CIS 490 Machine Learning

Lecture 9

Instructor: (Julia) Hua Fang

2

Last time

We are still in the phase of Supervised Learning

- Supervised Learning
  - Regression
    - Linear regression: Simple & Multiple

      R: Run simple and multiple linear regression using Auto MPG data
    - Regularized linear regression: Ridge & Lasso
      - ✓ Overfitting: Random error vs. deterministic error; Bias-Variance tradeoff
    - CV algorithm in general
      - ✓ CV for choosing optimal tuning parameter λ in the context of regularized linear regression

      R: Run Ridge and Lasso and K-fold CV for choosing λ* using Credit data

Adapted from Jeff Howbert, Greg Shakhnarovich, Patrick Breheny , M. Magdon-Ismail, Patrick Breheny, Jeff Schneider

3

Warning

Machine Learning methods ≠ R/Matlab/Python packages or functions

Eg.

- lasso and Ridge ≠ glmnet
- Introduced CV algorithm is essential to you; the R package glmnet includes the CV algorithm and facilitates users to do cross-validation!

You can develop and name your own packages based on introduced machine learning methods/algorithms, as these available R packages may not serve your special interests!

4

Outline

- Supised Learning
  - **Classification**
    - Logistic regression: Probability, Odds, Log Odds, Logit, logistic function.
    - Classification evaluation in general: Confusion table; classification accuracy (Sensitivity, Specificity, etc)
    - ROC: you are expected to know

      How to generate ROC and evaluate multiple classifiers; Area under the ROC curve (AUC)

Adapted from Jeff Howbert, Greg Shakhnarovich, Patrick Breheny , M. Magdon-Ismail, Patrick Breheny, Jeff Schneider

5

Supervised Learning:

Classification

Recall the contents reviewed in first two weeks:

Y: what type of random variable should be?
What type of distribution would it be?

6

Logistic regression:

7

Logistic regression: Applications

Examples of binary classification problems using logistic regression:

- **Spam Detection** : an email is Spam or not
- **Credit Card Fraud** : a given credit card transaction is fraud or not
- **Health** : a given mass of tissue is benign or malignant
- **Marketing** : a given user will buy an insurance product or not
- **Banking** : a customer will default on a loan.
- **Image:** Cat or dog; clean or dirty room

8

Logistic regression: Y and X

- Y, is discrete/binary, usually assuming binomial distribution.
- X, features/attributes can be categorical or continuous.

9

Logistic regression: must-know concepts and relationship

Must understand the relationship of these concepts:

Probability → Odds→ Log Odds

Logit→ Logistic regression

In class exercises next to help you understand!

10

Logistic Regression: Probability →Odds
in-class exercise

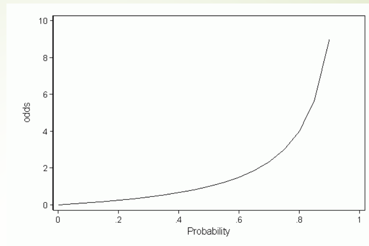
$$Odds = \frac{the\ probability\ of\ success}{the\ probability\ of\ \text{failure}}$$

a. Let's say that the probability of success of Patriots in the upcoming super bowl is .8, what is the odd of this success?

Logistic Regression: Probability-Odds Pattern in-class exercise

11

p	odds
.001	.001001
.01	.010101
.15	.1764706
.2	.25
.25	.3333333
.3	.4285714
.35	.5384616
.4	.6666667
.45	.8181818
.5	1
.55	1.222222
.6	1.5
.65	1.857143
.7	2.333333
.75	3
.8	4
.85	5.666667
.9	9
.999	999
.9999	9999

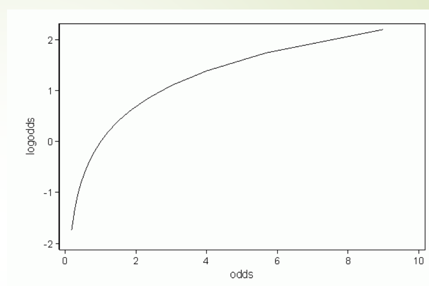


Q: Check the table and graph. What relationship between probability and odds can you find ?

