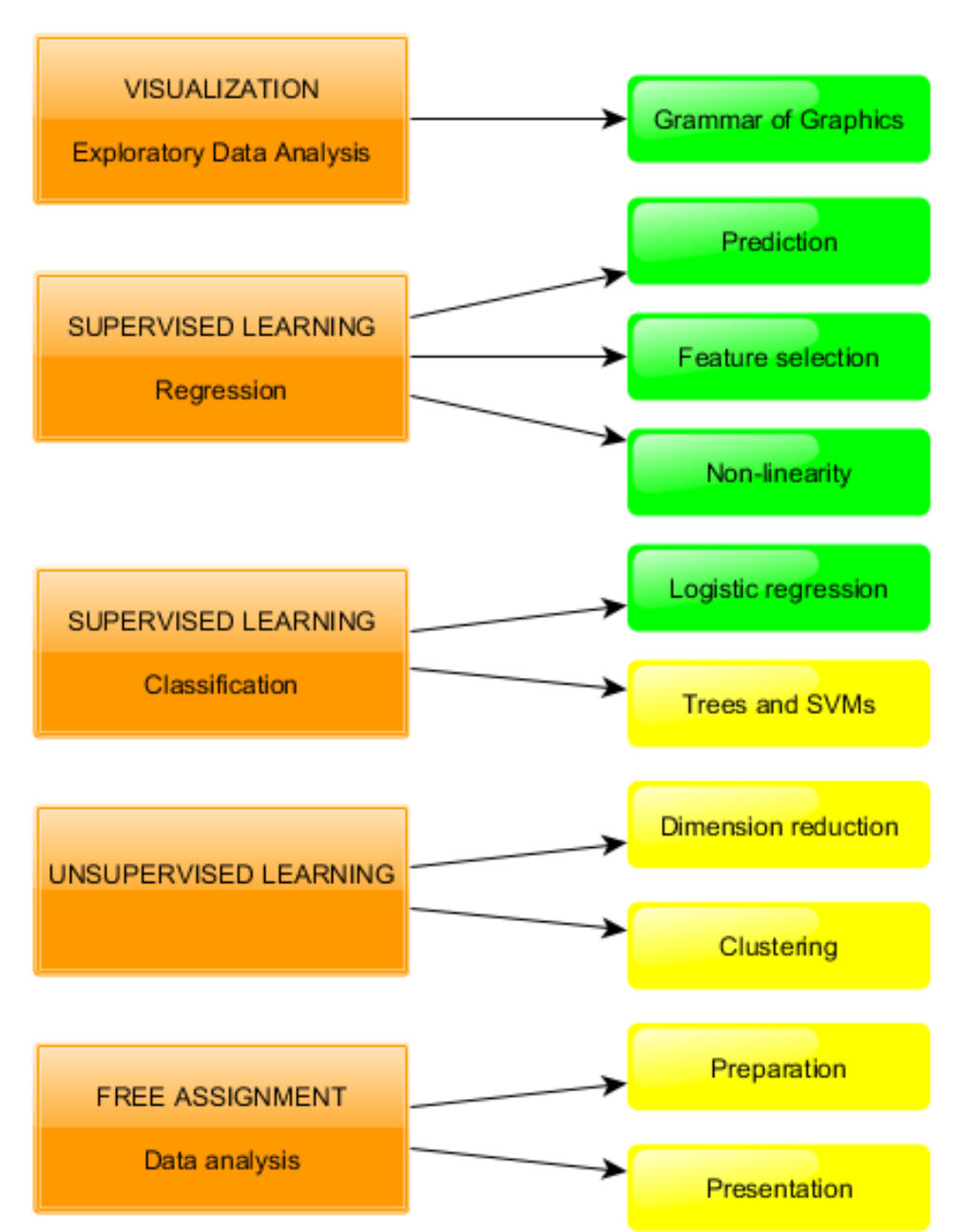
Supervised Learning: Classification

Logistic Regression and Discriminant Analysis

Maarten Cruyff

Program


VISUALIZATION
Exploratory Data Analysis
Grammar of Graphics
SUPERVISED LEARNING
Regression
Prediction
Feature selection
Non-linearity
SUPERVISED LEARNING
Classification
Logistic regression
Trees and SVMs
UNSUPERVISED LEARNING
Dimension reduction
Clustering
FREE ASSIGNMENT
Data analysis
Preparation
Presentation

