CSCI 6521
Advanced Machine Learning I

Chapter #2

LDA & QDA

Md Tamjidul Hoque

Objectives

- Here we will extend our previous models for classification problems.
- Approaches are similar to regression problem; however the outputs are broken into discrete ranges and labeled categorically:
    - Since our predictor *G*(*x*) takes values in a discrete set *G*, we can always divide the input space into a collection of regions labeled according to the classification.
    - For linear methods for classification, the decision boundaries are linear.

Idea(s) to Obtain Linear Decision Boundaries

- First to fit linear regression models to the class *indicator variables*, and then to classify to the largest fit.

- Let, there are *K* classes, labeled: 1, 2, …, *k*. The fitted linear model for the *k*th indicator response variable can be:

- $$\hat{f}_k(x) = \hat{\beta}_{k0} + \hat{\beta}_k^T x$$

- The decision boundary between class k and ℓ is that set of points for which

$$\hat{f}_k(x) = \hat{f}_l(x)$$

$\{x:(\hat{\beta}_{k0}-\hat{\beta}_{l0})+(\hat{\beta}_k-\hat{\beta}_l)^T x=0\}$, an affine set or hyperplane,

models *discriminant functions* $\delta_k(x)$ for each class *k*.

Idea(s) to Obtain (Linear) Decision Boundaries …

- We have seen posterior probabilities Pr($G = k \mid X = x$) as a discriminating function – if it is linear in x, then the decision boundary will be linear as well.

- if there are two classes, a popular model for the posterior probabilities is:

$$\Pr(G = 1|X = x) = \frac{\exp(\beta_0 + \beta^T x)}{1 + \exp(\beta_0 + \beta^T x)},$$

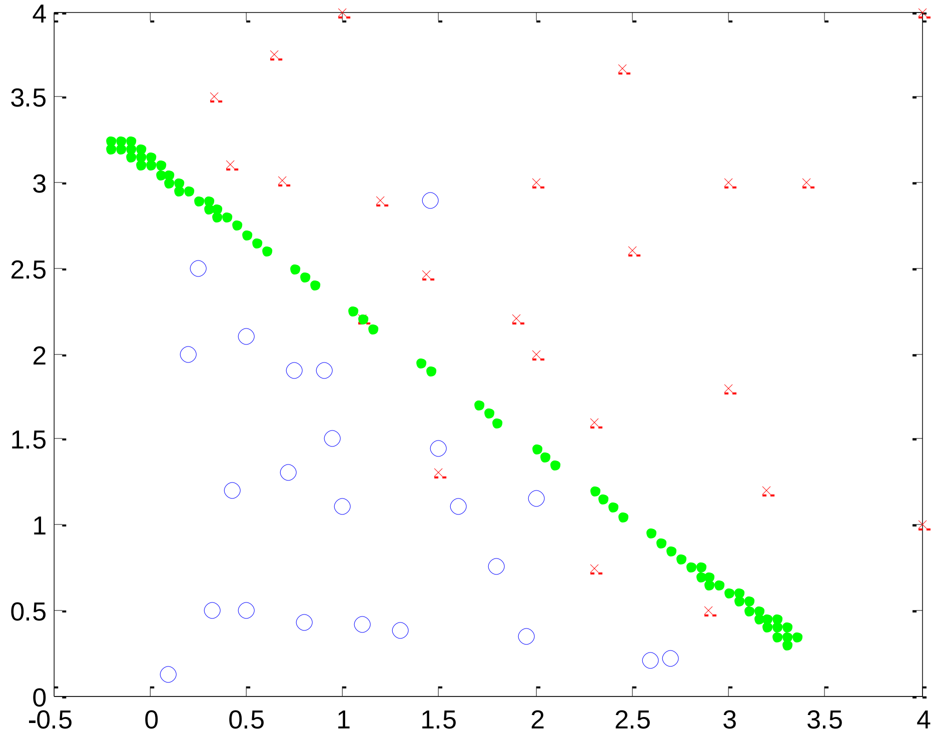
$$\Pr(G = 2|X = x) = \frac{1}{1 + \exp(\beta_0 + \beta^T x)}.$$

Idea(s) to Obtain Linear Decision Boundaries …

- For our cancer example, we can also try drawing a decision boundary by a simple approach: Pr(B |X=x) = Pr(M|X=x)

```
load B.txt;
plot(B(:,1), B(:,2), 'o');
hold on;
load M.txt
plot(M(:,1), M(:,2), 'rx');
Mu0 = [1.1052 1.1286];
Mu1 = [2.0552 2.4578];
Sigma= [1.09229995 0.03073048;
            0.03073048 1.29917801];
```

```
for i=-4:0.05:6,
  for j=-4:0.05:6,
   X= [ i j];
    Px0 = mvnpdf(X,Mu0,Sigma);
    Px1 = mvnpdf(X,Mu1,Sigma);
     if abs(Px0 - Px1) <0.001 & Px0 > 0.01
         plot(i,j, '.g');
      end
   end
end
```



Idea(s) to Obtain Linear Decision Boundaries …

- If we take the ***log-odds*** or ***logit*** of them, we get:

$$\log \frac{\Pr(G = 1 \mid X = x)}{\Pr(G = 2 \mid X = x)} = \beta_0 + \beta^T x$$

