x_i \in C_1} \log \frac{1}{(2\pi)^{\frac{p}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(x_i-\mu_1)^T\Sigma^{-1}(x_i-\mu_1)\right]$

$= \sum_{x_i \in C_1}\left[-\log(2\pi)^{\frac{p}{2}} - \frac{1}{2}\log|\Sigma| - \frac{1}{2}(x_i-\mu_1)^T\Sigma^{-1}(x_i-\mu_1)\right]$

$= -\frac{1}{2}N_1\log|\Sigma| - \frac{1}{2}\sum_{i=1}^{N_1}(x_i-\mu_1)^T\underbrace{\Sigma^{-1}(x_i-\mu_1)}_{} + c \qquad 1\times p \times p \times p \times p \times 1 \quad tr(c) = c$ 矩阵的迹

Formula: $\frac{\partial tr(AB)}{\partial A} = B^T$; $\frac{\partial |A|}{\partial A} = |A|A^{-1}$; $tr(ABC) = tr(BCA)$

$\therefore$ Original formular $= -\frac{1}{2}N_1\log|\Sigma| - \frac{1}{2}tr\left[\sum_{i=1}^{N_1}(x_i-\mu_1)(x_i-\mu_1)^T\Sigma^{-1}\right] + c$

$\because \sum_{i=1}^{N_1}(x_i-\mu_1)(x_i-\mu_1)^T = N_1S_1$ $\therefore$ Original formular $= -\frac{1}{2}N_1\log|\Sigma| - \frac{1}{2}N_1 tr(S_1\Sigma^{-1}) + c$

$\therefore \hat{\Sigma} = \arg\max_{\Sigma} -\frac{1}{2}(N_1+N_2)\log|\Sigma| - \frac{1}{2}N_1 tr(S_1\Sigma^{-1}) - \frac{1}{2}N_2 tr(S_2\Sigma^{-1}) + c$

$\frac{\partial f(\Sigma)}{\partial \Sigma} = 0 \quad \because \frac{\partial(\log|\Sigma|)}{\partial \Sigma} = \frac{|\Sigma|\Sigma^{-1}}{|\Sigma|} \quad \frac{\partial tr(S_1\Sigma^{-1})}{\partial \Sigma} = S^T(-1)\Sigma^{-2}$

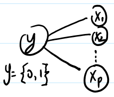
$\therefore -\frac{1}{2}N\Sigma^{-1} + \frac{1}{2}N_1S_1\Sigma^{-2} + \frac{1}{2}N_2S_2\Sigma^{-2} = 0 \Rightarrow N\Sigma - N_1S_1 - N_2S_2 = 0$

$\therefore \hat{\Sigma} = \frac{1}{N}(N_1S_1 + N_2S_2)$

P21 Naive Bayes Classifier

Alogrithm idea: Naive Bayesian Hypothesis $\Rightarrow$ Conditional Independence Hypothesis

$\Rightarrow$ The simplest probality graphcial model (directed graph)



$y=\{0,1\}$

$x_i \perp x_j \mid y \ (i \neq j) \qquad x \in \mathbb{R}^p$

$P(x|y) = \prod_{i=1}^{p} P(x_i|y)$

	Single Experiment	Multiple Experiments
Binary classification	Bernoulli (伯努力分布)	Binormial (二项式分布)
Multiclassification	Categorical (分类分布)	Mutinormial (多项式分布)

$$P(y=\Delta|x) = \frac{P(x|y=\Delta)P(y=\Delta)}{P(x)} \propto P(x|y=\Delta)P(y=\Delta)$$

The purpose is to compare the size of $P(y|x)$ of different categories and make a classification.

Without caring about the value of $P(y|x)$