Content

1. Logistic regression
2. Linear discriminant analysis
3. Quadratic dicriminant analysis
4. Classification criteria
5. Model comparisons

What's classification?

Outcome variable is categorical

- Predict **class membership** from feature set

Estimation

1. Estimate $P(class = j|X)$
2. Assign observation to class with largest probability

Models

- Logistic regression, Discriminant analysis, Trees, Random Forests, Bagging, Boosting, SVM's, etc.

Logistic Regression

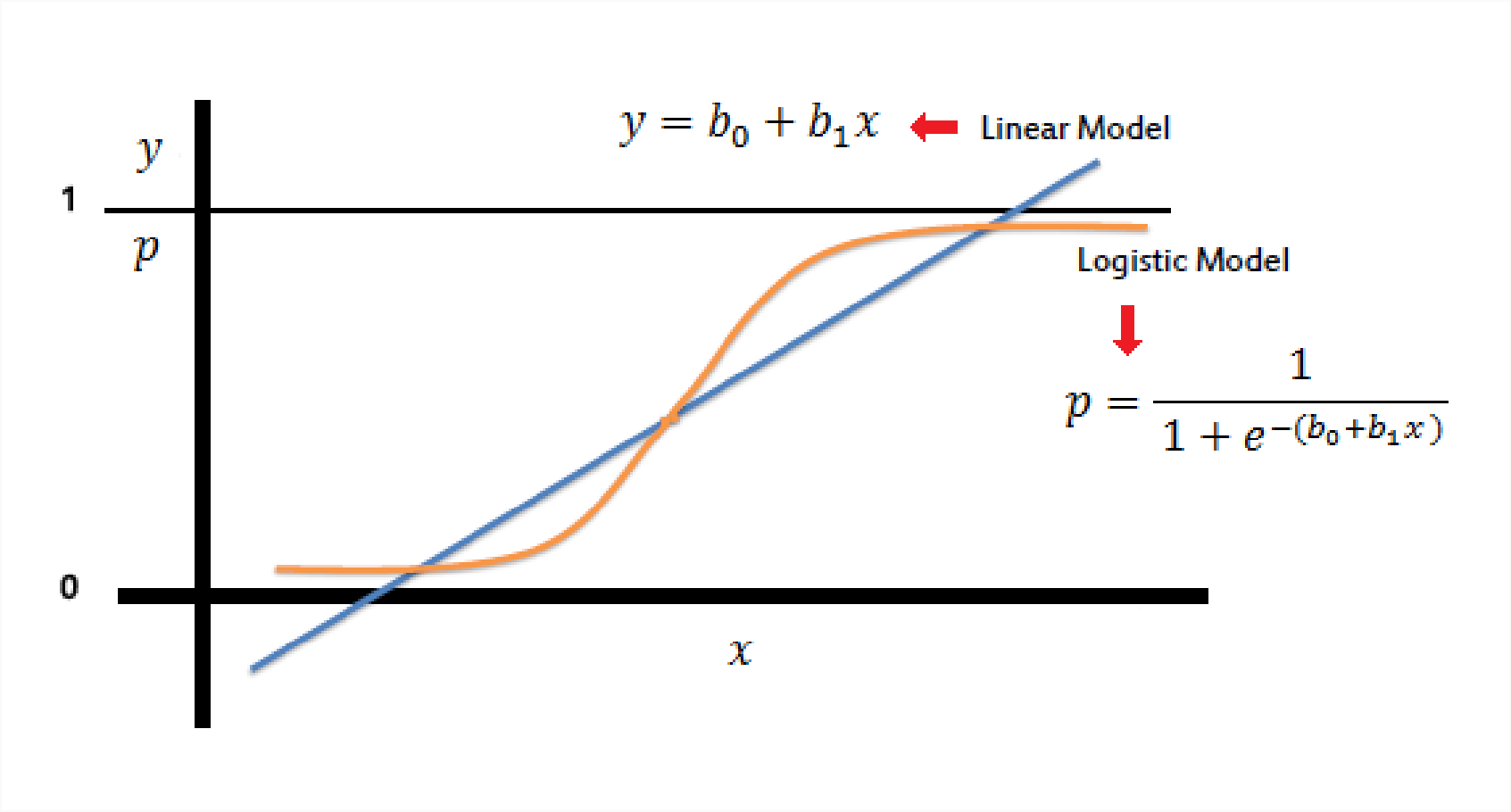
Binary logistic regression (BLR)

