Linear Classification

1 Linear Classification

1.1 Logistic Regression

$$
\begin{aligned}
& y=\sigma(w^T x), \text{ where } \sigma=\frac{1}{1+\exp(-w^T x)} \\
\because\ & \begin{cases} p_1=p(y=1|x)=\frac{1}{1+\exp(-w^T x)} \\ p_0=1-p(y=1|x)=\frac{\exp(-w^T x)}{1+\exp(-w^T x)} \end{cases} \\
\therefore\ & p(y|x)=p_0^{1-y} p_1^{y} \\
\therefore\ & P(Y|X)=\prod_{i=1}^{N} p(y_i|x_i) \\
\therefore\ & \text{MLE: } \arg\max_w \log P(Y|X) \\
& =\arg\max_w \sum_{i=1}^{N} \log p(y_i|x_i) \\
& =\arg\max_w \sum_{i=1}^{N} \log \left[p_0^{1-y_i} p_1^{y_i}\right] \\
& =\arg\max_w \sum_{i=1}^{N} \left[(1-y_i)\log p_0+y_i \log p_1\right] \\
& =\arg\min_w (\text{Cross Entropy})
\end{aligned} \tag{1}
$$

1.2 Perceptron

Model:

$$
\begin{aligned}
f(x)&=\text{sign}(w^T x+b) \\
\text{sign}&=\begin{cases} +1, a\geq 0 \\ -1, a\leq 0 \end{cases}
\end{aligned} \tag{2}
$$

Error driven model, therefore the loss function is,

$$
L(w,b)=-\sum_{x_i\in M} y_i(w^T x_i+b) \tag{3}
$$

where M is the set of error samples. Therefore,

$$
\begin{aligned}
\frac{\partial L}{\partial w}&=-\sum_{x_i\in M} y_i x_i \\
\frac{\partial L}{\partial b}&=-\sum_{x_i\in M} y_i \\
\therefore\ w_{i+1}&=w_i-\eta\frac{\partial L}{\partial w} \\
b_{i+1}&=b_i-\eta\frac{\partial L}{\partial b}
\end{aligned} \tag{4}
$$

1.3 Linear Discriminant Analysis (LDA or Fisher)

The idea of LDA is to find a direction, which can meet two requirements:

- the distances between samples of same categories are minimal;
- the distances between samples of different categories are maximal.

Assuming w is the best direction, so the projection of samples x in direction w is,

$$z = w^T x = |w| \cdot |x| \cos\theta \tag{5}$$

Given the training set (X, Y), where

$$\begin{aligned} X = (x_1, x_2, \cdots, x_n)^T = \begin{pmatrix} x_1^T, \\ x_2^T, \\ \vdots \\ x_n^T, \end{pmatrix}_{n \times p}, x_i \in \mathcal{R}^p, Y = \begin{pmatrix} y_1, \\ y_2, \\ \vdots \\ y_n, \end{pmatrix}_{n \times 1}, y_i \in \{0, 1\} \\ X_{C1} = \{x_i | y_i = 1\}, X_{C2} = \{x_i | y_i = -1\}, |X_{C1}| = N1, |X_{C2}| = N2, N1 + N2 = N \end{aligned} \tag{6}$$

The categories in the training set are $C1, C2$, and the number of samples in each category are $N1, N2$. Therefore, the distances between samples in the different categories are,

$$\begin{aligned} &(\frac{1}{N_1}\sum_{i=1}^{N1} z_i - \frac{1}{N_2}\sum_{i=1}^{N2} z_i)^2 \\ =&(\frac{1}{N_1}\sum_{i=1}^{N1} w^T x_i - \frac{1}{N_2}\sum_{i=1}^{N2} w^T x_i)^2 \\ =&\left[w^T(\frac{1}{N_1}\sum_{i=1}^{N1} x_i - \frac{1}{N_2}\sum_{i=1}^{N2} x_i)\right]^2 \\ =&w^T(\overline{X_{C1}} - \overline{X_{C2}})(\overline{X_{C1}} - \overline{X_{C2}})^T w \end{aligned} \tag{7}$$

The distances between samples in the same category are represented using variance values.