Logistic Regression: Probability-Odds-Log Odds Pattern in-class exercise

12

p	odds	logodds
.001	.001001	-6.906755
.01	.010101	-4.59512
.15	.1764706	-1.734601
.2	.25	-1.386294
.25	.3333333	-1.098612
.3	.4285714	-.8472978
.35	.5384616	-.6190392
.4	.6666667	-.4054651
.45	.8181818	-.2006707
.5	1	0
.55	1.222222	.2006707
.6	1.5	.4054651
.65	1.857143	.6190392
.7	2.333333	.8472978
.75	3	1.098612
.8	4	1.386294
.85	5.666667	1.734601
.9	9	2.197225
.999	999	6.906755
.9999	9999	9.21024



Q: Check the table and graph. What relationship among probability and odds and log odds can you find ?

Log base: e

13

Logistic Regression: Probability->Odds->Log Odds (in class exercise)

- Why bother to do the transformation from probability to log odds?
- What is this transformation called?

14

Logistic Regression: Probability->Odds->Log Odds (answer keys)

- Why bother do the transformation from probability to log odds?
  - Difficult to model a variable which has restricted range, eg. Probability
  - Get around the restricted range problem: [0, 1] → $[-\infty, +\infty]$
  - log odds is one of the easiest to understand and interpret.
- What is this transformation called?
  - logit transformation

15

Logistic regression: definition

Definition comes naturally as:

- models the logit-transformed probability as a linear relationship with the predictors/attributes/features.
- allows us to establish a relationship between a binary outcome variable and a group of predictors/attributes/features.

16

Logistic regression: in-class exercise

- Consider a two-Category outcome (Y) probability space, where:
  - $p(\, y_1\,) = p$
  - $p(\, y_2\,) = 1 - p = q$

  Can express probability of y_1 as:

	notation	range equivalents		
standard probability	p	0	0.5	1
odds	p / q	?	?	?
log odds (logit)	log(p / q)	?	?	?

17

Logistic regression: in-class exercise

Answer:

	notation	range equivalents		
standard probability	p	0	0.5	1
odds	p / q	0	1	+ ∞
log odds (logit)	log(p / q)	- ∞	0	+ ∞

18

Logistic regression functions:

$$\text{logistic function } p = \frac{e^z}{1+e^z} = \frac{1}{1+e^{-z}} \quad \text{where } \frac{p}{1-p} = e^z$$

$$\text{logit function} \quad z = \log\left(\frac{p}{1-p}\right)$$

Logistic expression of Y (0/1) with the probability (p) of Y to be 1 on X estimates the parameter values of β via this equation:

$$Y\ (0/1) = \text{logit}(p) = \log(p/(1-p)) = \beta_0 + \beta_1 * \mathbf{x_1} + \ldots + \beta_k * \mathbf{x_k}$$

In terms of probabilities, the equation above is translated into

$$p(x) = \exp(\beta_0 + \beta_1 * \mathbf{x_1} + \ldots + \beta_k * \mathbf{x_k}) / (1 + \exp(\beta_0 + \beta_1 * \mathbf{x_1} + \ldots + \beta_k * \mathbf{x_k})).$$

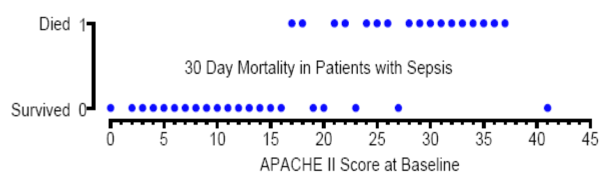
https://stats.idre.ucla.edu/other/mult-pkg/faq/general/faq-how-do-i-interpret-odds-ratios-in-logistic-regression/

19

Logistic regression: Example with one attribute

a) Example: APACHE II Score and Mortality in Sepsis

The following figure shows 30 day mortality in a sample of septic patients as a function of their baseline APACHE II Score. Patients are coded as 1 or 0 depending on whether they are dead or alive in 30 days, respectively.

