Linear Methods for Classification

Bayse Classifier

- Risk of classification rule g

$$\begin{aligned} E_{X,Y}[L(Y, g(X))] &= E_X E_{X|Y}[L(Y, g(X)|X)] \\ &= E_X \left[\sum_{k=1}^{K} L(k, g(X)) P(Y = k|X) \right] \end{aligned}$$

- X : input values, $Y \in \{1, \cdots, K\}$: output values with qualitiative response
- L : loss function, g : classification rule
- This risk function is minimized if the conditional risk is minimized for each x (= pointwise minimize)
- Bayes classifier is classification rule which minimize the conditional risk pointwisely

$$g(x) = \underset{g \in G}{argmin} \sum L(k, g) P(Y = k|X = x)$$

- which is came from Bayes rule

$$\begin{aligned} P(Y = k|X = x) &= \frac{f_{X,Y}(X = x, y = K)}{f_X(x = X)} \\ &= \frac{P(Y = K) f_X | y(X = x|Y = k)}{f_X(X = x)} \\ &= \frac{P(Y = k) f_k(x)}{\sum f_l(x) \pi_l} \\ \therefore P(Y = k|X = x) &\propto f_x(x) \pi_k \end{aligned}$$

- Example 1 : Linear regression of an indecator matrix
  1. coding $\mathbf{Y} = (Y + 1, \cdots, Y_K)$ with $Y_k = 1$ if $G = k$, else $Y_k = 0$ (= one hot encoding)
  2. Fit linear regression to each of the columns of $\mathbf{Y}$
  3. Classification rule

     compute the fitted output $\hat{f}(x)' = (1.x')\hat{\mathbf{B}}$

     identify the largest component and classify accordingly : $argmax_{x \in G} \hat{f}_k(x)$

     $\rightarrow$ follows baysian rule!!
  4. Since we design $E(Y_k|X)$ and select biggest one, this follows bayesian rule
  5. $\sum_{k \in G} \hat{f}(x) = 1$ for any x but $\hat{f}(x)$ can be negative or greater than 1

Linear Discriminant Analysis

- Assume X given $Y = k$ are distributed as multivarite Gaussian whose pdf is :

$$f_k(x) = |2\pi\Sigma|^{-1/2} exp\left\{-\frac{1}{2}(x-\mu_k)'\Sigma_k(x-\mu_k)\right\}$$

- By bayesian rule, We compare conditional expectation (= posterior probability)

$$P(Y=k)f_k(x) = \pi_k|2\pi\Sigma|^{-1/2} exp\left\{-\frac{1}{2}(x-\mu_k)'\Sigma_k(x-\mu_k)\right\}$$

- put a log function and compare its size (= QDA)

$$\delta_k(x) = log\pi_k - \frac{1}{2}log|\Sigma_k| - \frac{1}{2}(x-\mu_k)'\Sigma_k(x-\mu_k)$$

- Assume that $\Sigma_k = \Sigma$ for every $k \in G$ then the differences become

$$\delta_k(x) = x'\Sigma^{-1}\mu_k - \frac{1}{2}\mu_k'\Sigma^{-1}\mu_k + \log(\pi_k)$$

  parameters are estimated by

  $\hat{\pi} = \frac{N_k}{N}, \hat{\mu_k} = \frac{1}{N_k}\sum_{i:Y_i=k} x_i, \hat{\Sigma} = \dfrac{1}{N-K}\sum_{k=1}^{K}\sum_{i:Y_i=k}(x_i-\mu_k)(x_i-\mu_k)'$

- By ANOVA, total variances can be decomposed

$$\begin{aligned} T &= \sum_{k=1}^{K}\sum_{i:Y_i=k}(x_i-\mu_k)(x_i-\mu_k)' \\ &= \sum_{k=1}^{K}\sum_{i:Y_i=k}(x_i-\bar{x}_k)(x_i-\bar{x}_k)' + \sum_{k=1}^{K} N_k(\bar{x}_k-\bar{x})(\bar{x}_k-\bar{x})' \\ &= W + B \end{aligned}$$

  $\rightarrow$ In LDA, we can think that **Within Covariance is same for all class K!!!** ($W = \Sigma$)

- Although real data doesn't follow Gaussian distribution, LDA and QDA can make good performance.

  reasons are not clear but we can guess

  1. Simple decision boundaris (bias variance tradeoff) but not for the QDA
  2. This model does not estimate probability correctly, but can estimate the order of probability..