$$\begin{aligned} \text{var}(C1) = S1 &= \frac{1}{N1}\sum_{i=0}^{N1}(z_i - \overline{z_{C1}})(z_i - \overline{z_{C1}})^T \\ &= \frac{1}{N1}\sum_{i=0}^{N1}(w^T x_i - \overline{z_{C1}})(w^T x_i - \overline{z_{C1}})^T \\ &= \frac{1}{N1}\sum_{i=0}^{N1}(w^T x_i - \frac{1}{N1}\sum_{i=0}^{N1} w^T x_i)(w^T x_i - \frac{1}{N1}\sum_{i=0}^{N1} w^T x_i)^T \\ &= w^T \underbrace{\frac{1}{N1}\sum_{i=0}^{N1}(x_i - \frac{1}{N1}\sum_{i=0}^{N1} x_i)(x_i - \frac{1}{N1}\sum_{i=0}^{N1} x_i)^T}_{S_{C1}} w \\ &= w^T S_{C1} w \\ \text{var}(C1) = S2 &= w^T S_{C2} w \\ \therefore\ S1 + S2 = w^T (S_{C1} &+ S_{C2}) w \end{aligned} \tag{8}$$

Therefore, the loss function is,

$$
\begin{aligned}
L(w) &= \frac{w^T(\overline{X_{C1}} - \overline{X_{C2}})(\overline{X_{C1}} - \overline{X_{C2}})^T w}{w^T(S_{C1} + S_{C2})w} \\
&= (w^T(\overline{X_{C1}} - \overline{X_{C2}})(\overline{X_{C1}} - \overline{X_{C2}})^T w)(w^T(S_{C1} + S_{C2})w)^{-1} \\
&\therefore \text{ unconstrained optimization problem, } \frac{\partial L}{\partial w} = 0 \\
\therefore \ \frac{\partial L}{\partial w} &= 2(\overline{X_{C1}} - \overline{X_{C2}})(\overline{X_{C1}} - \overline{X_{C2}})w(w^T(S_{C1} + S_{C2})w)^{-1} - \\
&\quad (w^T(\overline{X_{C1}} - \overline{X_{C2}})(\overline{X_{C1}} - \overline{X_{C2}})^T w)(w^T(w^T(S_{C1} + S_{C2})w)^{-2}(2(S_{C1} + S_{C2})w) = 0 \\
&\text{Let } (\overline{X_{C1}} - \overline{X_{C2}})(\overline{X_{C1}} - \overline{X_{C2}})^T = S_a, \ \ (S_{C1} + S_{C2}) = S_b \\
\therefore \ \frac{\partial L}{\partial w} &= 2S_a w(w^T S_b w)^{-1} - (w^T S_a w)(w^T S_b w)^{-2}(2S_b w) = 0 \\
&\quad S_a w(w^T S_b w) - (w^T S_a w)(S_b w) = 0 \\
&\because \quad w \in \mathcal{R}^{P\times 1} \to w^T \in \mathcal{R}^{1\times P}, \ S_a \in \mathcal{R}^{P\times P} \\
&\therefore \quad w^T S_a w \in \mathcal{R} \\
&\because \quad \overline{X_{C1}}, \overline{X_{C2}} \in \mathcal{R}^{1\times P} \\
&\therefore \quad (\overline{X_{C1}} - \overline{X_{C2}})(\overline{X_{C1}} - \overline{X_{C2}})^T \in \mathcal{R}^{P\times P} \\
&\therefore \quad (\overline{X_{C1}} - \overline{X_{C2}})(\overline{X_{C1}} - \overline{X_{C2}})^T w \in \mathcal{R} \\
&\therefore \quad S_a w(w^T S_b w) = (w^T S_a w)(S_b w) \\
&\therefore \quad w = \frac{w^T S_b w}{w^T S_a w} S_b^{-1} S_a w \\
&\therefore \quad w \propto S_b^{-1} S_a w = S_b^{-1}(\overline{X_{C1}} - \overline{X_{C2}})(\overline{X_{C1}} - \overline{X_{C2}})^T w \\
&\because \quad (\overline{X_{C1}} - \overline{X_{C2}})^T \in \mathcal{R}^{1\times P}, \ w \in \mathcal{R}^{P\times 1} \\
&\therefore \quad (\overline{X_{C1}} - \overline{X_{C2}})^T w \in \mathcal{R} \\
&\therefore \quad w \propto S_b^{-1}(\overline{X_{C1}} - \overline{X_{C2}})
\end{aligned} \tag{9}
$$

1.4 Gaussian Discriminant Analysis (GDA)

Given

$$
\begin{aligned}
&\text{Data: } \{(x_i, y_i)\}_{i=1}^n, x \in \mathcal{R}^p, y_i \in \{0, 1\} \\
&y \sim \text{Bernoulli}(\phi), y = \begin{cases} \phi, & \text{if } y = 1, |\{x|y = 1\}| = n_1 \\ 1 - \phi, & \text{if } y = 0, |\{x|y = 0\}| = n_0 \\ n_0 + n_1 = n \end{cases} \\
&\therefore \ y = \phi^y (1 - \phi)^{1-y} \\
&P(x|y = 1) \sim \mathcal{N}(\mu_1, \Sigma) \\
&P(x|y = 0) \sim \mathcal{N}(\mu_2, \Sigma) \\
&\therefore \ P(x|y) = \mathcal{N}(\mu_1, \Sigma)^y \mathcal{N}(\mu_2, \Sigma)^{1-y}
\end{aligned} \tag{10}
$$