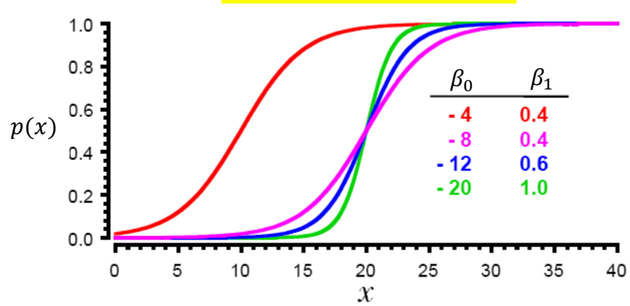


X: APACHEII Score at Baseline   Y: Died (1) or Survived (0)

20

Logistic regression: Example with one attribute in-class exercise

$p(x) = \exp(\beta_0 + \beta_1 x) / (1 + \exp(\beta_0 + \beta_1 x))$



$When\ x = -\beta_0/\beta_1,\quad \beta_0 + \beta_1 x = 0$ and hence $p(x) = 1/(1+1) = 0.5$

Midpoint: $P(x) = \exp[\beta_0 + \beta_1 * (-\beta_0/\beta_1)]/\{1 + \exp[\beta_0 + \beta_1 * (-\beta_0/\beta_1)]\}$
$= \exp(0)/[1+\exp(0)] = 1/2 = 0.5$

Sigmoid curve (S-curve)

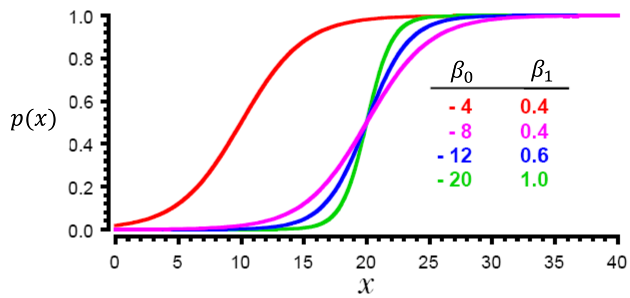
- β_0 controls location of midpoint
- β_1 controls slope of rise

Where is the x that points to the midpoint of red, purple, blue and green S-curve?

21

Logistic regression: Example with one attribute in-class exercise

$p(x) = \exp(\beta_0 + \beta_1 x) / (1 + \exp(\beta_0 + \beta_1 x))$



When $x = -\beta_0/\beta_1$, $\beta_0 + \beta_1 x = 0$ and hence $p(x) = 1/(1+1) = 0.5$

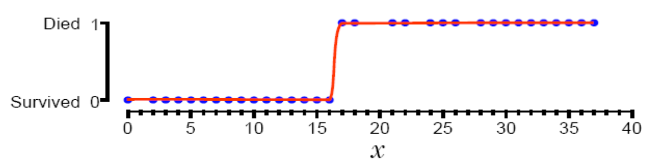
Where is the x that points to the midpoint of red, purple, blue and green S-curve?

Answer: Red: x = - (-4)/0.4 = 10
Pink: x = - (-8)/0.4= 20
Blue: x= - (-12)/0.6= 20
Green: x= - (-20)/1= 20

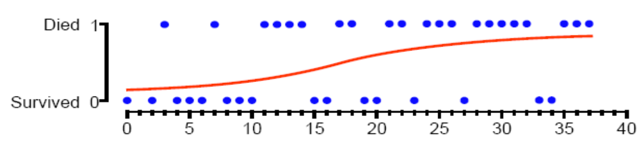
22

Logistic regression: Example with one attribute- - in class exercise

Data that has a sharp survival cut off point between patients who live or die should have a large value of β.



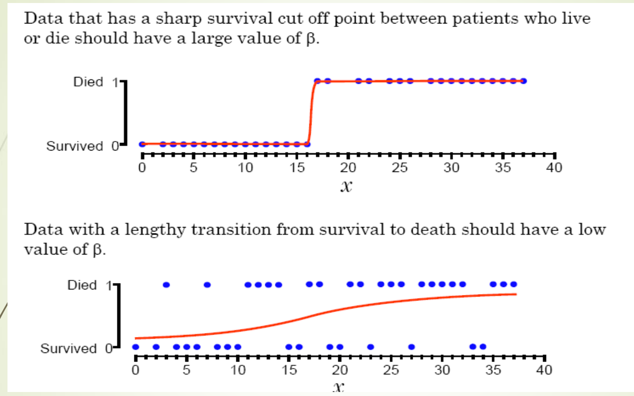
Data with a lengthy transition from survival to death should have a low value of β.



Q: From which graph can we find a clear cut-off score of X (APACHE II score) that can tell patients who live or die after 30-day admission to ICU?

