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

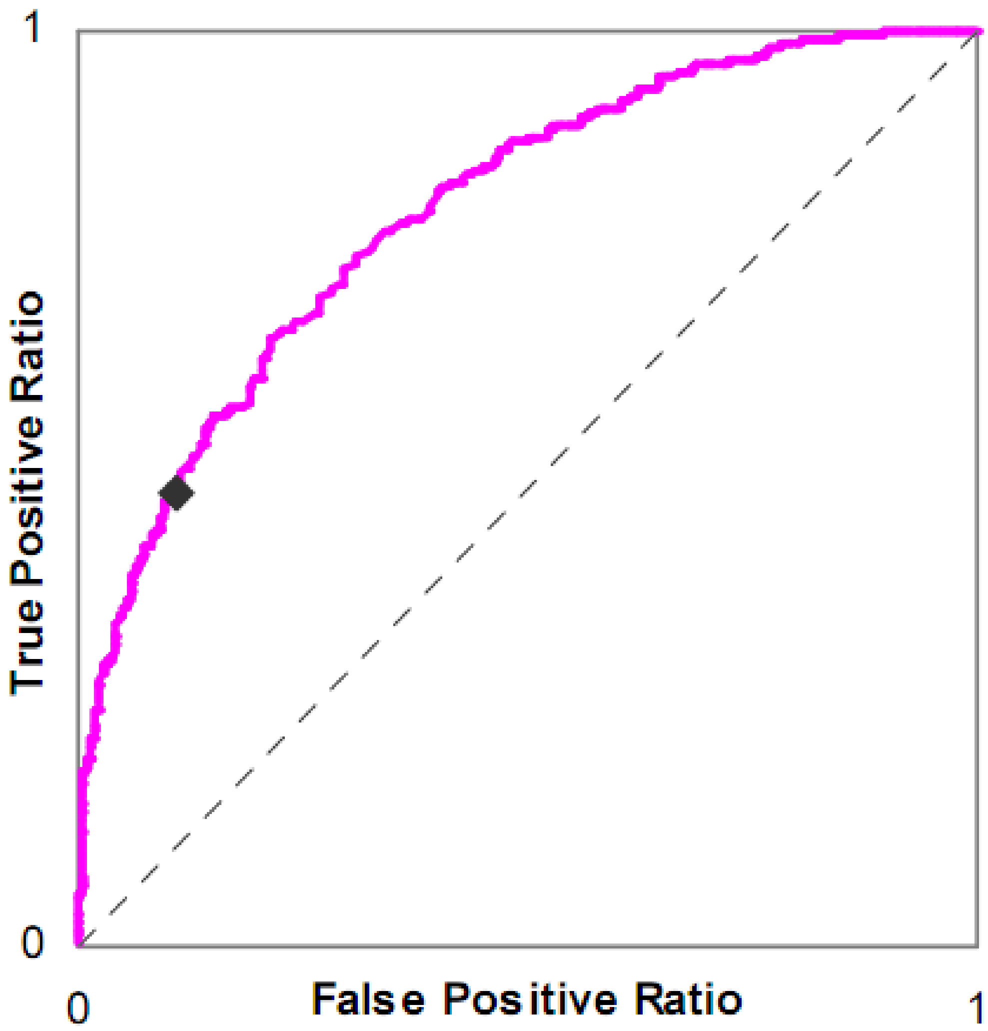
Threshold = 0.99

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

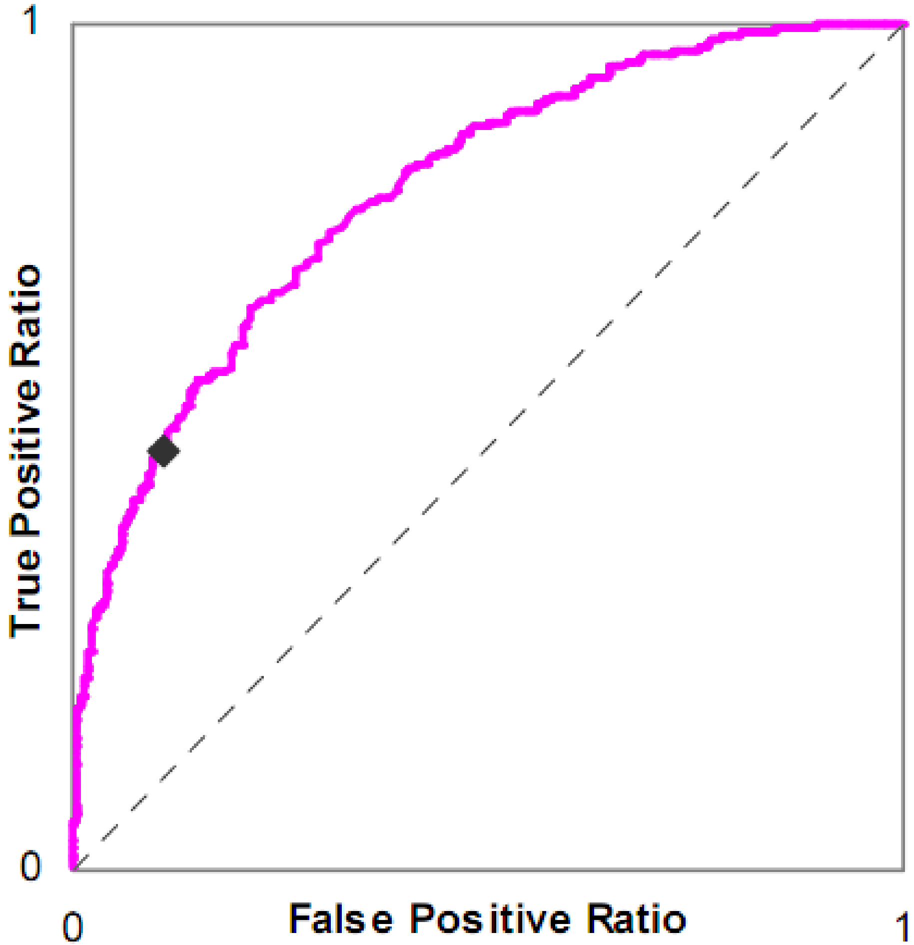