$P(x|y=1)$ is the identical independent distribution, therefore the log likelihood is

$$
\begin{aligned}
L(\theta) = \log P(X,Y) &= \log \prod_{i=1}^{n} p(x_i, y_i) \\
&= \sum_{i=1}^{n} \log p(x_i, y_i) \\
&= \sum_{i=1}^{n} [\log p(x_i|y_i) + \log p(y_i)] \\
&= \sum_{i=1}^{n} [y_i \log \mathcal{N}(\mu_1, \Sigma) + (1-y_i) \log \mathcal{N}(\mu_2, \Sigma) + \log p(y_i)] \\
&= \underbrace{\sum_{i=1}^{n} y_i \log \mathcal{N}(\mu_1, \Sigma)}_{a} + \underbrace{\sum_{i=1}^{n}(1-y_i) \log \mathcal{N}(\mu_2, \Sigma)}_{b} + \underbrace{\sum_{i=1}^{n} \log p(y_i)}_{c} \\
\therefore \quad & \hat{\theta} = \arg\max_{\theta} L(\theta),\ \theta \in \{\mu_1, \mu_2, \Sigma, \phi\}
\end{aligned}
\tag{11}
$$

Therefore, let $\frac{\partial L(\theta)}{\partial \theta} = 0$ is subjective to let a, b, c are 0 independently.

1.4.1 Compute $\hat{\phi}$

$$
\begin{aligned}
& \frac{\partial L(\mu_1, \mu_2, \Sigma, \phi)}{\partial \phi} = \frac{\partial c}{\partial \phi} = 0 \\
& \frac{\partial}{\partial \phi} \sum_{i=1}^{n} \log \left[\phi^{y_i}(1-\phi)^{1-y_i}\right] = 0 \\
& \frac{\partial}{\partial \phi} \sum_{i=1}^{n} [y_i \log \phi + (1-y_i) \log(1-\phi)] = 0 \\
& \sum_{i=1}^{n} \left[\frac{y_i}{\phi} - \frac{1-y_i}{1-\phi}\right] = 0 \\
& \sum_{i=1}^{n} [y_i - \phi] = 0 \\
& \therefore\ \hat{\phi} = \frac{1}{N} \sum_{i=1}^{n} y_i
\end{aligned}
\tag{12}
$$

1.4.2 Compute $\hat{\mu_1}$

$$
\begin{aligned}
&\frac{\partial L(\mu_1,\mu_2,\Sigma,\phi)}{\partial \mu_1} = \frac{\partial a}{\partial \mu_1} = 0 \\
&\sum_{i=1}^{n} \frac{\partial}{\partial \mu_1}\left[y_i \log \mathcal{N}(\mu_1,\Sigma)\right] = 0 \\
&\sum_{i=1}^{n} \frac{\partial}{\partial \mu_1} \log\left[y_i \frac{1}{(2\pi)^{\frac{p}{2}}\Sigma^{\frac{1}{2}}} \exp(-(x_i-\mu_1)^T\Sigma^{-1}(x_i-\mu_1))\right] = 0 \\
&\sum_{i=1}^{n} \frac{\partial}{\partial \mu_1}\left[y_1(x_i-\mu_1)^T\Sigma^{-1}(x_i-\mu_1)\right] = 0 \\
&\sum_{i=1}^{n} \frac{\partial}{\partial \mu_1}\left[y_1(x_i^T\Sigma^{-1}-\mu_1^T\Sigma^{-1})(x_i-\mu_1)\right] = 0 \\
&\sum_{i=1}^{n} \frac{\partial}{\partial \mu_1}\left[y_1(x_i^T\Sigma^{-1}x_i - x_i^T\Sigma^{-1}\mu_1 - \mu_1^T\Sigma^{-1}x_i + \mu_1^T\Sigma^{-1}\mu_1)\right] = 0 \\
&\because\ x_i \in \mathcal{R}^{p\times 1}, x_i^T \in \mathcal{R}^{1\times p}, \Sigma^{-1} \in \mathcal{R}^{p\times p}, \mu_1 \in \mathcal{R}^{p\times 1} \\
&\therefore\ x_i^T\Sigma^{-1}\mu_1 \in \mathcal{R},\ \text{and } x_i^T\Sigma^{-1}\mu_1 = \mu_1^T\Sigma^{-1}x_i \\
&\therefore\ \sum_{i=1}^{n} \frac{\partial}{\partial \mu_1}\left[y_1(x_i^T\Sigma^{-1}x_i - 2\mu_1^T\Sigma^{-1}x_i + \mu_1^T\Sigma^{-1}\mu_1)\right] = 0 \\
&\therefore\ \sum_{i=1}^{n} y_i\left[2\Sigma^{-1}\mu_1 - 2x_i\Sigma^{-1}\right] = 0 \\
&\therefore\ \hat{\mu_1} = \frac{\sum_{i=1}^{n} y_i x_i}{\sum_{i=1}^{n} y_i} = \frac{\sum_{i=1}^{n} y_i x_i}{n_1}
\end{aligned}
\tag{13}
$$