BLR predicts probability of a "success" with the logistic function

$$P(Y = \text{success}) = \frac{e^{\beta_0+\beta_1 x_1+\dots}}{1 + e^{\beta_0+\beta_1 x_1+\dots}} = \frac{1}{1 + e^{-(\beta_0+\beta_1 x_1+\dots)}}$$

- Y is coded as "1 = success" and "0 = failure"
- logistic (sigmoid) function ensures $\hat{P}(Y) \in (0, 1)$

Logistic vs linear regression



Reparametrizations

Logistic model is *generalized linear model* with link function

$$\text{logit}(Y) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

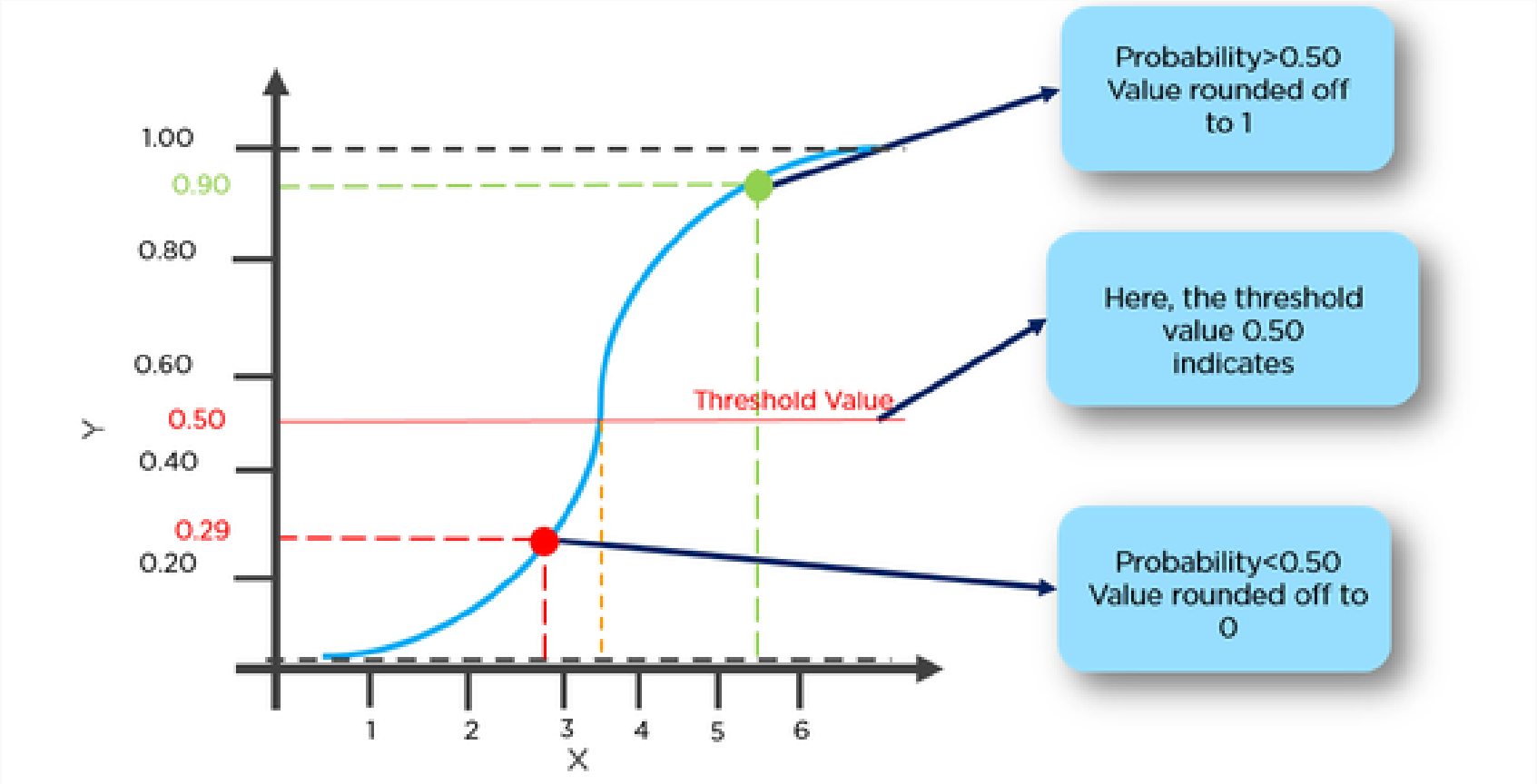
The "logit" is the log of the odds, so that