ROC curve

- 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

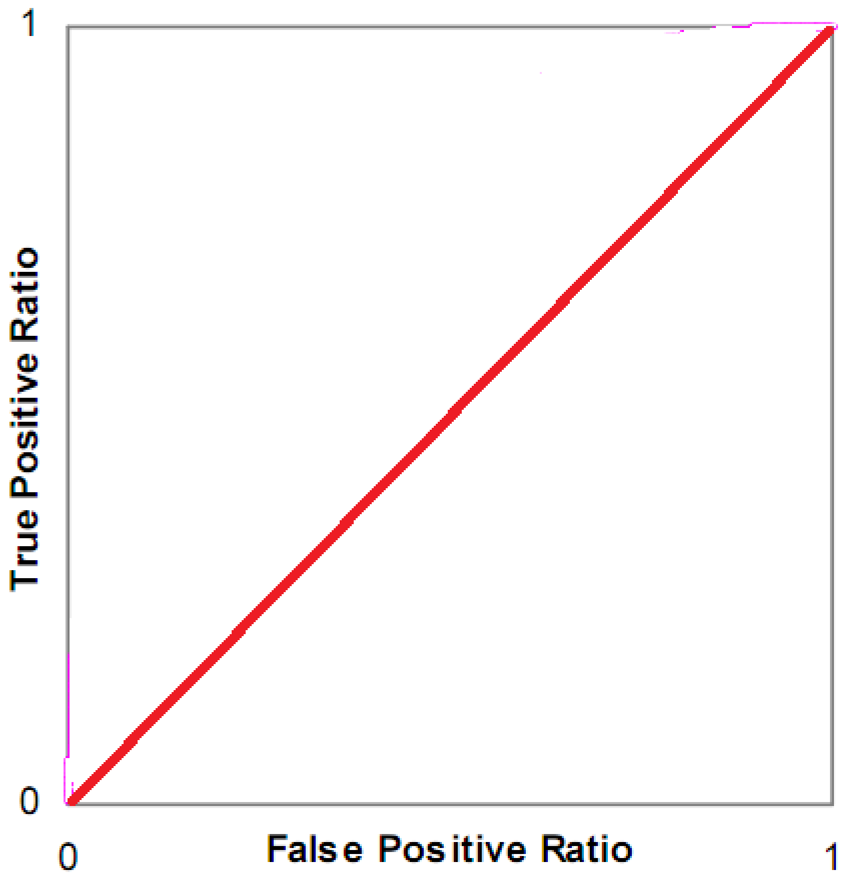
Example



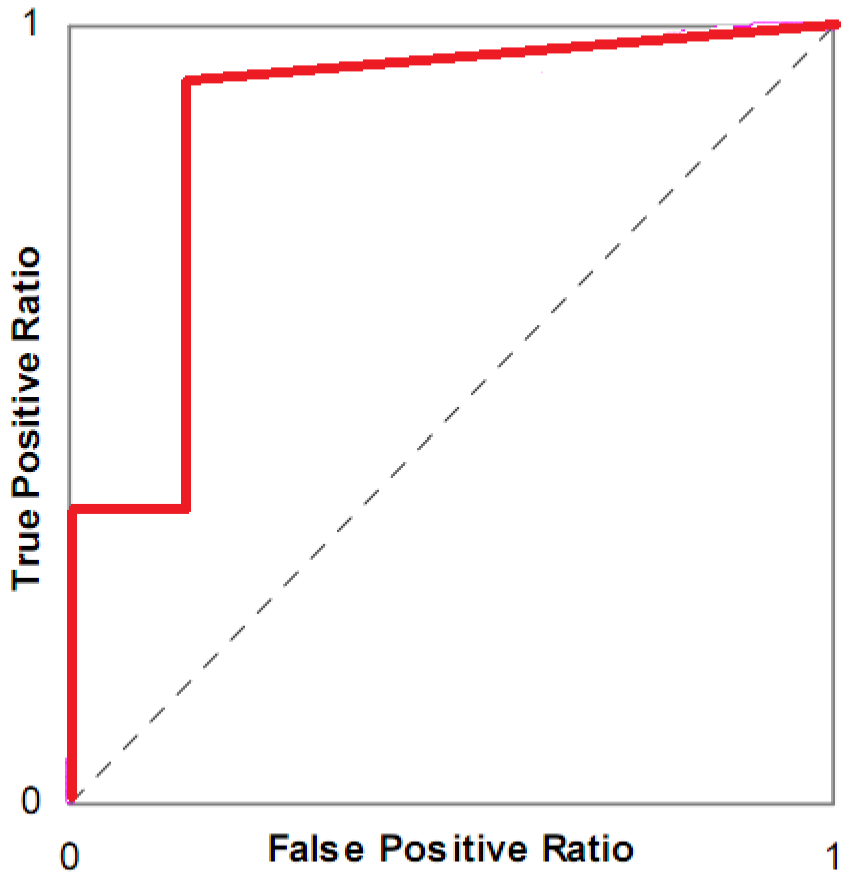
Is this a good model or a bad model?



