1 Basic Statistic Notions

1.1 Expected Value $E[X] = \mu$

$$E[X] = \begin{cases} \sum x_i p_i & \text{if discrete} \\ \int_{\mathbb{R}} x f(x) dx & \text{if continuous} \end{cases}$$

1.2 Variance $var[X] = \sigma_X^2$

$$var[X] = \begin{cases} \sum (x_i - \mu)^2 p_i & \text{if discrete} \\ \int_{\mathbb{R}} (x - \mu)^2 f(x) dx & \text{if continuous} \end{cases}$$

If X is continuous, $var[X] = E[(X - E[X])^2] = E[X^2] - E[X]^2$.

1.3 Standard Deviation σ_X

$$\sigma_X = \begin{cases} \sqrt{\sum (x_i - \mu)^2 p_i} & \text{if discrete} \\ \sqrt{\int_{\mathbb{R}} (x - \mu)^2 f(x) dx} & \text{if continuous} \end{cases}$$

Note that $\sigma_X = \sqrt{var[X]}$. If X is discrete and each $p_i = \frac{1}{N}$, $\sigma_X = \sqrt{\frac{\sum (x_i - \mu)^2}{N}}$.

1.4 Covariance $cov[X, Y]$

1.4.1 Two Variables

X, Y are random variables with space $\mathbb{R}$.

$$cov[X, Y] = \begin{cases} \sum p_i (x_i - \mu_X)(y_i - \mu_Y) & \text{if discrete} \\ \int_{\mathbb{R} \times \mathbb{R}} (x - \mu_X)(y - \mu_Y) f(x, y) & \text{if continuous} \end{cases}$$

Note that $cov[X, X] = var[X]$ and covariance matrix $\Sigma = \begin{pmatrix} cov[X, X] & cov[X, Y] \\ cov[Y, X] & cov[Y, Y] \end{pmatrix}$

1.4.2 More Variables

$X_1, X_2, \ldots, X_n$ are random variables with space $\mathbb{R}$.
The covariance matrix Σ is

$$\Sigma = \begin{pmatrix} cov[X_1, X_1] & cov[X_1, X_2] & \cdots & cov[X_1, X_n] \\ cov[X_2, X_1] & cov[X_2, X_2] & \cdots & cov[X_2, X_n] \\ \vdots & \vdots & \ddots & \vdots \\ cov[X_n, X_1] & cov[X_n, X_2] & \cdots & cov[X_n, X_n] \end{pmatrix}$$

Or let $X = (X_1, X_2, \ldots, X_n)$, $\Sigma = cov[X, X] = E[(X - E[X])(X - E[X])^T]$.

1.5 Correlation $corr[X, Y] = \rho_{X,Y}$

$corr[X, Y] = \frac{cov[X,Y]}{\sigma_X \sigma_Y} \in [-1, 1]$.

1.6 Assume sample come from one distribution

1.6.1 Gaussian Distribution

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

where x vector, D for dimension, μ mean vector, Σ covariance matrix.
An unknown $f_{\mu,\Sigma}(x)$ samples n points $x_1, \ldots, x_n$. The likelihood function

$$L(\mu, \Sigma) = f_{\mu,\Sigma}(x_1) f_{\mu,\Sigma}(x_2) \cdots f_{\mu,\Sigma}(x_n)$$

Higher value of the likelihood function means higher probability to sample $x_1, \ldots, x_n$. The maximum likelihood estimator (μ^*, Σ^*) is

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

Also, for $l(\mu, \Sigma) = \ln L(\mu, \Sigma)$,

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

Then Gaussian distribution

$$\mu^* = \frac{1}{n} \sum_{1}^{n} x_i, \quad \Sigma^* = \frac{1}{n} \sum_{1}^{n} (x_i - \mu^*)(x_i - \mu^*)^T$$

1.7 Two Classes C_1, C_2

C_1 has $\{x_i\}_1^{70}$, C_2 has $\{x_i\}_{71}^{180}$. The posterior probability of an unknown object x is

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

The terms $P(C_k)$ are easy to get. The terms $P(x|C_k)$ is based on the distribution we assumed. If it's Gaussian,

$$P(x|C_1) = f_{\mu_1^*,\Sigma_1^*}(x), \quad P(x|C_2) = f_{\mu_2^*,\Sigma_2^*}(x)$$

We can modify Σ^* by

$$\Sigma^* = \frac{70}{70+110}\Sigma_1^* + \frac{110}{70+110}\Sigma_2^*$$

And let C_1, C_2 share this Σ^*, the likelihood function becomes

$$L(\mu_1, \mu_2, \Sigma) = f_{\mu_1,\Sigma}(x_1) \cdots f_{\mu_1,\Sigma}(x_{70}) f_{\mu_2,\Sigma}(x_{71}) \cdots f_{\mu_2,\Sigma}(x_{180})$$

$\mu_1^*, \mu_2^*, \Sigma^*$ are maximum likelihood estimators. The new model will be a linear classifier. Explain in the following.

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

$$P(C_1|x) = \frac{1}{1+e^{-z}} = \sigma(z), \quad \text{the sigmoid function.}$$

$P(x|C_k) = f_{\mu_k^*,\Sigma^*}(x)$, $k = 1, 2$. So

$$z = \ln \frac{f_{\mu_1^*,\Sigma^*}(x)}{f_{\mu_2^*,\Sigma^*}(x)} + \ln \frac{P(C_1)}{P(C_2)}$$

Put the details in and compute, we get

$$z = (\mu_1^* - \mu_2^*)^T \Sigma^{*-1} x - \frac{1}{2}(\mu_1^*)^T \Sigma^{*-1} \mu_1^* + \frac{1}{2}(\mu_2^*)^T \Sigma^{*-1} \mu_2^* + \ln \frac{P(C_1)}{P(C_2)}$$

Let $w^T = (\mu_1^* - \mu_2^*)^T \Sigma^{*-1}$, $b = -\frac{1}{2}(\mu_1^*)^T \Sigma^{*-1} \mu_1^* + \frac{1}{2}(\mu_2^*)^T \Sigma^{*-1} \mu_2^* + \ln \frac{P(C_1)}{P(C_2)}$. Then

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