$$\text{odds}(Y) = \frac{P(Y)}{1 - P(Y)} = e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k}$$

Classification procedure in R

```
fit.glm  <- glm(y ~ x1 + x2 . . .,
                family = "binomial",
                data   = <data>)

pred.glm  <- predict(fit.glm,
                     data = <data>,
                     type = "response")  # alternative is "link"

class.glm <- factor(pred.glm > 0.5,
                    labels = c("success", "failure"))
```

Schematically

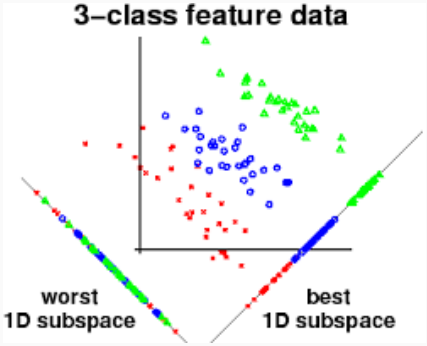


Discriminant Analysis

What's discriminant analysis

Estimation of discriminant functions

- directions in feature space that best separate between classes
- number of functions is $\min(J-1, p-1)$
  - J is number of classes
  - p is number of features



Linear Discriminant Analysis (LDA)

1. Estimate *posterior* probability $P(X = x|Y = j)$ of class $j = 1, \ldots, J$

$$P(Y = j|X = x) = \frac{\pi_j P(X = x|Y = j)}{\sum_{k=1}^{J} \pi_k P(X = x|Y = k)}$$

2. Use Bayes Theorem to find $P(Y = j|X = x)$

3. Assign x to class j with highest posterior probability

- π_j is *posterior probability* class j (estimated by sample proportion)
- $P(X = x|Y = j)$ estimated by conditional sample means

Linear discriminant functions

Linear discriminant functions

$$LD_j = c_{1j}X_1 + \cdots + c_{pj}X_p$$

- LD_1 separates the classes best, LD_2 second best, and so on
- LD's are orthogonal

Assumption $X|Y \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

- X is multivariate normal within each class
- X has covariance matrix $\boldsymbol{\Sigma}$ within each class

Quadratic Discriminant Analysis (QDA)

Estimates covariance matrix $\mathbf{\Sigma}_j$ for each class, with

- quadratic dicriminant functions
- more parameters, so less bias but higher variance