- Another model : Regularized discirminant analysis

  Idea :Compromise between LDA, QDA, and $\sigma^2 I$ in Σ

$$\hat{\Sigma}(\gamma) = \gamma\hat{\Sigma} + (1-\gamma)\hat{\sigma^2}I$$

$$\hat{\Sigma}_k(\alpha,\gamma) = \alpha\hat{\Sigma}_k + (1-\alpha)\hat{\Sigma}(\gamma)$$

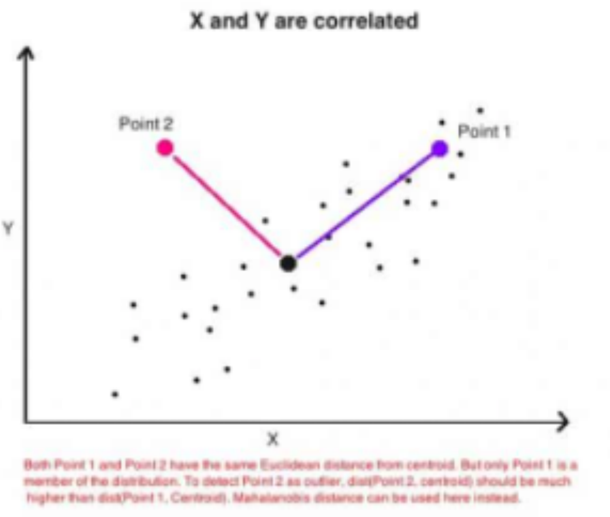
LDA as PC subspaces of the centroids

- In LDA, $\delta_k(x) = \log(\pi_k) - \frac{1}{2}(x-\mu_k)'\Sigma^{-1}(x-\mu_k)$
- Second term is the shape of mahalanobis distance
- Mahalanobis distance

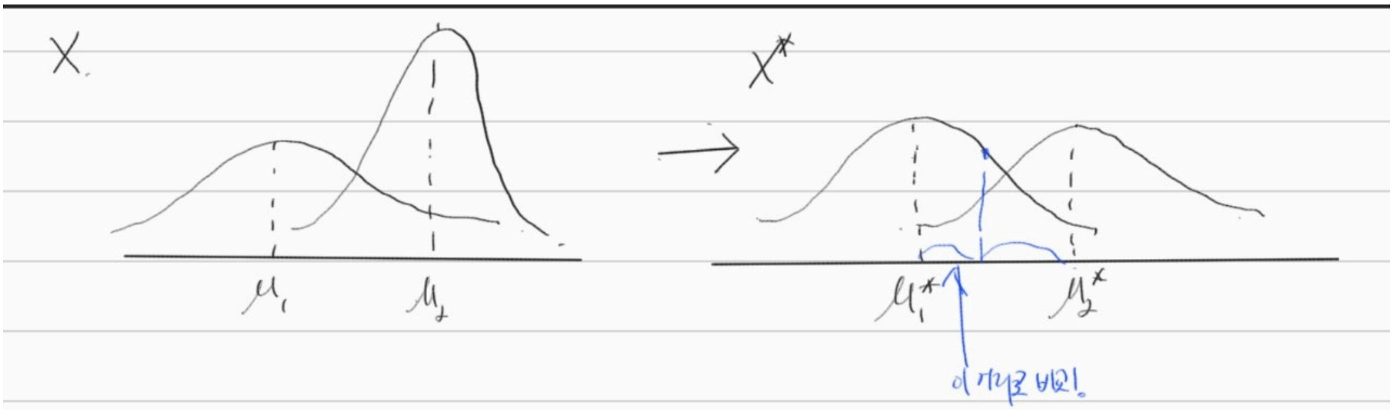
  def : measure of the distance between a point P and a distribution D

  property

  1. distance is zero for P at the mean of D, grows as P moves away from the mean along each PC axis



  2. If each of these axes is re-scaled to have unit variance, then the Mahalanobis distance corresponds to standard Euclidean distance in the transformed space.