23

Logistic regression: Example with one attribute-- in class exercise



Answer: Top graph: X (APACHE II score) <= 16 (Survived); X > 16 (Died), we see a clear cut off point;
Bottom graph: X ranges from 4 to 38 for Died; 0-34 for Survived; no clear cutoff point

Classification Evaluation

Classification Evaluation

- Classification error and accuracy: Creating a confusing matrix (a.k.a, crosstab/cross tabulation, contingency table)
  - For any data set we use to test the classification model on, we can build a **confusion matrix, e.g. for binary classification.**

  e.g., From logistic regression, we have

		Target/True	
		Y=1	Y=0
Predicted	$\hat{Y}$=1	140	17
	$\hat{Y}$=0	20	54

Classification Error = (20+7)/(140+54+20+17)= 37/231
Classification Accuracy = 1- Error = 194/231

Classification Evaluation

- Entries in a **confusion matrix** have names: e.g. for **binary** classification,

		Target/True	
		Y=1	Y=0
Predicted	$\hat{Y}$=1	TP	FP
	$\hat{Y}$=0	FN	TN

  - TP: True Positive (counts)
  - FP: False Positive (counts)
  - FN: False Negative (counts)
  - TN: True Negative (counts)

Classification Evaluation

- **Sensitivity (a.k.a. recall)** $SENS = \frac{TP}{TP + FN}$
- Specificity $SPEC = \frac{TN}{TN + FP}$
- **Positive predictive value(PPV) (a.k.a precision):** $PPV = \frac{TP}{TP + FP}$
- Negative predictive value (NPV) $NPV = \frac{TN}{TN + FN}$
- False Positive Rate: FPR (1- Specificity) FPR = 1- SPEC

Classification Evaluation:

Confusion matrix: Row and column quantities:

- Sensitivity (SENS)
- Specificity (SPEC)
- Positive predictive value (PPV)
- Negative predictive value (NPV)

	Y=1	Y=0	
$\hat{Y}=1$	140	10	$PPV = 140/150$
$\hat{Y}=0$	20	180	$NPV = 180/200$
	$SENS=140/160$ $SPEC=180/190$		

Receiver Operating Characteristic (ROC)

30

ROC: Brief History

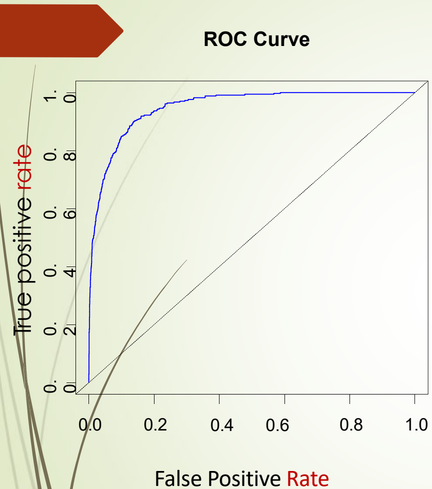
- From “Signal Detection Theory”: WWII for radar image analysis

  eg. Enemy target or A friendly ship

- Until 1970’s: recognized as useful for medical tests

- Now: become more generic evaluation for classification methods (Binary!)

ROC: True Positive vs. False Positive **rates**



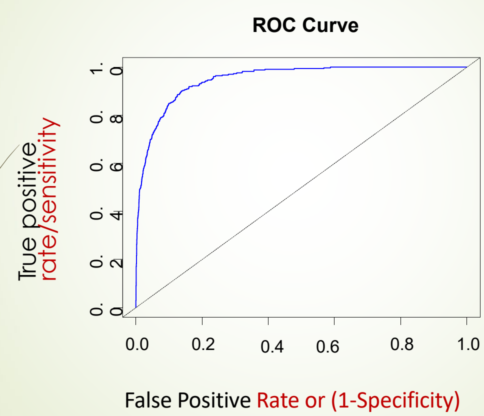
		Target/True	
		Y=1	Y=0
Predicted	$\hat{Y}$=1	TP (counts)	FP (counts)
	$\hat{Y}$=0	FN (counts)	TN (counts)

The *ROC plot* displays True & False positive rates simultaneously.