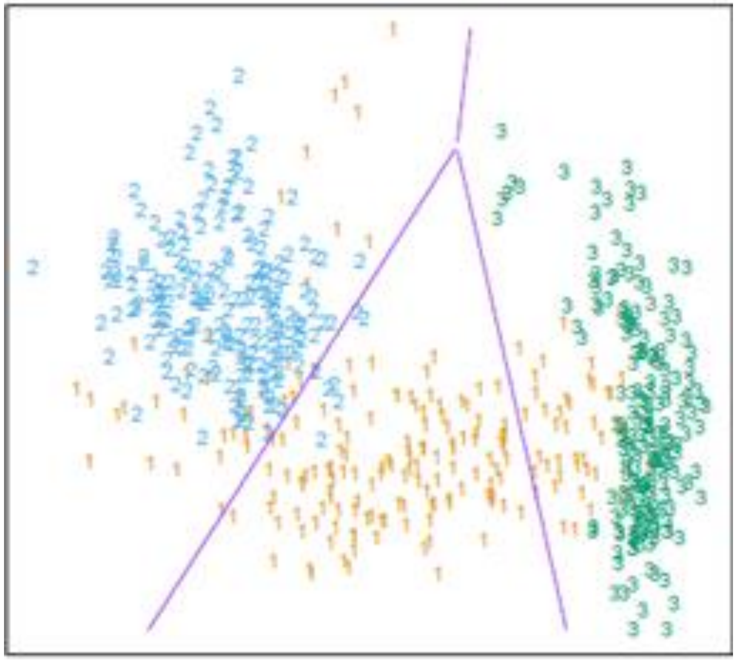
- The decision boundary is the set of points for which the log-odds are zero, and this is a hyperplane defined by $\{ x \mid \beta_0 + \beta^T x = 0\}$.

- We have two very popular but different methods that result in linear *log-odds* or *logits*:
  - linear discriminant analysis (LDA) and
  - linear logistic regression.

Idea(s) to Obtain Decision Boundaries

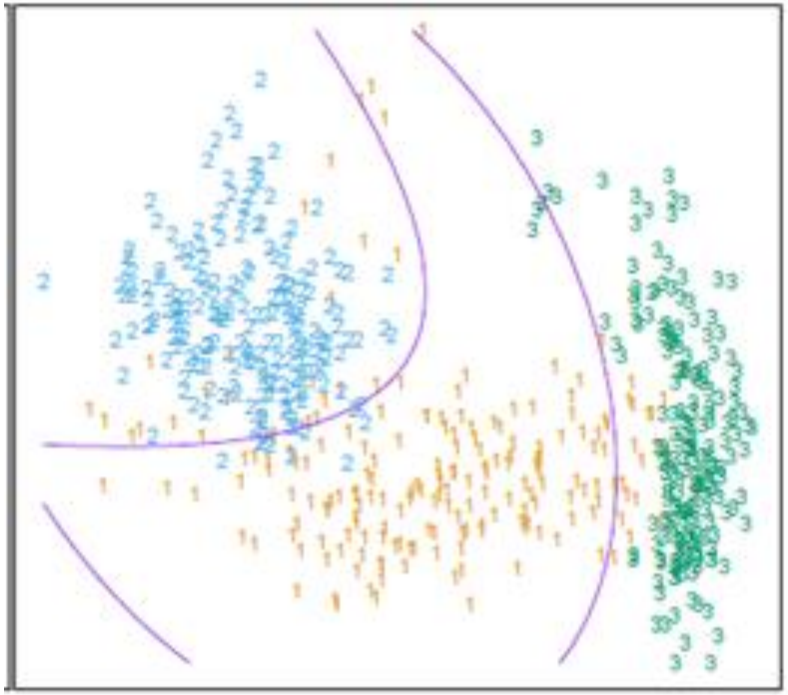
- We will mostly be focusing linear decision boundaries in the lecture. However, there is considerable scope for generalization:
  - For example, we can expand our variable set X_1, …, X_p by including their squares and cross-products X_1^2, X_2^2, … , X_1X_2, …,
  - Linear functions in the augmented space map down to quadratic functions in the original space—hence linear decision boundaries to quadratic decision boundaries

Idea(s) to Obtain Decision Boundaries



The left plot shows some data from three classes, with linear decision boundaries found by linear discriminant analysis

The right plot shows quadratic decision boundaries. These were obtained by finding linear boundaries in the five-dimensional space X_1, X_2, X_1X_2, X_1^2, X_2^2. Linear inequalities in this space are quadratic inequalities in the original space.



This approach can be used with any basis transformation $h(X)$ where $h : \Re^p \rightarrow \Re^q$ with q > p.

Linear Regression of an Indicator Matrix …

- There is a serious problem with the regression approach when the number of classes *K* ≥ 3, especially prevalent when *K* is large.