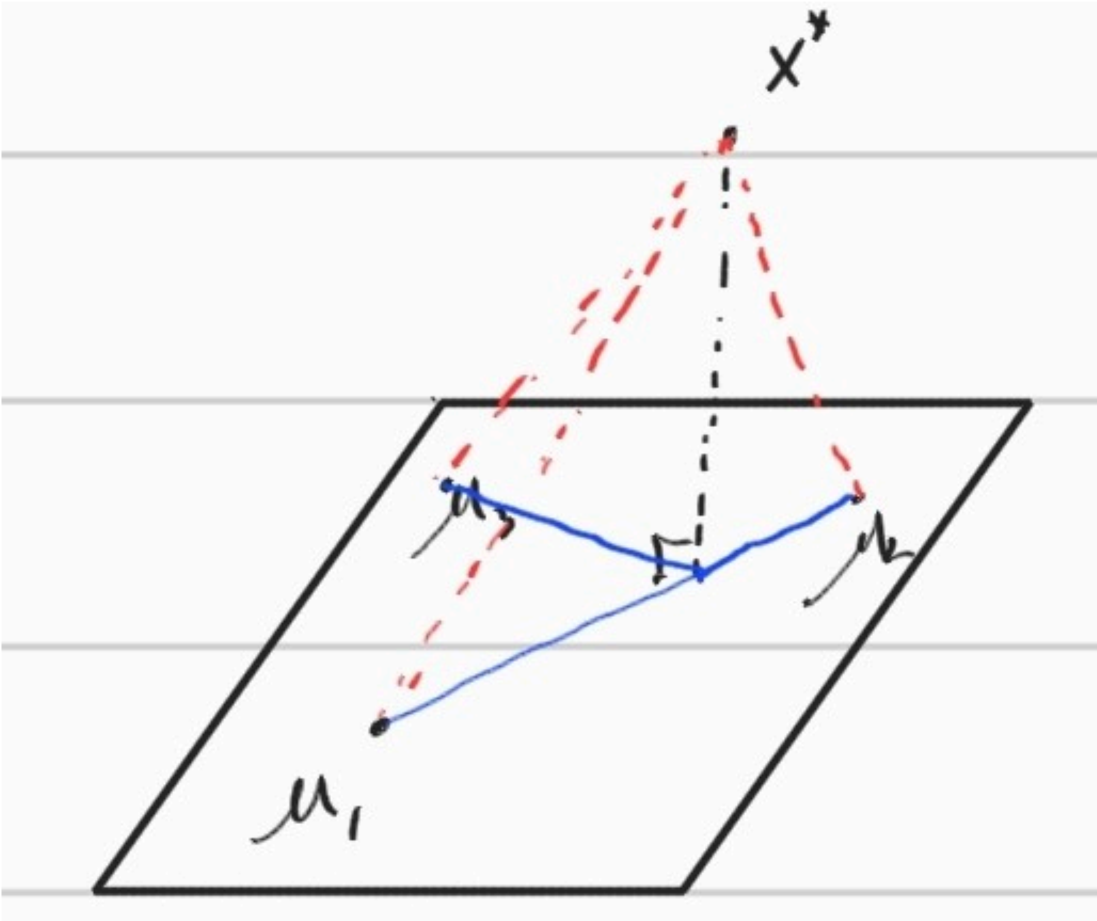


- assume that all π_k is same for simpler inference, then LDA classifier is to compare mahalanobis distance

