



# Week 2 Video 3

## Diagnostic Metrics

# Different Methods, Different Measures



- Today we'll continue our focus on classifiers
- Later this week we'll discuss regressors
  
- And other methods will get worked in later in the course

# Last class

---

- We discussed accuracy and Kappa
- Today, we'll discuss additional metrics for assessing classifier goodness

# ROC



- Receiver-Operating Characteristic Curve

# ROC



- You are predicting something which has two values
  - ▣ Correct/Incorrect
  - ▣ Gaming the System/not Gaming the System
  - ▣ Dropout/Not Dropout

# ROC



- Your prediction model outputs a probability or other real value
- How good is your prediction model?

# Example

PREDICTION	TRUTH
0.1	0
0.7	1
0.44	0
0.4	0
0.8	1
0.55	0
0.2	0
0.1	0
0.09	0
0.19	0
0.51	1
0.14	0
0.95	1
0.3	0

# ROC



- Take any number and use it as a cut-off
- Some number of predictions (maybe 0) will then be classified as 1's
- The rest (maybe 0) will be classified as 0's



# Threshold = 0.5

PREDICTION	TRUTH
0.1	0
0.7	1
0.44	0
0.4	0
0.8	1
0.55	0
0.2	0
0.1	0
0.09	0
0.19	0
0.51	1
0.14	0
0.95	1
0.3	0

# Threshold = 0.6

PREDICTION	TRUTH
0.1	0
0.7	1
0.44	0
0.4	0
0.8	1
0.55	0
0.2	0
0.1	0
0.09	0
0.19	0
0.51	1
0.14	0
0.95	1
0.3	0

# Four possibilities

---

- True positive
- False positive
- True negative
- False negative

# Threshold = 0.6

PREDICTION	TRUTH	
0.1	0	TRUE NEGATIVE
0.7	1	TRUE POSITIVE
0.44	0	TRUE NEGATIVE
0.4	0	TRUE NEGATIVE
0.8	1	TRUE POSITIVE
0.55	0	TRUE NEGATIVE
0.2	0	TRUE NEGATIVE
0.1	0	TRUE NEGATIVE
0.09	0	TRUE NEGATIVE
0.19	0	TRUE NEGATIVE
0.51	1	FALSE NEGATIVE
0.14	0	TRUE NEGATIVE
0.95	1	TRUE POSITIVE
0.3	0	TRUE NEGATIVE

# Threshold = 0.5

PREDICTION	TRUTH	
0.1	0	TRUE NEGATIVE
0.7	1	TRUE POSITIVE
0.44	0	TRUE NEGATIVE
0.4	0	TRUE NEGATIVE
0.8	1	TRUE POSITIVE
0.55	0	FALSE POSITIVE
0.2	0	TRUE NEGATIVE
0.1	0	TRUE NEGATIVE
0.09	0	TRUE NEGATIVE
0.19	0	TRUE NEGATIVE
0.51	1	TRUE POSITIVE
0.14	0	TRUE NEGATIVE
0.95	1	TRUE POSITIVE
0.3	0	TRUE NEGATIVE