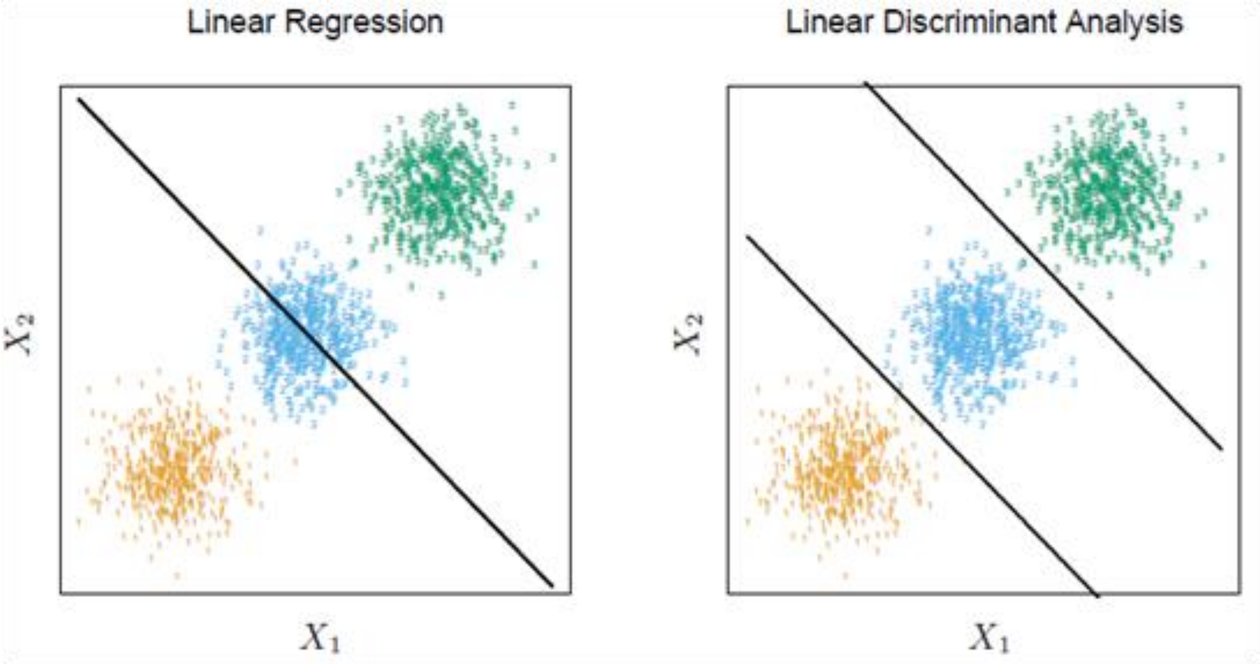


Figure: The data come from three classes in $\Re^2$ and are easily separated by linear decision boundaries. The right plot shows the boundaries found by linear discriminant analysis (*LDA*). The left plot shows the boundaries found by linear regression of the indicator response variables. The middle class is completely masked (never dominates).

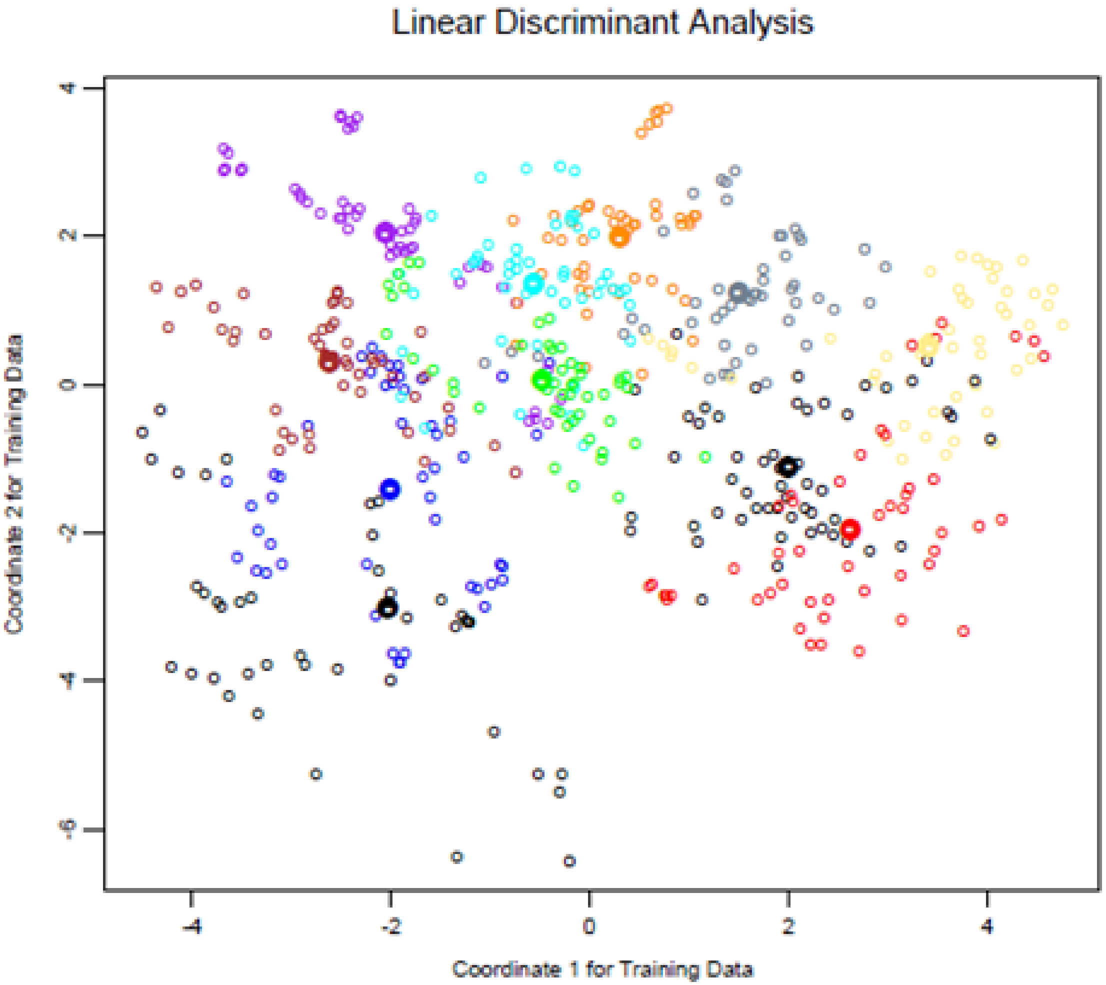
Figure shows an extreme situation when *K* = 3. The three classes are perfectly separated by linear decision boundaries, yet linear regression misses the middle class completely.

