Linear Discriminant Analysis

- Data: predictors matrix $X \in \mathcal{R}^{n \times p}$. Each row X_i represents an observation.
- Reponse y_i: $y_i = k$ then the ith observation belongs to class k, $k = 1, \cdots, K$.
- Logistic regression directly model the probability that a given observation belongs to a class k:$Pr(Y = k|X = x)$ using logistic function (for K=2).
- SVM looks for separating hyperplanes with largest margin to separate classes geometrically.
- Linear discriminant analysis (LDA) takes a different approach: modeling the predictors X separately in each response class, and then estimate $Pr(Y = k|X = x)$.

- Conditional probability: measure the probability of an event A given that another event B has occured: $P(A|B)$.

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

$$P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$$

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A)P(A)}{P(B)}$$

- Example: Draw two cards from a deck of card. Let S1='first card is a spade', S2='second card is a spade'. What is $P(S2|S1)$? $P(S2|S1) = 12/51$. $P(S1 \cap S2) = 13 * 12/(52 * 51) = 3/52$. P(S1)=1/4. Verified: $P(S2|S1) = 12/51$.

- For an observation X, try to classify it to one of the K classes. Response variable Y can take K possible values $1, 2, \cdots, K$.
- π_k be overall probability that a random observation belongs to class k (prior probability): $\sum_{i=1}^{K} \pi_k = 1$.
- π_k can be estimated from the data by $\hat{\pi_k}$:

$$\hat{\pi_k} = \frac{\text{number of samples in class k}}{\text{total number of samples}}$$

- $f_k(X) = Pr(X = x|Y = k)$ be the density function of X for an observation that belongs to class k

- We can show that:

$$
\begin{aligned}
Pr(Y = k|X = x) &= \frac{Pr(Y = k \cap X = x)}{Pr(X = x)} \\
&= \frac{Pr(X = x|Y = k)Pr(Y = k)}{\sum_{i=1}^{K} Pr(Y = i \cap X = x)} \\
&= \frac{\pi_k f_k(x)}{\sum_{i=1}^{K} \pi_i f_i(x)}
\end{aligned}
$$

- $Pr(Y = k|X = x)$ is called posterior probability that an observation X=x belongs to class k.
- If we can estimate all posterior probability, a new observation can classified to the class where it has greatest chance to belong to.
- To estimate $Pr(Y = k|X = x)$, need to estimate $f_k(x)$.

- Class density estimation: Assume that $f_k(x)$ has normal distribution (Gaussian). Also first consider having only one predictor p=1.

$$f_k(x) = \frac{1}{\sqrt{2\pi}\sigma_k} e^{\frac{-1}{2\sigma_k^2}(x-\mu_k)^2}$$

μ_k and σ_k^2 are the mean and variance of the distribution. For now we assume

$$\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2 = \sigma.$$

(all distributions is one normal distribution shifted around).

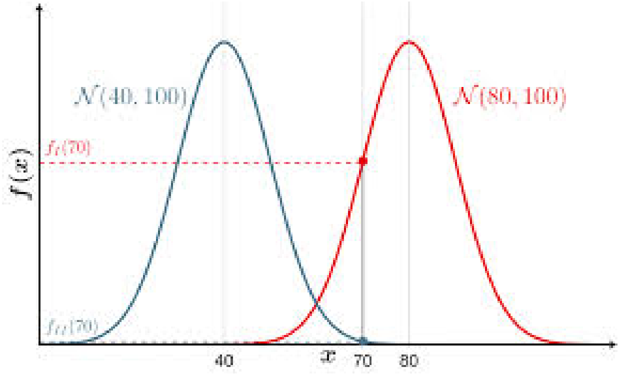


Figure: wikipedia.org

- $$Pr(Y = k|X = x) = \frac{\pi_k \frac{1}{\sqrt{2\pi\sigma}} e^{\frac{-1}{2\sigma^2}(x-\mu_k)^2}}{\sum_{i=1}^{k} \pi_i \frac{1}{\sqrt{2\pi\sigma}} e^{\frac{-1}{2\sigma^2}(x-\mu_i)^2}}$$
- Assign X=x to class k with largest $Pr(Y = k|X = x)$

  $$log(\pi_k \frac{1}{\sqrt{2\pi\sigma}} e^{\frac{-1}{2\sigma^2}(x-\mu_k)^2}) = log(\pi_k) - \frac{1}{2\sigma^2}(x - \mu_k)^2.$$

  which implies the class with largest $\delta_k = log(\pi_k) - \frac{\mu_k^2}{2\sigma^2} + \frac{x\mu_k}{\sigma^2}$
- $\delta_k = log(\pi_k) - \frac{\mu_k^2}{2\sigma^2} + \frac{x\mu_k}{\sigma^2}$ is called discriminant function.