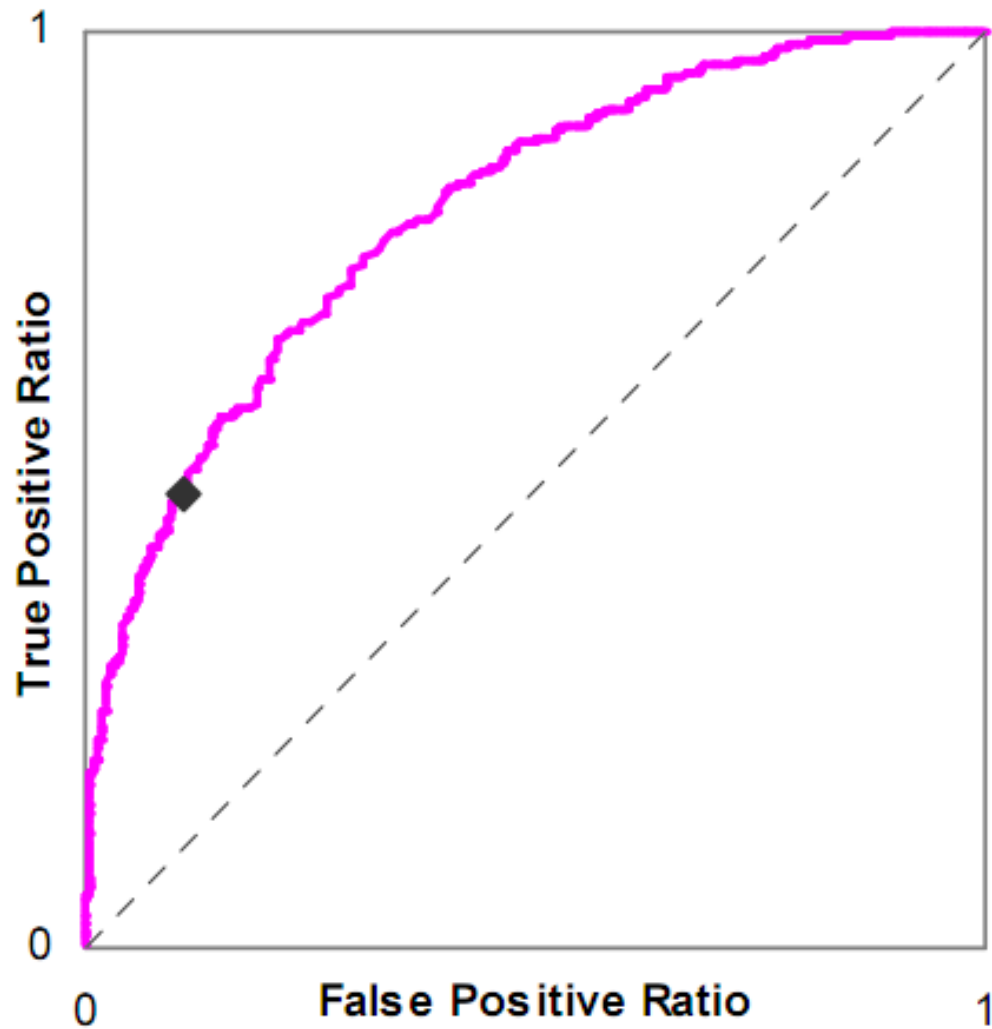
# Threshold = 0.99

PREDICTION	TRUTH	
0.1	0	TRUE NEGATIVE
0.7	1	FALSE NEGATIVE
0.44	0	TRUE NEGATIVE
0.4	0	TRUE NEGATIVE
0.8	1	FALSE NEGATIVE
0.55	0	TRUE NEGATIVE
0.2	0	TRUE NEGATIVE
0.1	0	TRUE NEGATIVE
0.09	0	TRUE NEGATIVE
0.19	0	TRUE NEGATIVE
0.51	1	FALSE NEGATIVE
0.14	0	TRUE NEGATIVE
0.95	1	FALSE NEGATIVE
0.3	0	TRUE NEGATIVE

# ROC curve

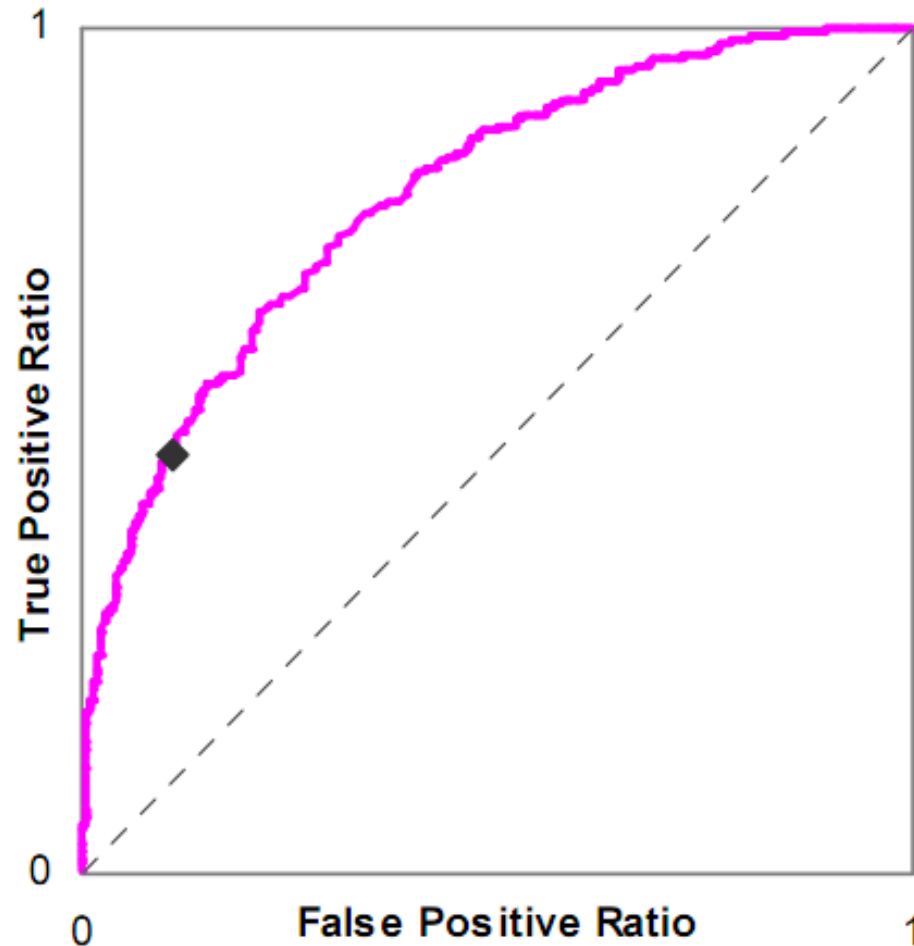
- X axis = Percent false positives (versus true negatives)
  - ▣ False positives to the right
- Y axis = Percent true positives (versus false negatives)
  - ▣ True positives going up