Linear Regression of an Indicator Matrix …

- For this simple example a quadratic rather than linear fit (for the middle class at least) would solve the problem.
- However, it can be seen that if there were four rather than three classes lined up like this, a quadratic would not come down fast enough, and a cubic would be needed as well.
- A loose but general rule is that if $K \geq 3$ classes are lined up, polynomial terms up to degree $(K - 1)$ might be needed to resolve them.
-

LDA for a Complex Problem

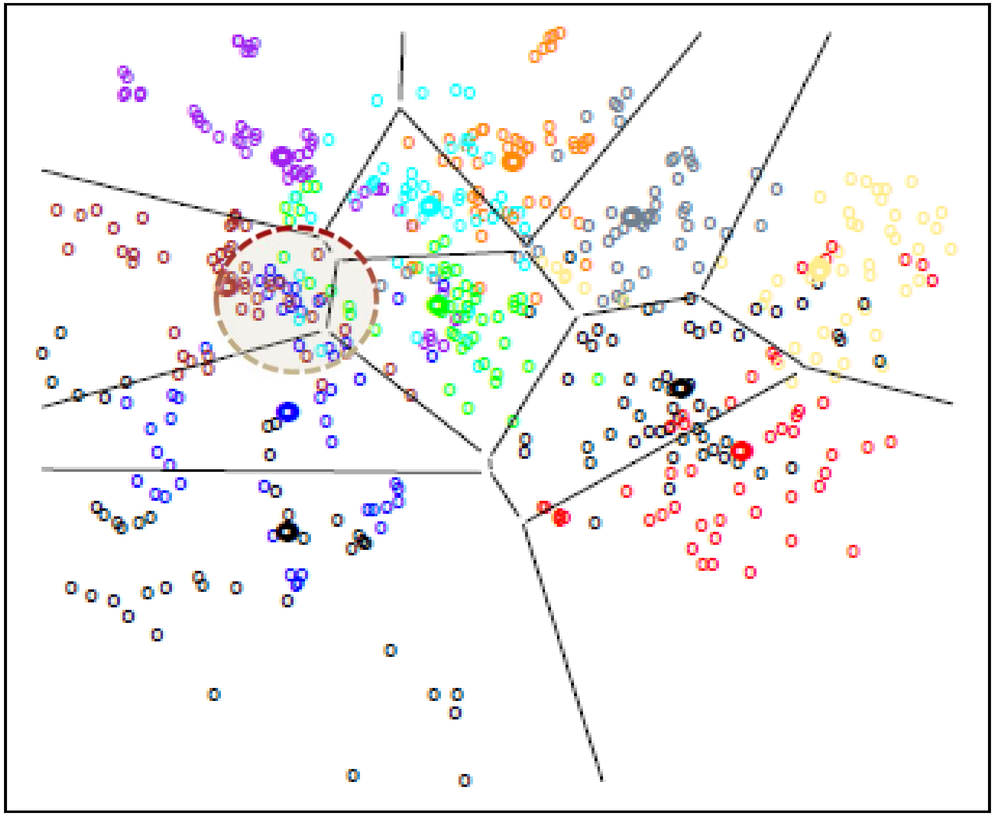
- However, for large *K* and small *p* such maskings naturally occur.



- The figure is a projection of the training data for a **vowel recognition** problem onto an informative two-dimensional subspace.
- There are $K = 11$ classes in $p = 10$ dimensions, of which three account for 90% of the variance.
- This is a difficult classification problem

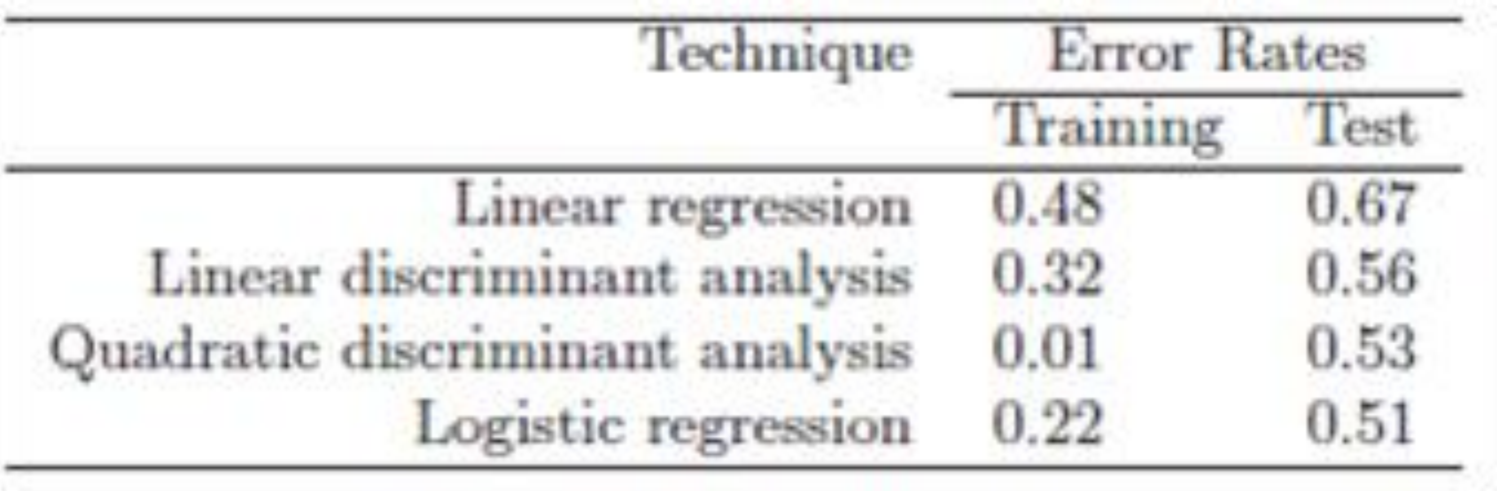
LDA for a Complex Problem …

- However, for large *K* and small *p* such maskings naturally occur.



- The figure is a projection of the training data for a **vowel recognition** problem onto an informative two-dimensional subspace.
- There are $K = 11$ classes in $p = 10$ dimensions, of which three account for 90% of the variance.
- This is a difficult classification problem

Performance Comparisons



Technique	Error Rates	
	Training	Test
Linear regression	0.48	0.67
Linear discriminant analysis	0.32	0.56
Quadratic discriminant analysis	0.01	0.53
Logistic regression	0.22	0.51

- Training and test error rates using a variety of linear techniques on the vowel data.