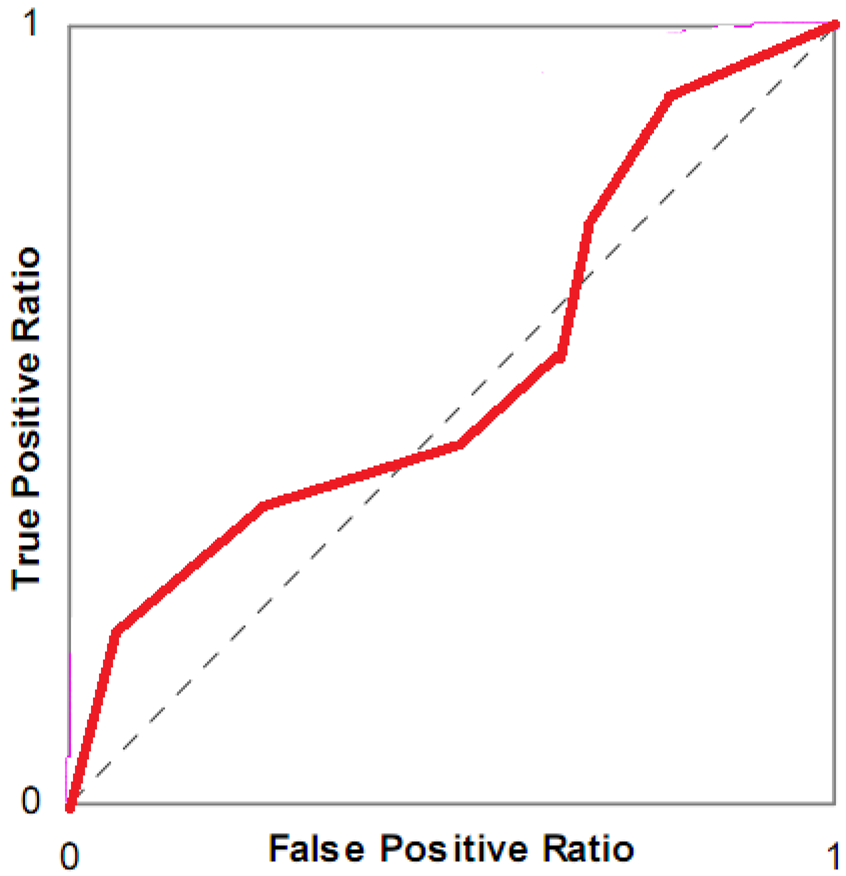
Chance model



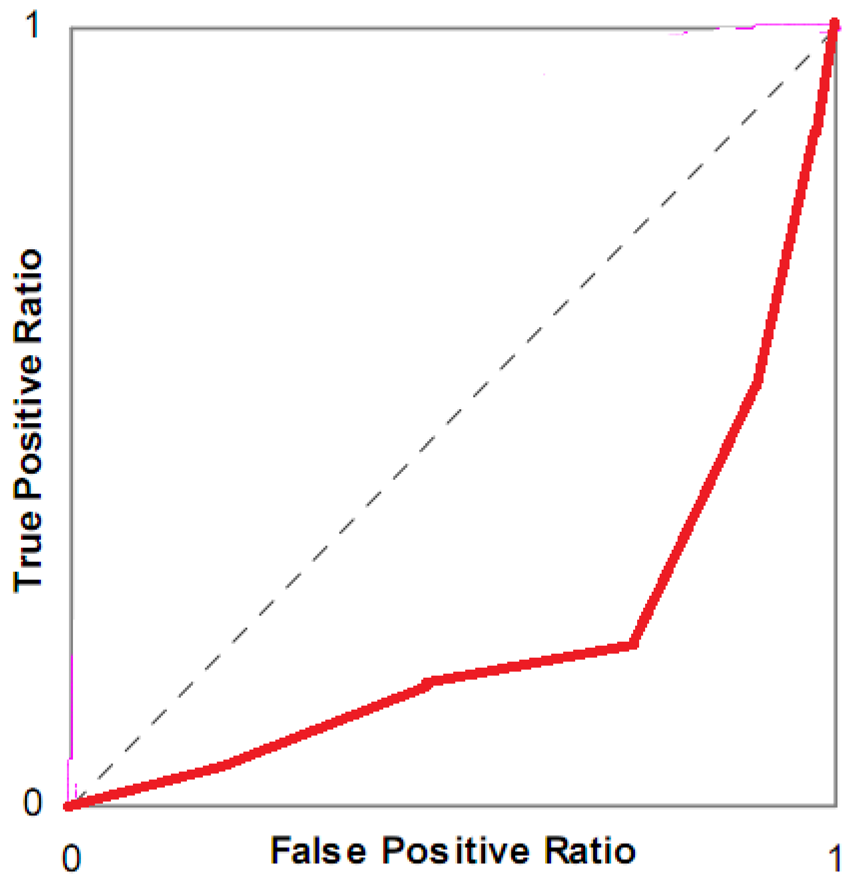
Good model (but note stair steps)



Poor model



So bad it's good



AUC ROC

- Also called AUC, or A'
- The area under the ROC curve

AUC

- Is mathematically equivalent to the Wilcoxon statistic (Hanley & McNeil, 1982)
  - The probability that if the model is given an example from each category, it will accurately identify which is which

AUC

- Equivalence to Wilcoxon is useful

- It means that you can compute statistical tests for
  - Whether two AUC values are significantly different
    - Same data set or different data sets!
  - Whether an AUC value is significantly different than chance

Notes

- Not really a good way to compute AUC for 3 or more categories
  - There are methods, but the semantics change somewhat

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

$$Z = \frac{AUC_1 - AUC_2}{\sqrt{SE(AUC_1)^2 + SE(AUC_2)^2}}$$

Comparing Model to Chance

$$Z = \frac{AUC_1 - 0.5}{\sqrt{SE(AUC_1)^2 + 0}}$$

Equations

$$D_p = (n_p - 1)(\frac{AUC}{2 - AUC} - AUC^2)$$

$$D_n = (n_n - 1)(\frac{2 * \ AUC^2}{1 + AUC} - AUC^2)$$

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

Complication

- This test assumes independence
- If you have data for multiple students, you usually should compute AUC and significance for each student and then integrate across students (Baker et al., 2008)
  - There are reasons why you might not want to compute AUC within-student, for example if there is no intra-student variance (see discussion in Pelanek, 2017)
  - If you don't do this, don't do a statistical test

More Caution

- The implementations of AUC remain buggy in many data mining and statistical packages in 2018

- But it works in sci-kit learn
- And there is a correct package for r called auctestr
- If you use other tools, see my webpage for a command-line and GUI implementation of AUC
  http://www.upenn.edu/learninganalytics/ryanbaker/edmtools.html

AUC and Kappa

AUC and Kappa

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

AUC

- AUC values are almost always higher than Kappa values

- AUC 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

Terminology

- FP = False Positive = Type 1 error
- FN = False Negative = Type 2 error

Still active debate about these metrics

- (Jeni et al., 2013) finds evidence that AUC is more robust to skewed distributions than Kappa and also several other metrics
- (Dhanani et al., 2014) finds evidence that models selected with RMSE (which we'll talk about next time) come closer to true parameter values than AUC
- (Pelanek, 2017) argues that AUC only pays attention to relative differences between models and that absolute differences matter too

Next lecture

- Metrics for regressors