  $\rightarrow$ distance from centroid of each class is target of comparison

- Transform the space with respect to the common covariance estimate $\hat{\Sigma}$: let $X^* = \hat{\Sigma}^{-1/2}X$
- K centroids in p-dimensional input space lie in an affine substpace of dimension $\le K-1$
- Then, we can project X^* onto this centroid_spanning subspace H_{K-1} and make distance comparison

  $\rightarrow$ regardless of dimension of P, we can compare distances in $K-1$ dimension (Dimension reduction)



- **How to find subspace H_{K-1} ??**
  1. Compute $K \times p$ matrix of centroids M and common covariance matrix W
  2. Compute $M^* = MW^{-1/2}$
  3. Compute B^* which is between-class covariance (= covariance matrix of M^*)
  4. eigen values of B^* is orthogonal basis for H_{K-1}
  5. These basis are
- Ralationship between B and B^*

$$
\begin{aligned}
B &= (M - \frac{1}{k}JM)'(M - \frac{1}{k}JM) \\
&= ((I - \frac{1}{k}J)M)'((I - \frac{1}{k}J)M) \\
&= M'(I - \frac{1}{k}J)M \\
B^* &= M^{*\prime}(I - \frac{1}{k})M^* \\
&= W^{-1/2}M'(I - \frac{1}{k})MW^{-1/2} \\
&= W^{-1/2}BW^{-1/2}
\end{aligned}
$$

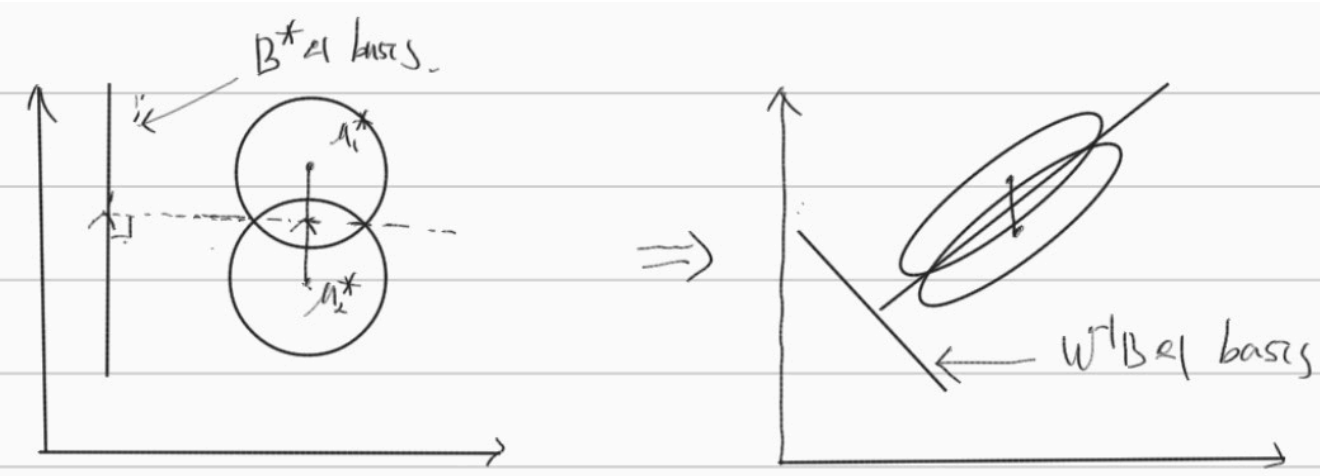
- Eigenvalues of B and B^*

$$
\begin{aligned}
W^{-1/2}BW^{-1/2}v^* &= \lambda v^* \\
W^{-1/2}W^{-1/2}BW^{-1/2}v^* &= \lambda W^{-1/2}v^* \\
W^{-1}Bv &= \lambda v
\end{aligned}
$$

eigen vector of B^* = $W^{-1/2}$ * eigen vector of $W^{-1}B$ ($v_j^* = W^{-1/2}v_j$)