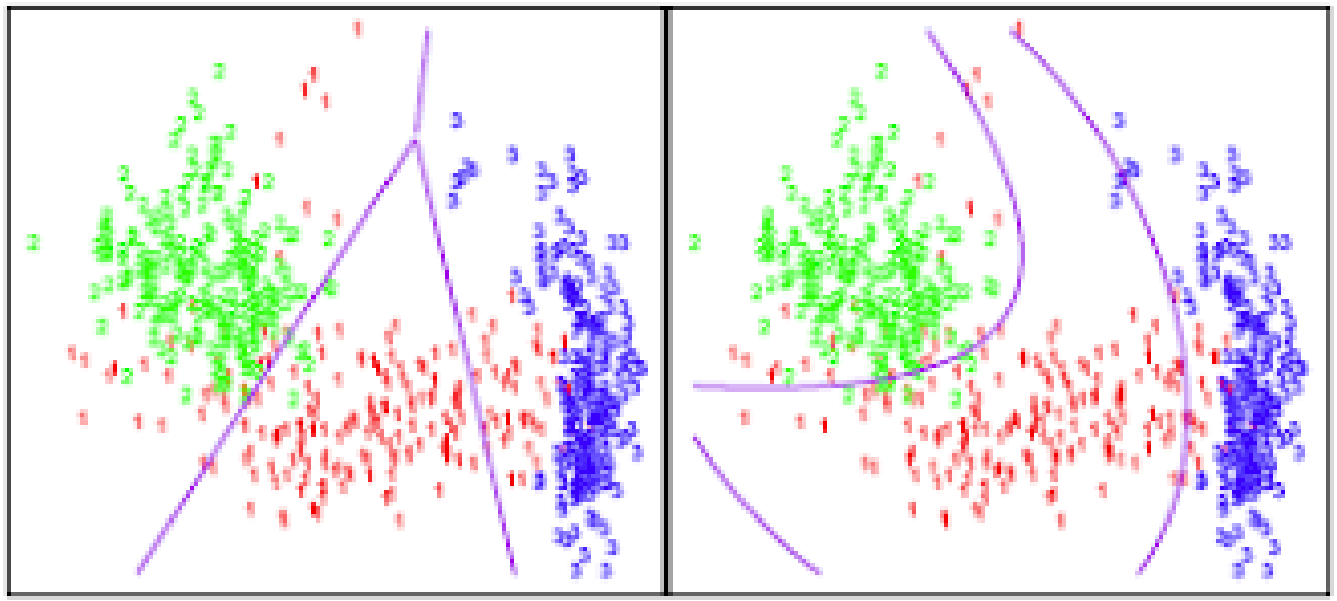


Figure 1: Linear vs quadratic discriminant functions

DA in R

`lda()`and `qda()`

- for LDA:

```
fit.lda   <- lda(formula, data = <data>)

pred.lda  <- predict(fit.lda, newdata = <data>)

prob.lda  <- pred.lda$posterior

class.lda  <- pred.lda$class
```

Functions in basic R package `MASS`

Classification criteria

Goodness-of-fit criteria

a. Deviance statistic $dev = 2 \sum y \log(y/\hat{P}(y))$
   - distance measure observed and estimated

b. AIC
   - deviance plus penalty model complexity (2 times # parameters)

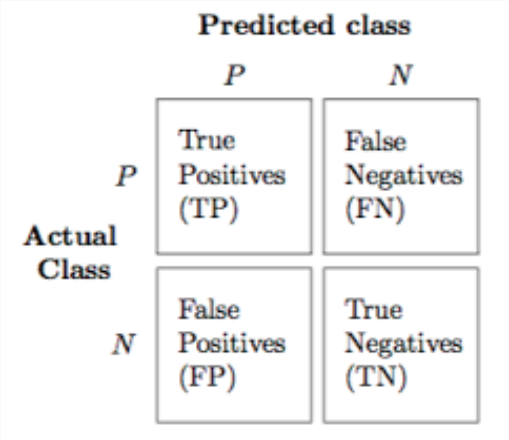
c. Confusion matrix
   - proportions correctly/incorrectly classified

d. ROC curve/AUC
   - Receiving Operation Characteristic (ROC) curve
   - Area Under Curve (AUC)