- There are eleven classes in ten dimensions, of which three account for 90% of the variance.

- We see that linear regression is hurt by masking, increasing the test and training error by over 10%.

Overview of the Next Parts of the Presentation

- We will mostly cover by chalk and talk for the derivations of the following approaches:
  - Linear Discriminant Analysis (LDA)
  - Quadratic Discriminant Analysis (QDA)
  - ~~Logistic Regression (Logistic Classification)~~

Linear Discriminant Analysis

- ***Decision theory for classification*** (Section 2.4) tells us that we need to know the class posteriors Pr(G|X) for optimal classification.

- Suppose $f_k(x)$ is the class-conditional density of X in class $G = k$

- let π_k be the prior probability of class k, with $\sum \pi_l$=1.

- Bayes theorem gives us:

$$\Pr(G = k|X = x) = \frac{f_k(x)\pi_k}{\sum_{l=1}^{K} f_l(x)\pi_l}$$

Linear Discriminant Analysis …

- In comparing two classes *k* and *ℓ*, it is sufficient to look at the log-ratio, and we see that:

$$\log \frac{\Pr(G=k \mid X=x)}{\Pr(G=l \mid X=x)} = \log \frac{f_k(x)}{f_l(x)} + \log \frac{\pi_k}{\pi_l}$$

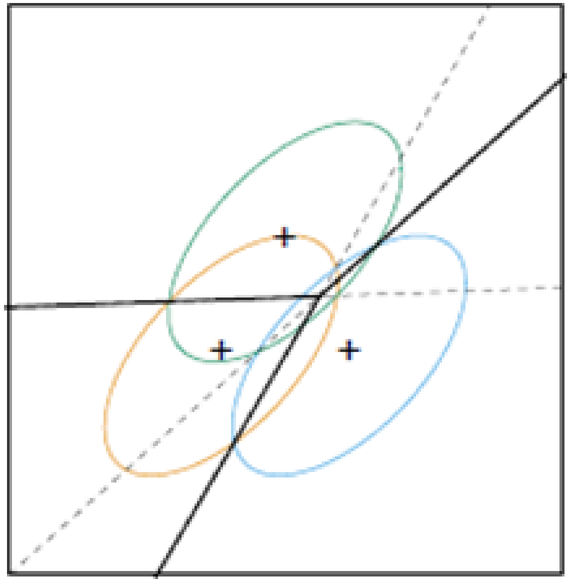
- Here we assumes, that we model each class density as multivariate Gaussian, that is:

$$f_k(\mathbf{x}) = \frac{1}{(2\pi)^{p/2} \left|\mathbf{\Sigma}_k\right|^{1/2}} \exp\left\{ -\frac{1}{2} (\mathbf{x}-\boldsymbol{\mu}_k)^T \mathbf{\Sigma}_k{}^{-1} (\mathbf{x}-\boldsymbol{\mu}_k) \right\}$$

- we see that the linear discriminant functions

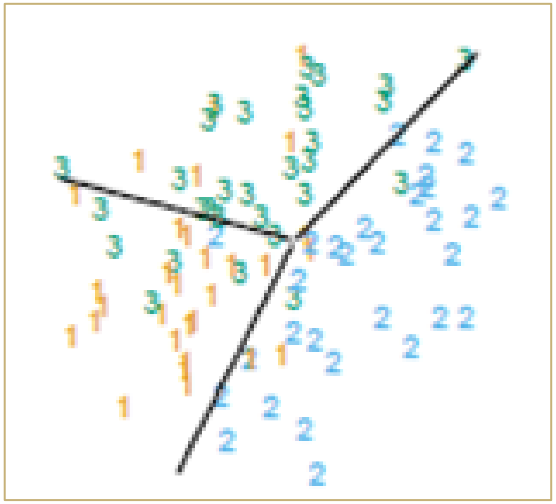
$$\delta_k(x) = x^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + \log \pi_k$$

Linear Discriminant Analysis …



- The left panel shows three Gaussian distributions, with the same covariance and different means.
- Included are the contours of constant density enclosing 95% of the probability in each case.
- The Bayes decision boundaries between each pair of classes are shown (broken straight lines), and the Bayes decision boundaries separating all three classes are the thicker solid lines (a subset of the former).

- Notice that the decision boundaries are not the perpendicular bisectors of the line segments joining the centroids. This would be the case if the covariance $\mathbf{\Sigma}$ were spherical $\sigma^2\mathbf{I}$, and the class priors were equal.
- On the right we see a sample of 30 drawn from each Gaussian distribution, and the fitted LDA decision boundaries.



Quadratic Discriminant Analysis …

- If the Σ_k are not assumed to be equal, then the pieces quadratic in *x* remain. We then get *quadratic discriminant functions:*