$\rightarrow$ by this property, We can find basis of B^* by calculating eigen values of $W^{-1}B$

$\rightarrow$ which means the best classification axis in original space is v can be found as eigenvector of $W^{-1/2}B$



In general, $W^{-1}B$ is not symmetric. So, it can not be said to eigenvetors of it is orthogonal

However, their eigenvalues are orthogonal because

$$v_j^*\prime v_i^* = 0$$
$$v_j W^{-1} v_i = 0$$

-
- Eigen vector of $W^{-1}B$ is same as $T^{-1}B$

  By generallized eigenvalue problem

$$
\begin{aligned}
&|\ \ Bv = \lambda W v \\
+&|\ \ \lambda Bv = \lambda Bv \\
&--------\\
&(1+\lambda)Bv = \lambda(B+W)v \\
&\therefore Bv = \frac{\lambda}{1+\lambda}Tv
\end{aligned}
$$

  Use this property when we link LDA to CCA

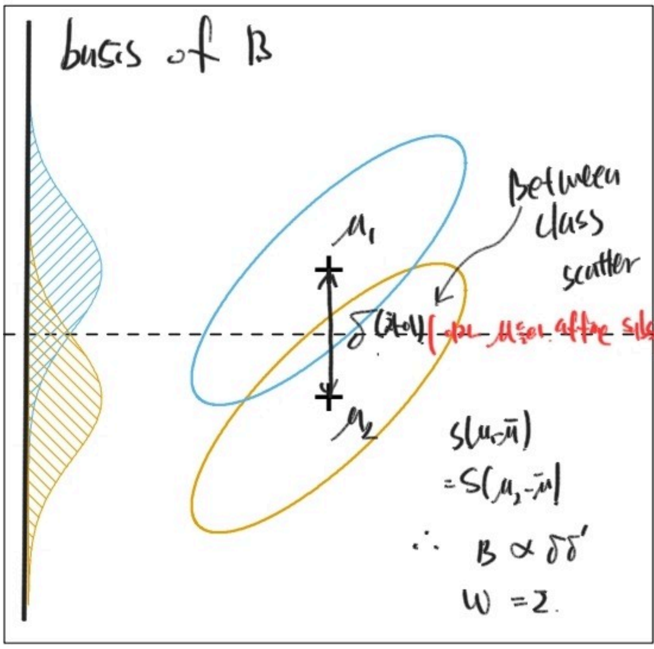
- Example 2 : Discriminant in Binary Class

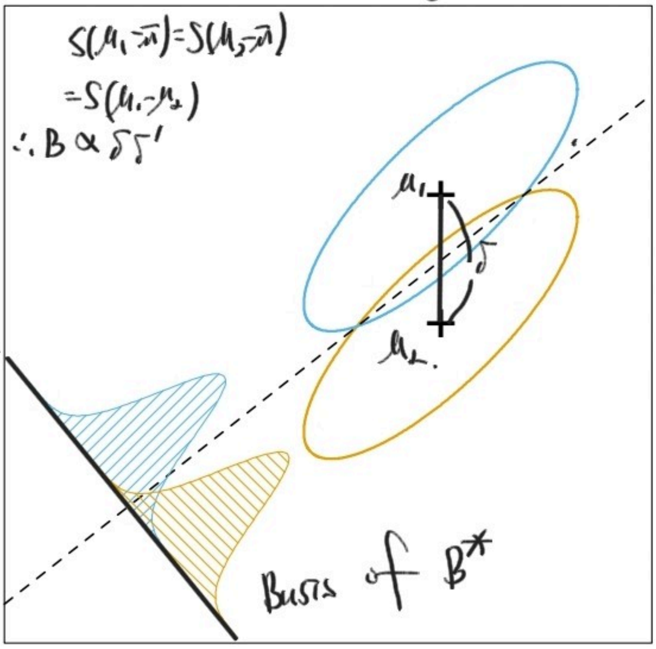
  We want to know eigenvectors of B^* which is same as eigenvectors of $W^{-1}B$