# Example

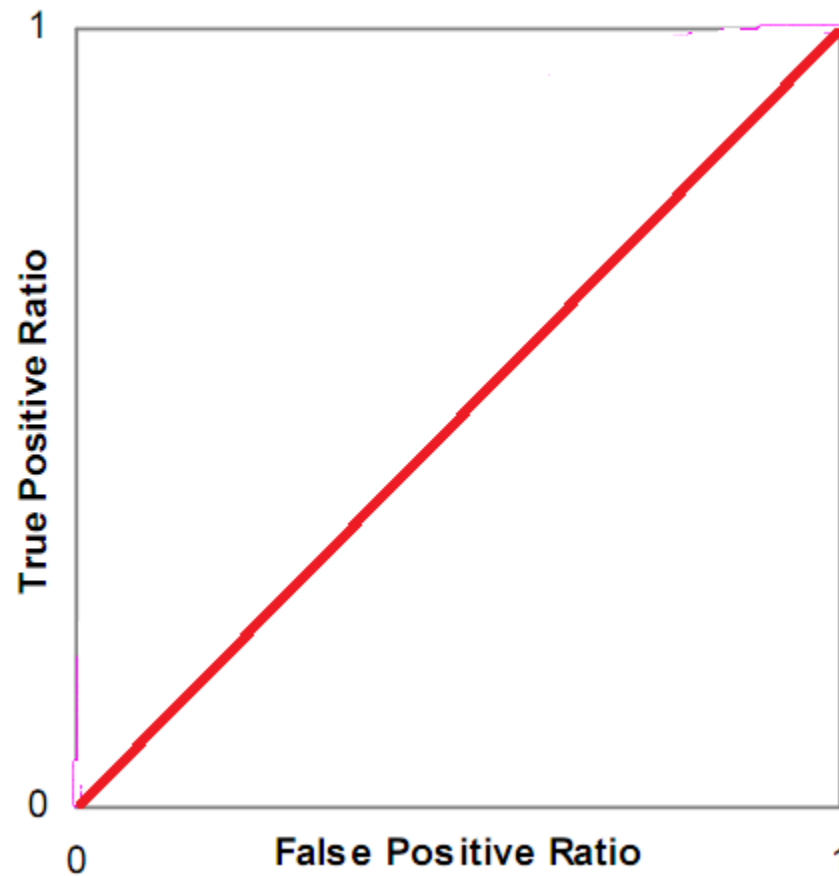




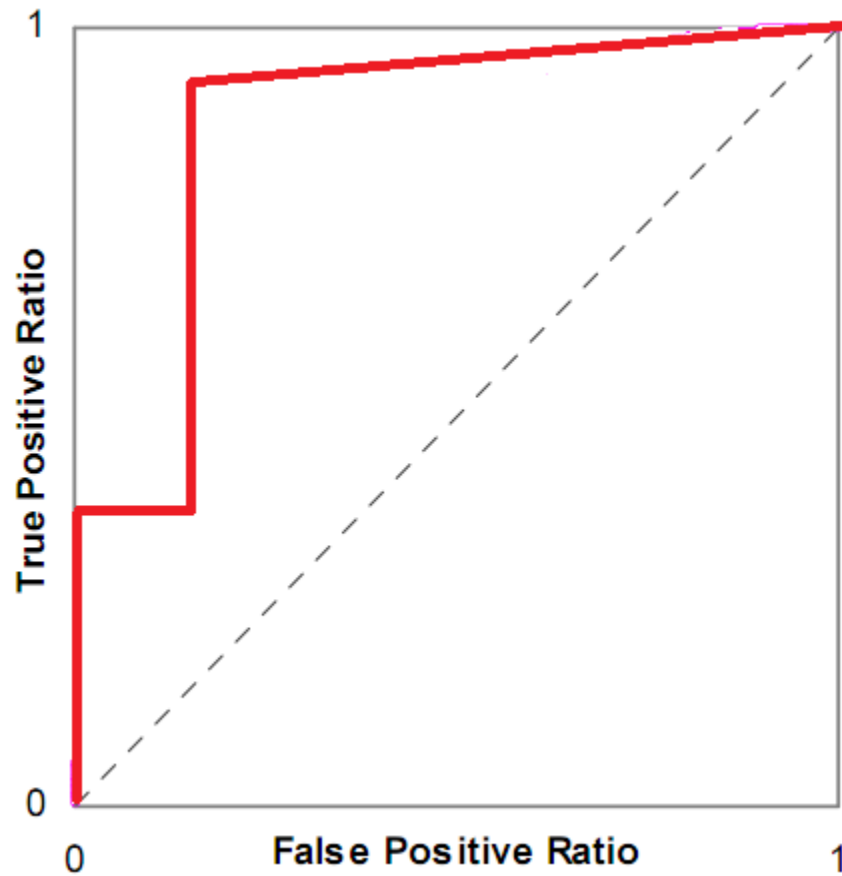
# Is this a good model or a bad model?



