MLA

EXAM PAPER 2011 Q2

2A. - Assume there exists a set of K groups in the data
- Assume the form of distribution over the data — parametric
- In order to classify a new observation x into one of known K groups we need to know $P(x \in k | x)$ for $k = 1 \dots K$
- Need to know posterior probability of belonging to a group
- Obs from group k follow a MVN distribution with mean μ_k and cov Σ
- Assume value for $\pi_k = P(x \in k)$ which is the proportion of population objects belonging to class k.
- In LDA equal cov assumed
- Not in QDA - different cov matrix $\Rightarrow$ different between groups

B CROSS VALIDATION
- Process of testing a model on more than one sample
Plots the two numeric variables of the salmon data with colours being determined by true classification and symbols being determined by the resulting classification of "leave one out" LDA

C FL=13 RW=12 CL=29 CW=33 BD=12

$x^T = (13, 12, 29, 33, 12)$

$\left(x - \frac{1}{2}(\mu_K + \mu_L)\right)^T \Sigma^{-1} (\mu_K - \mu_L) > 0$

$13(-0.9) + 12(-1.67) + 29(0.91) + 33(0.13) + 12(-0.18)$
$= -11.8 - 3.22$

$\Rightarrow$ female ogSlyn to group 2

D Logistic regression uses a logit link function to link the probability of class assignment to a linear function of data points
$\log \frac{P(Y=k)}{P(Y=L)} = \alpha + \beta_1 x_1 + \beta_2 x_2 + \dots$
- Class is yes/no - Binary outcome - LDA class
- Parameters found through MLE

In LDA we have $\log \frac{P(Y=k)}{P(Y=L)} = \log \frac{\pi_k}{\pi_L} + \log \frac{f(x|x \in k)}{f(x|x \in L)} > 0$
- Difference is how to find $P(Y=L)$
LDA - no relation to which class belong to.

KNN - classify to nearest group
LDA uses mean distance
LDA is parametric KNN not

KNN doesn't allow for probability assignment of class membership

As number of KNN increase, will be assigned to class which is most prevalent in data

LDA - separated by hyper plane

LDA - minimize variance within each group, maximize distance between groups

KNN - classify to nearest group
LDA uses mean distance
LDA is parametric KNN not

KNN doesn't allow for probability assignment of class membership

As number of KNN increase, will be assigned to class which is most prevalent in data

LDA - separated by hyper plane

LDA - minimize variance within each group, maximize distance between groups