True Positive Rate= Sensitivity
False positive Rate= 1- Specificity

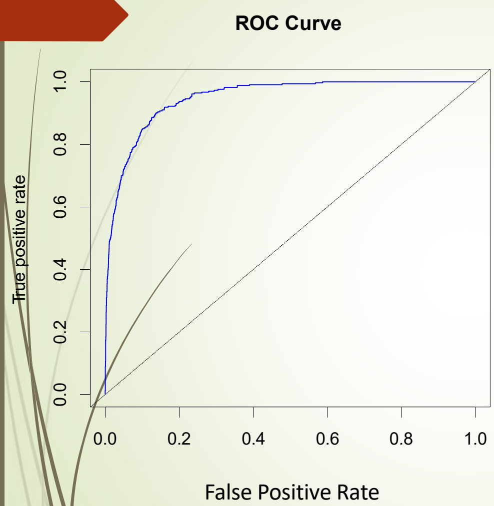
ROC: True Positive vs. False Positive **rates**

True Positive Rate= Sensitivity
False positive Rate= 1- Specificity



The *ROC plot* displays Sensitivity & Specificity simultaneously.

ROC: 45° line/dotted line vs ideal ROC curve



The dotted line/ 45° line represents the "no information" classifier:

The ideal ROC curve(better classifier) hugs the top left corner:

indicating a high true positive rate and a low false positive rate.

How to generate the ROC curve? (see next illustrative example)

ROC: Working with an example

New measure	Disease	Non-Disease
5 or less	18	1
5.1 - 7	7	17
7.1 - 9	4	36
9 or more	3	39
Totals:	32	93

ROC: Working with an example

Set a cut-off score at 5 for disease/non-disease

New measure	Disease	Non-Disease
5 or less	18	1
5.1 - 7	7	17
7.1 - 9	4	36
9 or more	3	39
Totals:	32	93

Cut point	Disease	Non-Disease
5 or less	18	1
> 5	14	92
Totals:	32	93

Sensitivity =18/32= **0.56**

Specificity = 92/93= **0.99**

ROC: Working with an example (In-class exercise)

Set a cut-off score at 7 for disease/non-disease: less strigent

New measure	Disease	Non-Disease
5 or less	18	1
5.1 - 7	7	17
7.1 - 9	4	36
9 or more	3	39
Totals:	32	93

Cut point	Disease	Non-Disease
7 or less	25	18
> 7	7	75
Totals:	32	93

Sensitivity = **0.78** Specificity = **0.81**

ROC: Working with an example

Set a cut-off score at 9 for disease/non-disease

New measure	Disease	Non-Disease
5 or less	18	1
5.1 - 7	7	17
7.1 - 9	4	36
9 or more	3	39
Totals:	32	93

Cut point	Disease	Non-Disease
9 or less	?	?
> 9	?	?
Totals:	?	?

Sensitivity = **0.91** Specificity = **0.42**

ROC: Working with an example

Put the sensitivity and specificity values into a table

New measure	Disease	Non-Disease
5 or less	18	1
5.1 - 7	7	17
7.1 - 9	4	36
9 or more	3	39
Totals:	32	93

Cut point	Sensitivity	Specificity
5	0.56	0.99
7	0.78	0.81
9	0.91	0.42

What sensitivity and specificity relationship do you see?

ROC: Working with an example (In-class exercise)

What sensitivity and specificity relationship do you see? (Answer key)

- Improve the sensitivity by moving the cut point to a *higher* value, ie., you can make the criterion for a positive test *less* strict.
- You can improve the specificity by moving the cut point to a *lower* value—ie. you can make the criterion for a positive test *more* strict.
- There's tradeoff between sensitivity and specificity. You can change the definition of a positive test to improve one but the other will decline.

ROC: Working with an example

Plot ROC: Convert Specificity to False Positive Rate

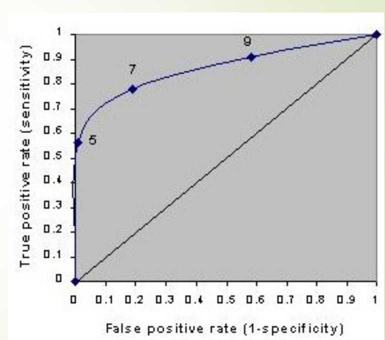
Cut point	Sensitivity	Specificity
5	0.56	0.99
7	0.78	0.81
9	0.91	0.42

Cut point	True Positive Rate	False Positive Rate
5	0.56	0.01
7	0.78	0.19
9	0.91	0.58