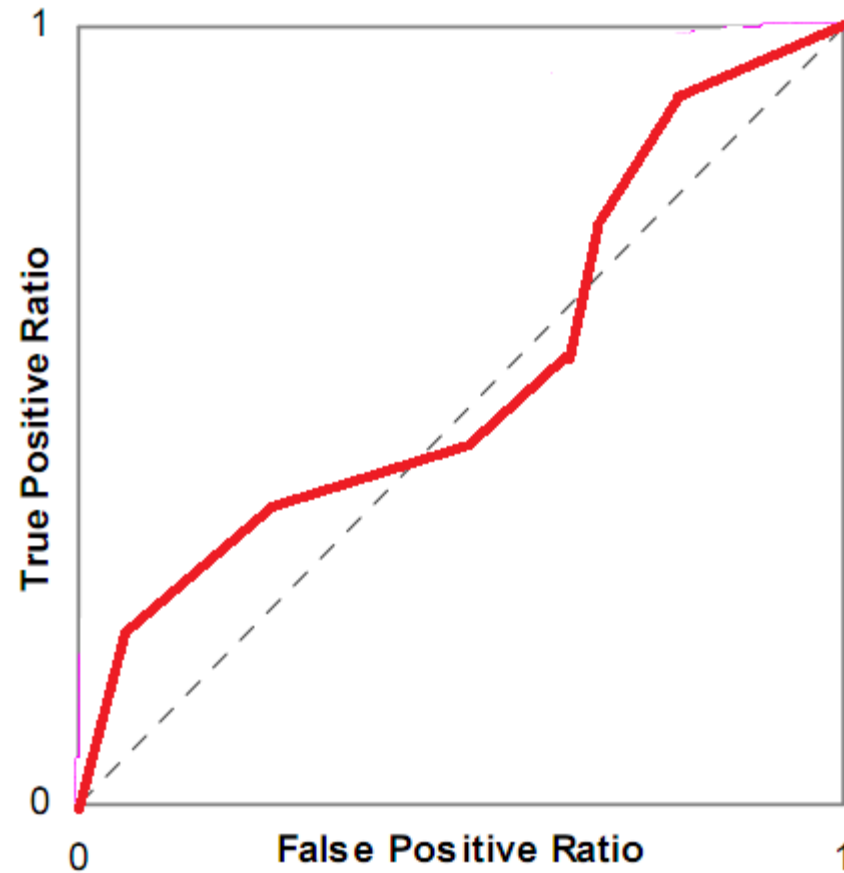
# Chance model



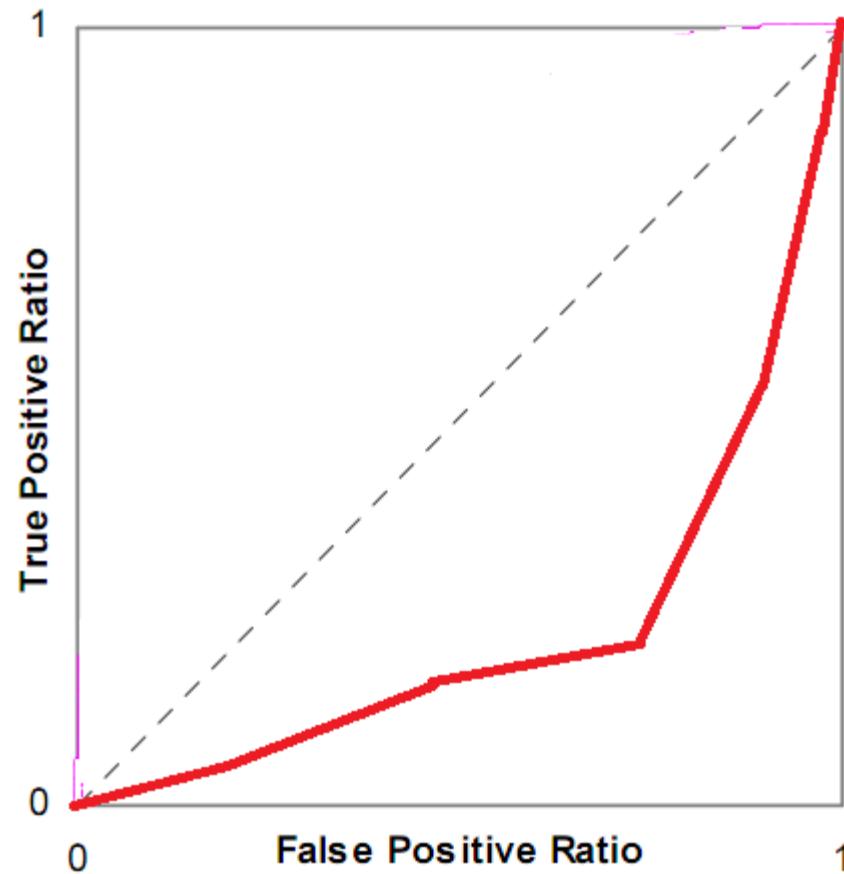
# Good model (but note stair steps)



# Poor model



# So bad it's good



# A': A close relative of ROC

- The probability that if the model is given an example from each category, it will accurately identify which is which

# A'

- Is mathematically equivalent to the Wilcoxon statistic (Hanley & McNeil, 1982)
- Useful result, because it means that you can compute statistical tests for
  - ▣ Whether two A' values are significantly different
    - Same data set or different data sets!
  - ▣ Whether an A' value is significantly different than chance

# Notes

- Not really a good way (yet) to compute  $A'$  for 3 or more categories
  - ▣ There are methods, but the semantics change somewhat



## Comparing Two Models (**ANY** two models)

$$Z = \frac{A'_1 - A'_2}{\sqrt{\text{SE}(A'_1)^2 + \text{SE}(A'_2)^2}}$$

# Comparing Model to Chance

$$Z = \frac{A'_1 - 0.5}{\sqrt{\text{SE}(A'_1)^2 + 0}}$$

# Equations

$$D_p = (n_p - 1) \left( \frac{A'}{2 - A'} - A'^2 \right) \quad D_n = (n_n - 1) \left( \frac{2 * A'^2}{1 + A'} - A'^2 \right)$$

$$SE(A') = \sqrt{\frac{A'(1 - A') + D_p + D_n}{n_p * n_n}}$$

# Complication

- This test assumes independence
- If you have data for multiple students, you usually should compute  $A'$  and significance for each student and then integrate across students (Baker et al., 2008)