Deviance/AIC

```
Call:
glm(formula = type ~ glu, family = binomial, data = Pima.te)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2343  -0.7270  -0.4985   0.6663   2.3268

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.946808   0.659839  -9.013   <2e-16 ***
glu          0.042421   0.005165   8.213   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 420.30  on 331  degrees of freedom
Residual deviance: 325.99  on 330  degrees of freedom
AIC: 329.99
```

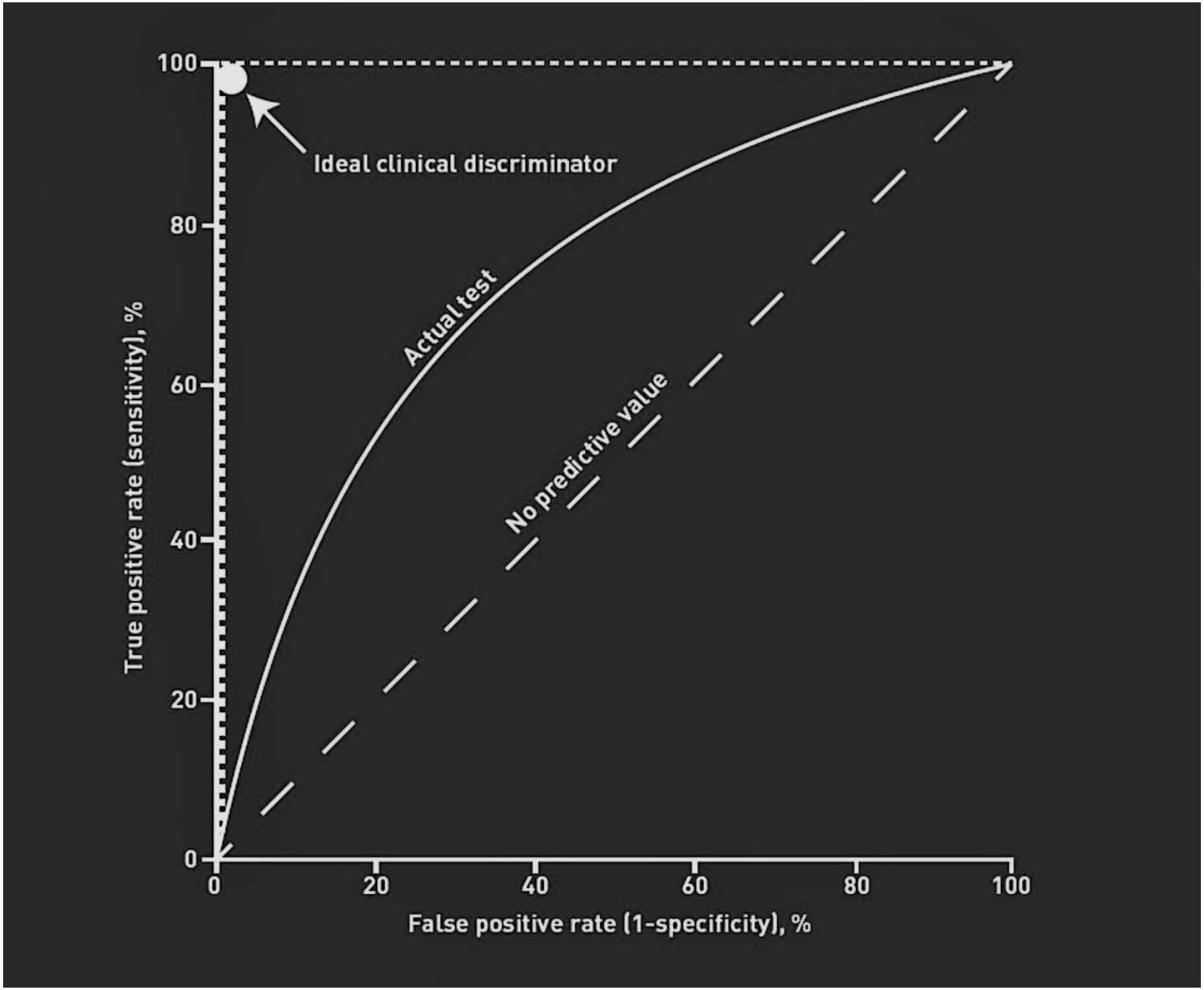
Confusion matrix



- Accuracy: $(TP + TN)/(TP + TN + FP + FN)$
- Sensitivity: $TP/(TP + FN)$
- Specificity: $TN/(TN + FP)$

Misclassification error rate: 1 - Accuracy

ROC and AUC


100
80
60
40
20
0
0
20
40
60
80
100
True positive rate (sensitivity), %
False positive rate (1-specificity), %
Ideal clinical discriminator
Actual test
No predictive value

Cross validation with `train()`

```
fit_cv <- train(formula   = y ~ .,
                data      = <data>,
                method    = c("glm", "lda", "qda"),
                metric    = "Accuracy",
                trControl = trainControl(method = "cv",
                                         number = 5))
```

Cross-validated accuracy:

- `fit_cv$results$Accuracy`

Estimates of model fitted to <data>:

- `fit_cv$finalModel`

Model comparisons

Example data `Pima.te`

Data set on Indian tribe where women often have diabetes

- 332 observations
- 7 predictors

Small data set and no tuning parameters, so

- no train/test partition
- just cross validation with 'caret::train()‘

Estimates BLR

```
Call:  NULL

Coefficients:
(Intercept)        npreg          glu           bp         skin