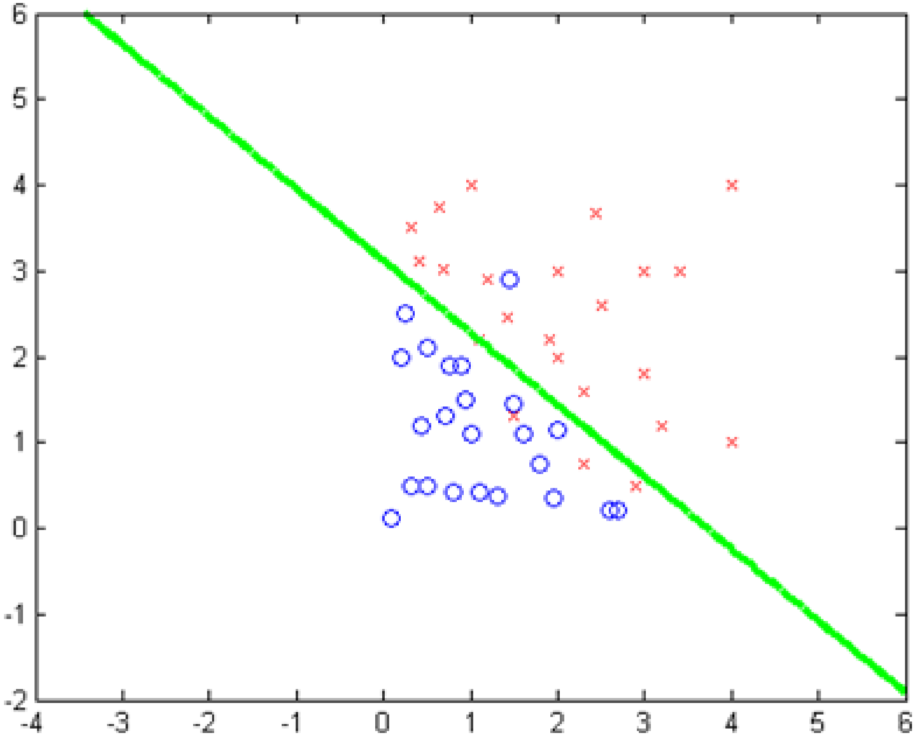
$$\delta_k(x) = -\frac{1}{2}\log|\Sigma_k| - \frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1}(x-\mu_k) + \log \pi_k$$

- The decision boundary between each pair of classes *k* and *ℓ* is described by a quadratic equation:

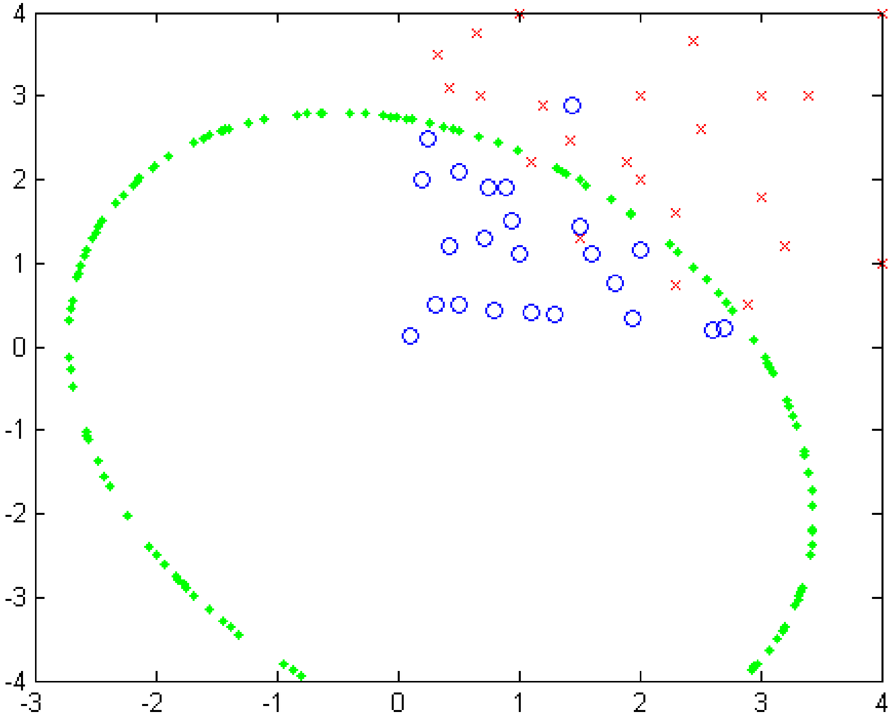
$$\{x : \delta_k(x) = \delta_l(x)\}$$

Cancer Dataset : LDA versus QDA

- We can implement LDA and QDA for our cancer example (see the **EXERCISE_Chapter_5_Classification** and the data section for details):