- For K=2 and $\pi_1 = \pi_2$ (2 classes have equal probabilities to occur), the decision boundary correspond to:

$$\delta_1(x) = \delta_2(x)$$
$$x = \frac{\mu_1^2 - \mu_2^2}{2(\mu_1 - \mu_2)} = \frac{\mu_1 + \mu_2}{2}$$

- If $2x(\mu_1 - \mu_2) > \mu_1^2 - \mu_2^2$, x is assigned to class 1. Otherwise, assign x to class 2.

- When $p > 1$, observation $X = (x_1, x_2, \cdots, x_p)$ has multivariate Gaussian (normal) distribution, with a mean vector μ and covariance matrix specific to its class.
  Assume $f_k(x)$ has multivariate normal distribution:

$$f_k(x) = \frac{1}{\sqrt{(2\pi)^p|\Sigma|}} e^{\frac{-1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)}$$

- Observation X=x is assigned to the class k with largest $Pr(Y = k|X = x)$:

$$log(\pi_k \frac{1}{\sqrt{(2\pi)^p|\Sigma|}} e^{\frac{-1}{2}(x-\mu_k)^T\Sigma^{-1}(x-\mu_k)}) = log(\pi_k)$$
$$+x^T\Sigma^{-1}\mu_k - \frac{1}{2}\mu_k^T\Sigma^{-1}\mu_k - \frac{1}{2}x^T\Sigma^{-1}x$$

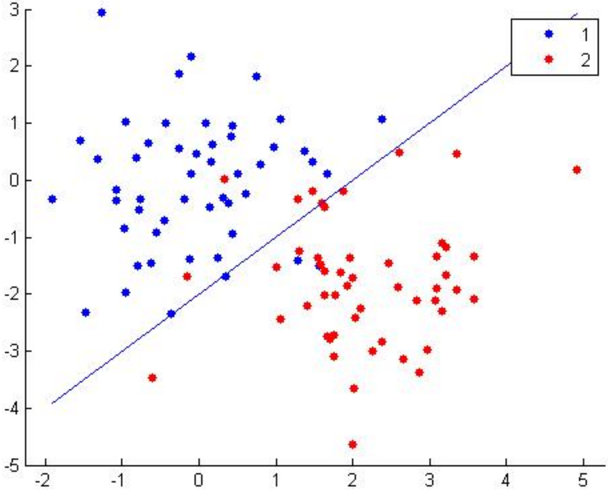
- We have discriminant function:

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

- Given an observation X=x, predict X to be in class k with largest $\delta_k(x)$.
- Decision boundary between class 1 and class 2: $\delta_1(x) = \delta_2(x)$.

$$log(\frac{\pi_1}{\pi_2}) - \frac{1}{2}(\mu_1 + \mu_2)\Sigma^{-1}(\mu_1 - \mu_2) + x^T \Sigma^{-1}(\mu_1 - \mu_2) = 0.$$

- Example: $\pi_1 = 0.5$,$\pi_2 = 0.5$.
- $\mu_1 = [0; 0]$,$\mu_2 = [2; -2]$,$\Sigma = \begin{pmatrix} 1 & 0.1 \\ 0.1 & 1 \end{pmatrix}$
- Let $a = \Sigma^{-1}(\mu_1 - \mu_2)$, $a0 = \frac{1}{2}(\mu_1 - \mu_2)^T\Sigma^{-1}(\mu_1 - \mu_2)$
- Decision boundary: $log(\frac{\pi_1}{\pi_2}) - a0 + x^T a = 0$.


1
2
3
2
1
0
-1
-2
-3
-4
-5
-2
-1
0
1
2
3
4
5

- We need to estimate all the parameters.
- $\hat{\pi}_k = N_k/N$ (number of observations in class k / total number of observations).
- $\hat{\mu}_k = \sum_{y_i=k} x_i/N$.
- $\hat{\Sigma} = \frac{1}{N-K}\sum_{k=1}^{K}\sum_{y_i=k}(x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T$.

- LDA assume observations from each class are draw from Gaussian distribution with a common covariance matrix.
- Assume that each class has its own covariance matrix Σ_k: $X \sim N(\mu_k, \Sigma_k)$.
- Quadratic discriminant function:

$$\begin{aligned}\delta_k(x) &= log(\pi_k) - \frac{1}{2}log(|\Sigma_k|) - \frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k). \\ &= log(\pi_k) - \frac{1}{2}log(|\Sigma_k|) - \frac{1}{2}x^T\Sigma_k^{-1}x - \frac{1}{2}\mu_k^T\Sigma_k^{-1}\mu_k.\end{aligned}$$

- Disriminant function is quadratic in terms of x. (Quadratic discriminant analysis).
- QDA fits data better than LDA but it has more parameters to estimate (Σ_k).

- LDA and QDA in Matlab.
- Function: fitsdiscr.m.
- File:lda_example_1.m, lda_example_2.m