MVA

LDA AND QDA

- Unlike KNN (and all other techniques) LDA and QDA assume the use of distribution over the data
- Allows us to quantify uncertainty over the structure of the data
- i.e. can start to talk about the probability of group assignment

Multivariate normal distribution
- Let $X^T = (X_1, X_2 \dots X_m)$ where $X_1 \dots X_m$ are random variables
- MVN distribution has two parameters:
  - ⇒ Mean μ, an m-dimensional vector
  - ⇒ Covariance matrix Σ with dimension $m \times m$
- A vector X is said to follow MVN $X \sim MVN(\mu, \Sigma)$ if its pdf:

$$f(x \mid \mu, \Sigma) \quad \frac{1}{2\pi^{m/2} |\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right]$$

$|\Sigma|$ is used to denote determinant of Σ

- MVN useful for modelling multivariate data
- If it assumed that the data within a group k follow a MVN distribution with mean μ_k and covariance Σ_k, then the scatter of data should be roughly elliptical
- The mean fixes location of the scatter and covariance affects the shape of the ellipsoid

Mahalanobis Distance
- M distance from point x to mean μ is D where;

$$D^2 = (x-\mu)^T \Sigma^{-1} (x-\mu).$$

Two points have same distance if they are on the same ellipsoid centred on μ

Which is closest?
- Suppose we wish to find the mean μ_k that a point x is closest to as measured by M distance
- Want to find k that minimises $(x-\mu_k)^T \Sigma_k^{-1} (x-\mu_k)$
- Point x is closer to μ_k than it is to μ_l (under M distance) when:

$$(x-\mu_k)^T \Sigma_k^{-1} (x-\mu_k) < (x-\mu_l)^T \Sigma_l^{-1} (x-\mu_l)$$

- This is a quadratic expression for x.

When Covariance is Equal

- If $\Sigma_k = \Sigma$ for all k, previous expression becomes:

$$(x-\mu_k)^T \Sigma^{-1} (x-\mu_k) < (x-\mu_l)^T \Sigma^{-1} (x-\mu_l)$$

- Can be simplified to

$$-2x^T \Sigma^{-1} \mu_k + \mu_k^T \Sigma^{-1} \mu_k < -2x^T \Sigma^{-1} \mu_l + \mu_l^T \Sigma^{-1} \mu_l$$

$$\Rightarrow -2\mu_k^T \Sigma^{-1} x + \mu_k^T \Sigma^{-1} \mu_k < -2\mu_l^T \Sigma^{-1} x + \mu_l^T \Sigma^{-1} \mu_l$$

- Now a linear expression for x
- NOTE named LINEAR and QUADRATIC

Estimating Equal Covariance

- In LDA we need to pool the covariance matrices of individual classes
- Sample Cov matrix Q for set of n observations of dimension m is matrix with elements:

$$q_{ij} = \frac{1}{n-1} \sum_{k=1}^{n} (x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j) \quad \text{for } i=1..m \quad j=1..m$$

- Pooled covariance matrix defined as $Q_p = \frac{1}{n-g} \sum_{l=1}^{g} (n_l - 1) Q_l$

G is number of classes

Q_l is estimated sample covariance matrix for class l

n_l is number of data points in class l

n is total number of data points

- Arises from summing the squares and cross products over data points in all classes

$$W_{ij} = \sum_{l=1}^{g} \sum_{k=1}^{n_l} (x_{ki} - \bar{x}_{li})(x_{kj} - \bar{x}_{lj}) \quad i=1..m \quad j=1..m$$

Hence $W = \sum_{l=1}^{g} (n_l - 1) Q_l$

- n-g degrees of freedom, estimated g group means
- Result for pooled cov matrix: $Q_p = \frac{W}{n-g}$

/04/15

MLA

LDA and QDA

Modelling Assumptions

- Both LDA and QDA are parametric statistical methods
- In order to classify a new observation x into one of the known k groups, we need to know $\mathbb{P}(x \in k | x)$ for $k = 1 \dots K$

  $\Rightarrow$ need to know posterior probability of belonging to each of possible groups in the data
- Then classify new observation as belonging to the class with largest posterior probability
- Baye's Theorem: posterior probability of observation x belonging to group k is

$$\mathbb{P}(x \in k | x) = \frac{\overbrace{\pi_k f(x_i | x \in k)}^{\text{sum product}}}{\underbrace{\sum_{l=1}^{k} \pi_l f(x_i | x \in l)}_{\text{sum product}}} \leftarrow \text{MVN}$$

- Discriminant Analysis assumes that observations from group k follow a MVN distribution with mean μ_k and covariance Σ_k
- That is $f(x | x \in k) = f(x | \mu_k, \Sigma_k) = \frac{1}{(2\pi)^{p/2} |\Sigma_k|^{1/2}} \exp\left[-\frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1} (x-\mu_k)\right]$
- Discriminant analysis assumes values for $\pi_k = \mathbb{P}(x \in k)$, which is proportion of population objects belonging to class k (can be known or estimated)
- Note that $\sum_{k=1}^{K} \pi_k = 1$
- Typically $\pi_k = 1/K$ is used
- π_k sometimes referred to as prior probabilities
- Using all this we can compute $\mathbb{P}(x \in k | x)$ and assign data points to groups so as to maximise this probability

Calculations

- Probability of x belonging to group k conditional on x being known satisfies:

  $\mathbb{P}(x \in k | x) \propto \pi_k f(x | \mu_k, \Sigma_k)$
- Hence: $\mathbb{P}(x \in k | x) > \mathbb{P}(x \in L | x) \Leftrightarrow \pi_k f(x | \mu_k, \Sigma_k) > \pi_L f(x | \mu_L, \Sigma_L)$
- Taking Logs and Substituting in the pdf of MVN, after simplification:

$$\log \pi_k - \frac{1}{2} \log |\Sigma_k| - \frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1} (x-\mu_k)$$
$$> \log \pi_L - \frac{1}{2} \log |\Sigma_L| - \frac{1}{2}(x-\mu_L)^T \Sigma_L^{-1} (x-\mu_L$$

If equal covariance are assumed, then $\mathbb{P}(X \in k|x) > \mathbb{P}(X \in L|x)$ iff:
$$\log \pi_k + x^T \Sigma^{-1} \mu_k - \tfrac{1}{2}\mu_k^T \Sigma^{-1} \mu_k > \log \pi_L + x^T \Sigma^{-1} \mu_L - \tfrac{1}{2}\mu_L^T \Sigma^{-1} \mu_L$$
- Hence name Linear DA.
- If $\pi_k = 1/k$ for all k, reduces to
$$(x - \tfrac{1}{2}(\mu_k + \mu_L))^T \Sigma^{-1} (\mu_k - \mu_L) > 0$$

QDA
- no simplification arises
- covariances different
- If $\pi_k = 1/k$ for all k, some simplification arises

Summary:
- In LDA the decision boundary between class k and L given by
$$\log \frac{\mathbb{P}(k|x)}{\mathbb{P}(L|x)} = \log \frac{\pi_k}{\pi_L} + \frac{\log f(x|k)}{f(x|L)} = 0$$

- Unlike K-nearest neighbour, both LDA and QDA are model based classifiers where $P(\text{data}|\text{group})$ is assumed to follow a MVN:
  - ⇒ Model based assumption allows for generation of the probability for class membership
  - ⇒ MVN assumption means that groups are assumed to follow an eliptical shape
- Whilst LDA assumes groups have the same covariance matrix, QDA permits different covariance structures between groups