1.4.3 Compute $\hat{\Sigma}$

$$
\begin{aligned}
&\frac{\partial L(\mu_1,\mu_2,\Sigma,\phi)}{\partial \Sigma} = \frac{\partial (a+b)}{\partial \Sigma} = 0 \\
&\frac{\partial}{\partial \Sigma}\left[\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)\right] = 0 \\
&\rightarrow \ \frac{\partial}{\partial \Sigma}\left[\sum_{i=1}^{n} \log \mathcal{N}(\mu_1,\Sigma) + \sum_{i=1}^{n}\log \mathcal{N}(\mu_2,\Sigma)\right] = 0 \\
&\therefore \ \sum_{i=1}^{n} \log \mathcal{N}(\mu_1,\Sigma) \\
&= \sum_{i=1}^{n_1} \log\left[\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))\right] \\
&= \sum_{i=1}^{n_1}\left[C - \frac{1}{2}\log|\Sigma| - \frac{1}{2}(x_i-\mu_1)^T\Sigma^{-1}(x_i-\mu_1)\right] \\
&= C - \frac{n_1}{2}\log|\Sigma| - \frac{1}{2}\sum_{i=1}^{n_1}(x_i-\mu_1)^T\Sigma^{-1}(x_i-\mu_1) \\
&\because \ \frac{\partial |A|}{\partial A} = |A|A^{-1}, \frac{\partial A^{-1}}{\partial A} = -A^{-2}, \frac{\partial Au}{\partial x} = A\frac{\partial u}{\partial x}, \text{ where } A \text{ is not a function of } x, u = u(x) \\
&\therefore \ \frac{\partial}{\partial \Sigma}\left[C - \frac{n_1}{2}\log|\Sigma| - \frac{1}{2}\sum_{i=1}^{n_1}(x_i-\mu_1)^T\Sigma^{-1}(x_i-\mu_1)\right] \\
&= \frac{n_1}{|\Sigma|}|\Sigma|\Sigma^{-1} - \sum_{i=1}^{n_1}(x_i-\mu_1)(x_i-\mu_1)^T\Sigma^{-2} \\
&= n_1\Sigma^{-1} - \sum_{i=1}^{n_1}(x_i-\mu_1)(x_i-\mu_1)^T\Sigma^{-2} \\
&\therefore \ \frac{\partial (a+b)}{\partial \Sigma} = n_1\Sigma^{-1} - \sum_{i=1}^{n_1}(x_i-\mu_1)(x_i-\mu_1)^T\Sigma^{-2} + n_2\Sigma^{-1} - \sum_{i=1}^{n_2}(x_i-\mu_2)(x_i-\mu_2)^T\Sigma^{-2} = 0 \\
&(n_1+n_2)\Sigma^{-1} - \sum_{i=1}^{n_1}(x_i-\mu_1)(x_i-\mu_1)^T\Sigma^{-2} - \sum_{i=1}^{n_2}(x_i-\mu_2)(x_i-\mu_2)^T\Sigma^{-2} = 0 \\
&(n_1+n_2)\Sigma - (\sum_{i=1}^{n_1}(x_i-\mu_1)(x_i-\mu_1)^T + \sum_{i=1}^{n_2}(x_i-\mu_2)(x_i-\mu_2)^T) = 0 \\
&\hat{\Sigma} = \frac{n_1S_1 + n_2S_2}{n} \\
&\text{where } S_1 = \frac{1}{n_1}\sum_{i=1}^{n_1}(x_i-\mu_1)(x_i-\mu_1)^T, S_2 = \frac{1}{n_2}\sum_{i=1}^{n_2}(x_i-\mu_2)(x_i-\mu_2)^T
\end{aligned}
\tag{14}
$$