  - ▣ There are reasons why you might not want to compute  $A'$  within-student, for example if there is no intra-student variance
  - ▣ If you don't do this, don't do a statistical test

# A'

- Closely mathematically approximates the area under the ROC curve, called AUC (Hanley & McNeil, 1982)
- The semantics of  $A'$  are easier to understand, but it is often calculated as AUC
  - ▣ Though at this moment, I can't say I'm sure why –  $A'$  actually seems mathematically easier

# More Caution

- The implementations of AUC are buggy in **all** major statistical packages that I've looked at
- Special cases get messed up
- There is A' code on my webpage that is more reliable for known special cases
  - ▣ Computes as Wilcoxon rather than the faster but more mathematically difficult integral calculus

# A' and Kappa



# A' and Kappa

## □ A'

- more difficult to compute
- only works for two categories (without complicated extensions)
- meaning is invariant across data sets ( $A'=0.6$  is always better than  $A'=0.55$ )
- very easy to interpret statistically



# A'

- A' values are almost always higher than Kappa values
- A' takes confidence into account

# Precision and Recall

□ Precision = 
$$\frac{TP}{TP + FP}$$

□ Recall = 
$$\frac{TP}{TP + FN}$$

# What do these mean?

- Precision = The probability that a data point classified as true is actually true
- Recall = The probability that a data point that is actually true is classified as true

# Next lecture

---

- Metrics for regressors