ROC: Working with an example

Plot ROC

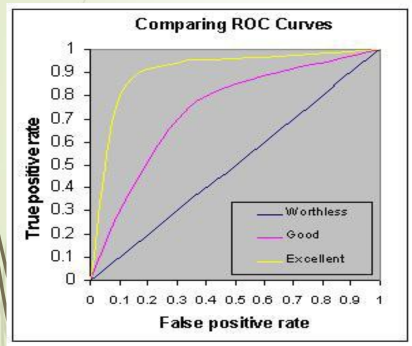
Cut point	True Positives	False Positives
5	0.56	0.01
7	0.78	0.19
9	0.91	0.58



ROC: Comparing multiple classifiers

- If you compare three or more classifiers, do the same steps to draw the ROC for each classifier.

Remember: The Y is binary, though! The curve represents each classifier.



e.g. three ROC curves : excellent, good, and worthless classifiers plotted on the same graph.

What if Y has more than two categories?

(Remember: you can always compare one category vs. the rest)

ROC: Area Under the ROC curve (AUC)

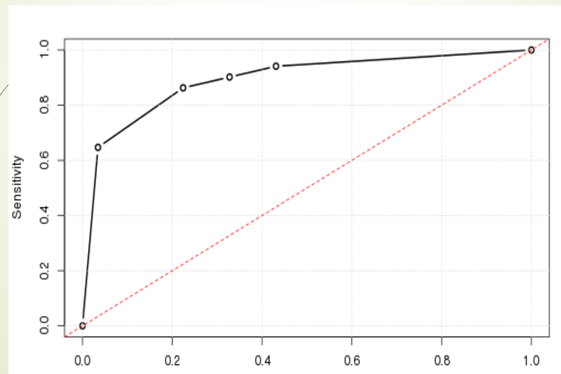
- AUC measures accuracy and discrimination: the ability of the classifier to correctly classify those with and without the disease in our example.
  - The higher the better; 0.5 is the base;
  - AUC range [0.5, 1]

  eg. A rough guide for classifying the accuracy of a diagnostic test is the traditional academic point system:

  - .90-1 = excellent (A)
  - .80-.90 = good (B)
  - .70-.80 = fair (C)
  - .60-.70 = poor (D)
  - .50-.60 = fail (F)

Area Under the Curve (AUC)

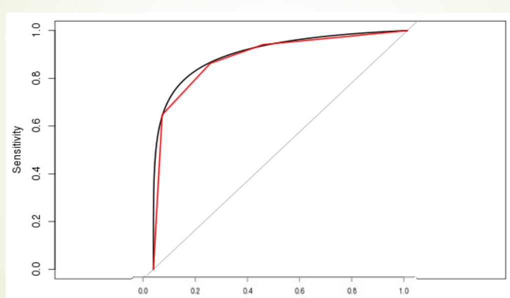
- Manually calculating the AUC (illustrative graph)
  -- We can calculate the area under the ROC curve, using the formula for the area of a trapezoid (a+b)h/2:



False Positive Rate

Area Under the Curve (AUC)

- **Smoothed ROC:** integrals. When using normalized units, AUC is equal to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one. This can be seen as follows:



False Positive Rate

$$A = \int_{\infty}^{-\infty} \mathrm{TPR}(T)\mathrm{FPR}'(T)\,dT = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} I(T' > T) f_1(T') f_0(T)\,dT'\,dT = P(X_1 > X_0)$$

where X_1 is the score for a positive instance and X_0 is the score for a negative instance.

R : Area Under the Curve (AUC)

- R: Use The pROC package. It can smooth the ROC estimate and calculate an AUC estimate based on the smoothed ROC.

ROC: Summary

- The *ROC plot* displays True & False positive rates (or sensitivity and specificity) simultaneously; shows the tradeoff between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity).
- The closer the curve follows the left-hand border and then the top border (ie., the top left corner) of the ROC space, the more accurate the classifier; the closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the classifier.
- ROC is useful for comparing different classifiers, as they take into account different possible thresholds.
- *AUC* (*area under the ROC curve)* is used to summarize the overall performance. The higher *AUC the better the Classifier*

49

R: Run logistic regression, confusion table and ROC/AUC in Rstudio using UCLA admission data

Let's go through the instruction file "R_logistic_CF_ROC&AUC_LS9.docx" posted with LS9 slides at myCourses.

50

Next time

- CART: **CART**: Classification **a**nd **R**egression Trees (Decision Trees)