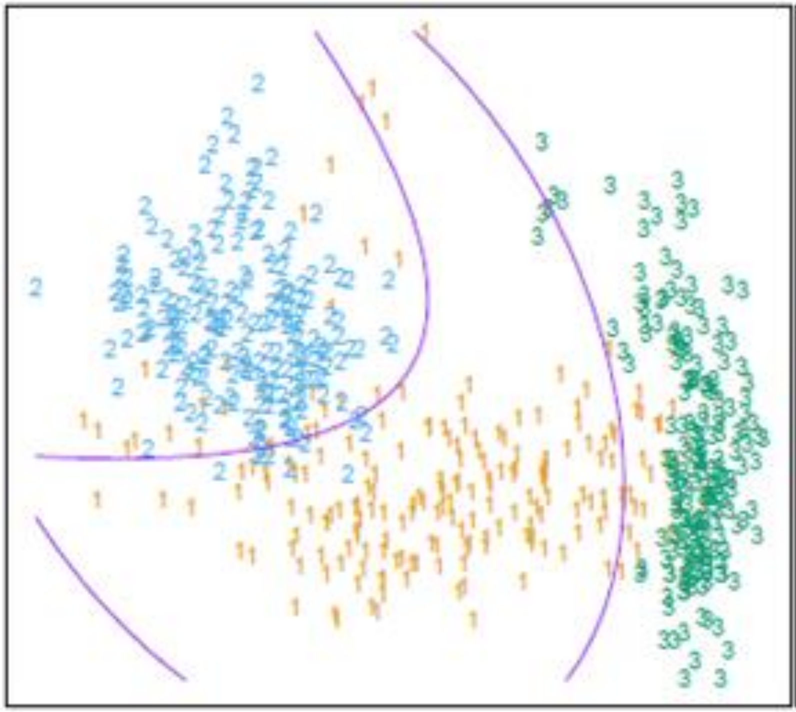


LDA

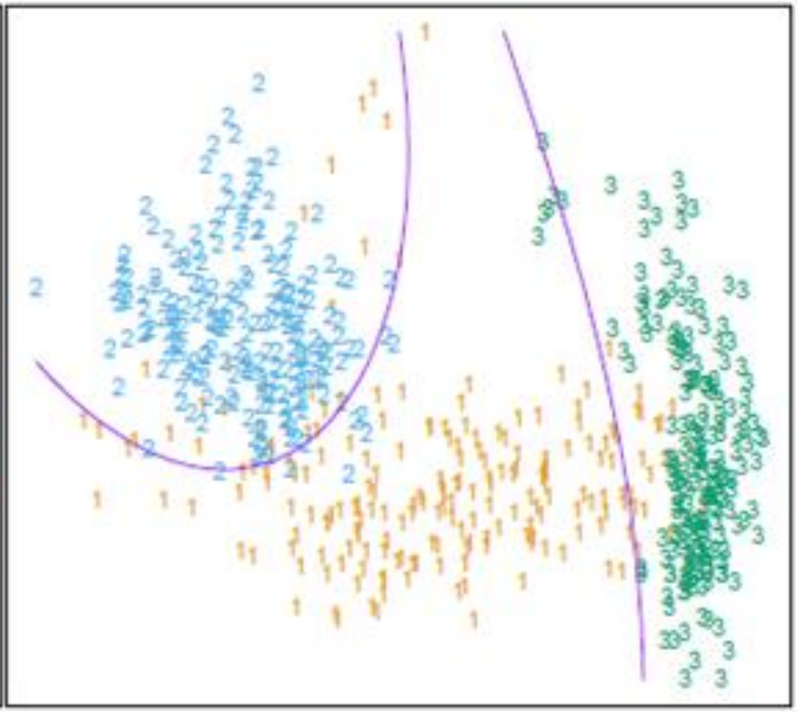


QDA

LDA in Augmented (Quadratic) space Versus QDA



LDA …



QDA

- The left plot shows the quadratic decision boundaries for the data obtained using LDA in the five-dimensional space $X_1, X_2, X_1X_2, X_1^2, X_2^2$.
- The right plot shows the quadratic decision boundaries found by QDA.
- The differences are small, as is usually the case.
- QDA is the preferred approach, with the LDA method a convenient substitute.

LDA versus QDA

- The estimates for QDA are similar to those for LDA, except that separate covariance matrices must be estimated for each class.
- When p is large this can mean a dramatic increase in parameters. For K classes there will be
  - there are (K − 1) × (p + 1) parameters for LDA.
  - there will be (K − 1) × {p (p + 3)/2 + 1} parameters for QDA.
- Both LDA and QDA perform well on an amazingly large and diverse set of classification tasks.
- why LDA and QDA have such a good track record?
  - the data can only support simple decision boundaries such as linear or quadratic, and the estimates provided via the Gaussian models are stable.

Logistic Regression (Classification)

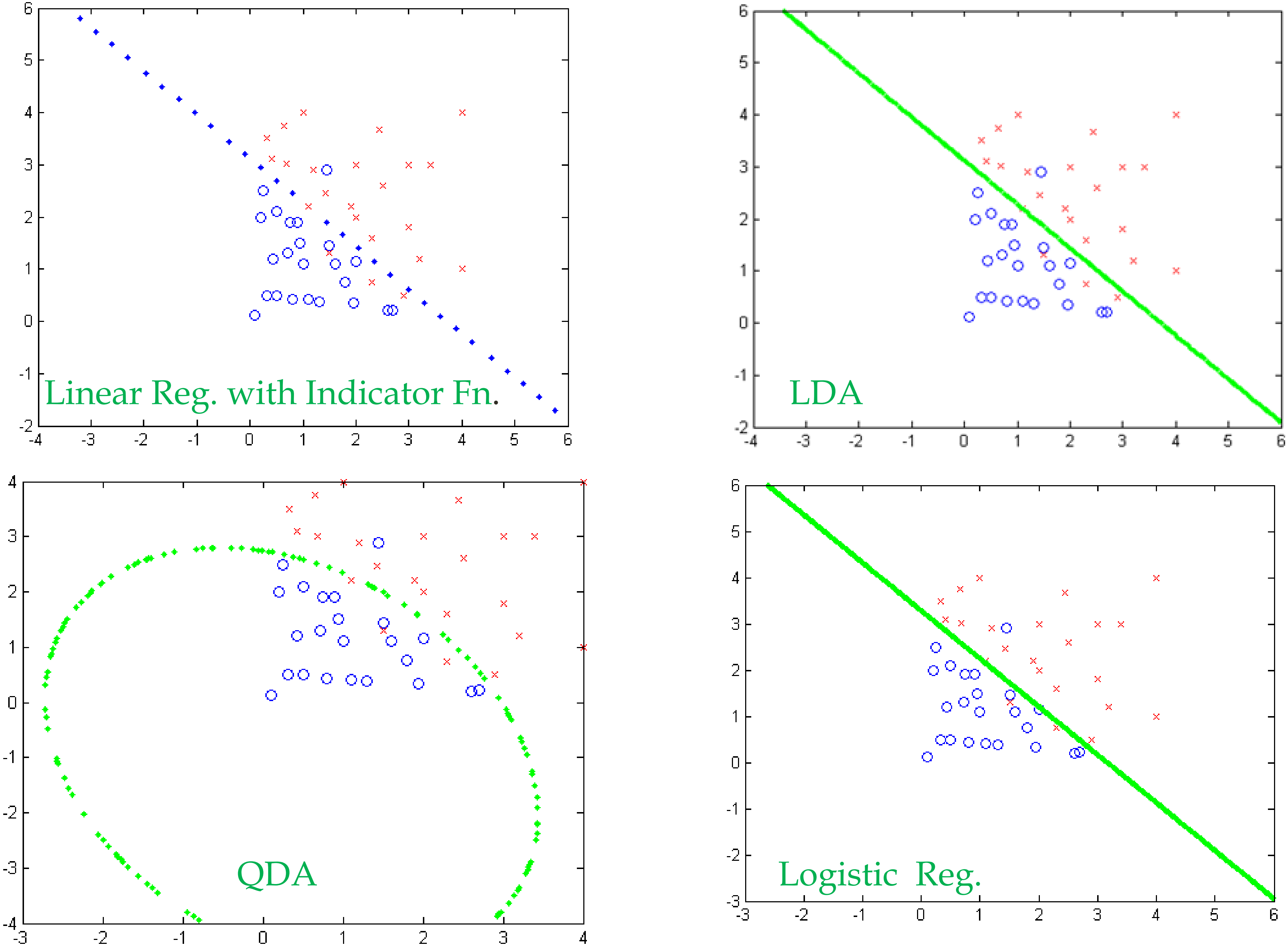
- Logistic regression is widely used because:
  - it is easy to build and it is easy to use
  - it is easy to interpret
  - In terms of complexity, parameters remain proportional to the input features, therefore, complexity remain simple while the number of input feature increases.