  $B \propto \delta\delta'$ where δ is difference between centroids

$$\Sigma^{-1}\delta\delta' v = \lambda v$$
$$\Sigma^{-1}\delta C = \lambda v$$
$$\therefore v \propto \Sigma^{-1}\delta$$

-





LDA as an optimization problem

- Canonnical LDA (Fisher 1936)
- In 2 dimensional problem. Fisher's original idea : Find the linear combination $Z = a'X$ such that the betweenclass variance is maximized relateve to the within-class variance
- maximizing the Rayleigh quotient

$$\underset{a}{argmax} \frac{a'Ba}{aWa}$$

  which is same as

$$\underset{a}{argmax} \;\; a'Ba \;\;\; \text{subject to } a'Wa = 1$$

- In Fisher's idea, Gaussian distribution is not assumed.
- This is generalized by Rau(1948)

  in order to generalize, Gaussian distribution and have same with-in cov assumption is needed

$$\underset{a}{argmax} \;\; a'Ba \;\;\; \text{subject to } a'Wa = 1$$

$$\underset{a}{argmax} \;\; a'Ba \;\;\; \text{subject to } a'Wa = 1, a'Wa_1 = 0$$

  $\rightarrow$ Which is same as finding eigen values of B^*

$$a^* = W^{1/2}a$$

$$a'W^{1/2}W^{-1/2}BW^{-1/2}W^{1/2}a = a^{*'}B^*a^*$$

$$\therefore \underset{a^*}{argmax} \;\; a^{*'}B^*a^* \;\;\; \text{subject to } a^{*'}a = 1$$

  Reason why this can be generallized

$$\delta_k(x) = x'\Sigma^{-1}\mu_k - \frac{1}{2}\mu_k'\Sigma^{-1}\mu_k + \log(\pi_k)$$

$\delta(x)$ haverΣ^{-1} and μ_k whoose space is made by W^{-1} and B

Two class LDA obtaind by the regression

- In two class response, with class sizes n_1, n_2 and the target coded as $-\frac{n}{n_1}, -\frac{n}{n_2}$
- LDA rule classifies to class 2 if

$$x'\hat{\Sigma}^{-1}(\hat{\mu}_2 - \hat{\mu}_1) > \frac{1}{2}\hat{\mu}_2'\hat{\Sigma}^{-1}\hat{\mu}_2 - \frac{1}{2}\hat{\mu}_1'\hat{\Sigma}^{-1}\hat{\mu}_1 + \log\left(\frac{n_1}{n}\right) - \log\left(\frac{n_2}{n}\right)$$

$\rightarrow$ coefficient of $x \propto \hat{\Sigma}^{-1}(\hat{\mu}_2 - \hat{\mu}_1) \propto W^{-1}B$

- Linear regression classification, by normal equation $X_c'X_c\beta = X_c'y_c$

$$\begin{gathered} T\hat{\beta} = n(\hat{\mu}_2 - \hat{\mu}_1) \\ \left[(n-2)\hat{\Sigma} + \frac{n_1 n_2}{n}\hat{\Sigma}_B\right]\hat{\beta} = n(\hat{\mu}_2 - \hat{\mu}_1) \\ \text{where } \hat{\Sigma}_B = (\hat{\mu}_2 - \hat{\mu}_1)(\hat{\mu}_2 - \hat{\mu}_1)' \end{gathered}$$

$\rightarrow \hat{\beta} \propto \hat{\Sigma}^{-1}(\hat{\mu}_2 - \hat{\mu}_1)$

- Therefore the least squares regression coefficient is identical to the LDA coefficient up to scalar multiple
- This results holds for any distinct coding of the two classes
- Reason why this results is important
  1. We can solve LDA problem not as maximize of convex function but as minimize of convex function

     $\rightarrow$ Computing time is faster
  2. We can add penalty to give sparsity.