  -9.514019     0.140944     0.037481    -0.008675     0.013167
        bmi          ped          age
   0.078951     1.110131     0.018055

Degrees of Freedom: 331 Total (i.e. Null);  324 Residual
Null Deviance:      420.3
Residual Deviance: 285.8    AIC: 301.8
```

Estimates LDA

```
Call:
lda(x, grouping = y)

Prior probabilities of groups:
       No       Yes
0.6716867 0.3283133

Group means:
       npreg      glu       bp     skin      bmi       ped      age
No  2.932735 108.1883 70.13004 27.34081 31.63991 0.4645650 29.21525
Yes 4.614679 141.9083 74.77064 32.88991 36.51284 0.6589633 35.61468

Coefficients of linear discriminants:
               LD1
npreg  0.099984437
glu    0.028389957
bp    -0.004602843
skin   0.004726554
bmi    0.052002256
ped    0.615724201
age    0.012235784
```

Estimates QDA

```
Call:
qda(x, grouping = y)

Prior probabilities of groups:
       No       Yes
0.6716867 0.3283133

Group means:
       npreg      glu       bp     skin      bmi       ped      age
No  2.932735 108.1883 70.13004 27.34081 31.63991 0.4645650 29.21525
Yes 4.614679 141.9083 74.77064 32.88991 36.51284 0.6589633 35.61468
```

Confusion matrices of final models

BLR

```
        estimated
observed  No Yes
     No  201  22
     Yes  46  63
```

LDA

```
        estimated
observed  No Yes
     No  199  24
     Yes  47  62
```

QDA

```
        estimated
observed  No Yes
     No  199  24
     Yes  40  69
```

Accuracy

Accuracy crossvalidation versus final model

	CV	Final model
BLR	0.783	0.795
LDA	0.786	0.786
QDA	0.765	0.807

- BLR best CV accuracy
- QDA best final model accuracy

ROC's and AUC's