- we are trying to model

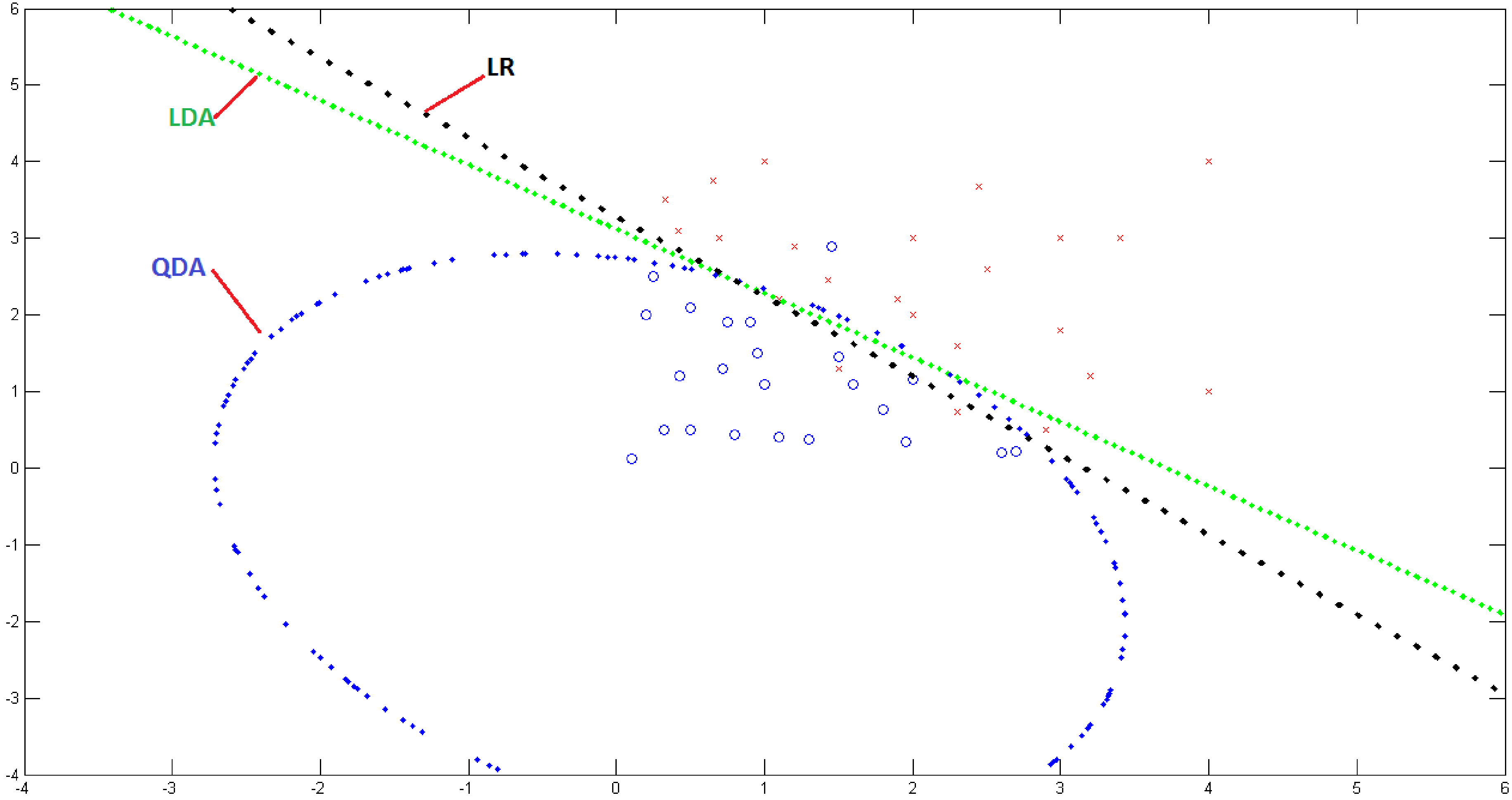
$$\Pr(G \mid X)$$

where we assumed, $X = \{x_1, x_2\}$ and $\hat{G} \in \{\text{ Benign, Malignant}\}$,
x_1= *Size of the tumor* and x_2 = *Age of the tumor*.

Various Boundaries (for Cancer Data)



Various Boundaries (for Cancer Data)



Comparing LDA, QDA and LR by superimposing

-------- x ----------