LDA as Optimal Scoring

- Suppose we have K classes, and we code the class K as indicator s-vector $Y = (Y_1, \cdots, Y_{K-1})$
- Let θ is scoring vector and S_{11}, S_{22}, S_{12} is sample covariance matrices for $X, Y, (X, Y)$
- Then

$$\begin{gathered} T = NS_{11} \\ B = NS_{11}S_{22}^{-1}S_{21} \end{gathered}$$

- As in CCA(Cannonical Correlation Anaysis), let

$$\begin{gathered} K = S_{11}S_{22}^{-1}S_{21} \\ KK' = S_{11}^{-1/2}S_{12}S_{22}^{-1}S_{21}S_{11}^{-1/2} \end{gathered}$$

The CC vectors $a_k = S_{11}^{-1/2}$ are the eigen vectors of $S_{11}^{-1/2} S_{12} S_{22}^{-1} S_{21} = S_T^{-1} S_B$

- by this method, We can make LDA problem as RSS(Regression) problem

$$\underset{\theta,\beta,\beta_0}{\operatorname{argmin}} \sum_i^n (\theta' y_i - \beta X_i - \beta_0)^2$$

- We can have sparse solution!!

Logistic Regression

- Model

$$p_k(x) = P(Y = k | X = x)$$
$$pdf_{Y|X} = exp\left[\sum_{k=1}^{K-1} I(y = k) \log(\frac{p_k(x)}{p_K(x)}) + \log p_K(x)\right]$$
$$\text{multinomial distribution}$$

- In logistic regression, we assume

$$\log \frac{p_k(x)}{p_K(x)} = \beta_{k,0} + \beta_k' x$$

- MLE

$$\hat{p}_k(x) = \frac{\exp\left(\hat{\beta}_{k,0} + \hat{\beta}_k' x\right)}{1 + \sum_{j=1}^{K-1} \exp\left(\hat{\beta}_{j,0} + \hat{\beta}_j' x\right)}$$
$$\hat{p}_K(x) = \frac{1}{1 + \sum_{j=1}^{K-1} \exp\left(\hat{\beta}_{j,0} + \hat{\beta}_j' x\right)}$$

- Classification rule

$$\hat{k} = \underset{1 \le k \le K}{argmax}\ \hat{p}_k(x)$$
$$\text{which is equivalent to}$$
$$\hat{k} = \underset{1 \le k \le K}{argmax}\ \delta_k(x)$$
$$\text{where}$$
$$\delta_k(x) = \log \hat{p}_k(x) - \log \hat{p}_K(x) = \hat{\beta}_{k,0} + \hat{\beta}_k' x$$

Logistic regression vs LDA

- LDA can be expressed as

$$\begin{aligned}\log \frac{P(Y=k|x)}{P(Y=K|x)} &= \log \frac{\pi_k}{\pi_K} - \frac{1}{2}(\mu_k+\mu_K)'\Sigma^{-1}(\mu_k+\mu_K) + x'\Sigma^{-1}(\mu_k-\mu_K) \\ &= \alpha_{k,0} + \alpha_k' x\end{aligned}$$

- Which is same form as logistic regression. But two methods are not the same!!
- Common component : $P(Y=k|X=x)$ has the same ligit linear form
- Difference

  Logistic : leaves the marginal density of X as as arbitrary, only maximize conditional likelihood

  LDA : fit the parameters by maximizing the full likelihood based on the joint density

$$P(X, Y=k) = P(X|Y=k) \times P(Y=k) = \phi(X; \mu_k, \Sigma)\pi_k$$

- Advantage of LDA
  1. If true $f_k(x)$ is Gaussian, LDA is better (loss of efficiency of about 30% asymptotically in the error rate)
  2. Marginal likelihood can be thought of as a regularizer. $\rightarrow$ will not permit degeneracies
- Disadbantage of LDA

  LDA uses all the points $\rightarrow$ Not robust to outliers