Week 2 Video 2

Diagnostic Metrics, Part 1

Different Methods, Different Measures

- Today we'll focus on metrics for classifiers
- Later this week we'll discuss metrics for regressors

- And metrics for other methods will be discussed later in the course

Metrics for Classifiers

Accuracy

Accuracy

- One of the easiest measures of model goodness is ***accuracy***
- Also called ***agreement***, when measuring inter-rater reliability

$$\frac{\text{\# of agreements}}{\text{total number of codes/assessments}}$$

Accuracy

- There is general agreement across fields that accuracy is not a good metric

Accuracy

- Let's say that my new Kindergarten Failure Detector achieves 92% accuracy

- Good, right?

Non-even assignment to categories

- Accuracy does poorly when there is non-even assignment to categories
  - Which is almost always the case

- Imagine an extreme case
  - 92% of students pass Kindergarten
  - My detector always says ***PASS***

- Accuracy of 92%
- But essentially no information

Kappa

Kappa

$$\frac{\text{(Agreement – Expected Agreement)}}{\text{(1 – Expected Agreement)}}$$

Computing Kappa (Simple 2x2 example)

	Detector Off-Task	**Detector On-Task**
Data Off-Task	20	5
Data On-Task	15	60

Computing Kappa (Simple 2x2 example)

	Detector Off-Task	**Detector On-Task**
Data Off-Task	20	5
Data On-Task	15	60

- What is the percent agreement?

Computing Kappa (Simple 2x2 example)

	Detector Off-Task	**Detector On-Task**
Data Off-Task	20	5
Data On-Task	15	60

- What is the percent agreement?
  - 80%

Computing Kappa (Simple 2x2 example)

	Detector Off-Task	**Detector On-Task**
Data Off-Task	20	5
Data On-Task	15	60

- What is Data's expected frequency for on-task?

Computing Kappa (Simple 2x2 example)

	Detector Off-Task	**Detector On-Task**
Data Off-Task	20	5
Data On-Task	15	60

- What is Data's expected frequency for on-task?
  - *75%*

Computing Kappa (Simple 2x2 example)

	Detector Off-Task	**Detector On-Task**
Data Off-Task	20	5
Data On-Task	15	60

- What is Detector's expected frequency for on-task?

Computing Kappa (Simple 2x2 example)

	Detector Off-Task	**Detector On-Task**
Data Off-Task	20	5
Data On-Task	15	60

- What is Detector's expected frequency for on-task?
  - 65%

Computing Kappa (Simple 2x2 example)

	Detector Off-Task	**Detector On-Task**
Data Off-Task	20	5
Data On-Task	15	60

- What is the expected on-task agreement?

Computing Kappa (Simple 2x2 example)

	Detector Off-Task	**Detector On-Task**
Data Off-Task	20	5
Data On-Task	15	60

- What is the expected on-task agreement?
  - 0.65*0.75= 0.4875

Computing Kappa (Simple 2x2 example)

	Detector Off-Task	**Detector On-Task**
Data Off-Task	20	5
Data On-Task	15	60 (48.75)

- What is the expected on-task agreement?
  - 0.65*0.75= 0.4875

Computing Kappa (Simple 2x2 example)

	Detector Off-Task	**Detector On-Task**
Data Off-Task	20	5
Data On-Task	15	60 (48.75)

- What are Data and Detector's expected frequencies for off-task behavior?

Computing Kappa (Simple 2x2 example)

	Detector Off-Task	**Detector On-Task**
Data Off-Task	20	5
Data On-Task	15	60 (48.75)

- What are Data and Detector's expected frequencies for off-task behavior?
  - 25% and 35%

Computing Kappa (Simple 2x2 example)

	Detector Off-Task	**Detector On-Task**
Data Off-Task	20	5
Data On-Task	15	60 (48.75)

- What is the expected off-task agreement?

Computing Kappa (Simple 2x2 example)

	Detector Off-Task	**Detector On-Task**
Data Off-Task	20	5
Data On-Task	15	60 (48.75)

- What is the expected off-task agreement?
  - 0.25*0.35= 0.0875

Computing Kappa (Simple 2x2 example)

	Detector Off-Task	**Detector On-Task**
Data Off-Task	20 (8.75)	5
Data On-Task	15	60 (48.75)

- What is the expected off-task agreement?
  - 0.25*0.35= 0.0875

Computing Kappa (Simple 2x2 example)

	Detector Off-Task	**Detector On-Task**
Data Off-Task	20 (8.75)	5
Data On-Task	15	60 (48.75)

- What is the total expected agreement?

Computing Kappa (Simple 2x2 example)

	Detector Off-Task	**Detector On-Task**
Data Off-Task	20 (8.75)	5
Data On-Task	15	60 (48.75)

- What is the total expected agreement?
  - 0.4875+0.0875 = 0.575

Computing Kappa (Simple 2x2 example)

	Detector Off-Task	**Detector On-Task**
Data Off-Task	20 (8.75)	5
Data On-Task	15	60 (48.75)

- What is kappa?

Computing Kappa (Simple 2x2 example)

	Detector Off-Task	**Detector On-Task**
Data Off-Task	20 (8.75)	5
Data On-Task	15	60 (48.75)

- What is kappa?
  - (0.8 – 0.575) / (1-0.575)
  - 0.225/0.425
  - 0.529

So is that any good?

	Detector Off-Task	**Detector On-Task**
Data Off-Task	20 (8.75)	5
Data On-Task	15	60 (48.75)

- What is kappa?
  - (0.8 – 0.575) / (1-0.575)
  - 0.225/0.425
  - 0.529

Interpreting Kappa

- Kappa = 0
  - Agreement is at chance
- Kappa = 1
  - Agreement is perfect
- Kappa = -1
  - Agreement is perfectly inverse
- Kappa > 1
  - You messed up somewhere

Kappa<0

- This means your model is worse than chance

- Very rare to see unless you're using cross-validation
- Seen more commonly if you're using cross-validation
  - It means your model is junk

0<Kappa<1

- What's a good Kappa?

- There is no absolute standard

0<Kappa<1

- For data mined models,
  - Typically 0.3-0.5 is considered good enough to call the model better than chance and publishable
  - In affective computing, lower is still often OK

Why is there no standard?

- Because Kappa is scaled by the proportion of each category

- When one class is much more prevalent
  - Expected agreement is higher than
- If classes are evenly balanced

Because of this…

- Comparing Kappa values between two data sets, in a principled fashion, is highly difficult
  - It ***is*** OK to compare Kappa values within a data set
- A lot of work went into statistical methods for comparing Kappa values in the 1990s
- No real consensus
- Informally, you can compare two data sets if the proportions of each category are "similar"

Quiz

	Detector Insult during Collaboration	**Detector No Insult during Collaboration**
Data Insult	16	7
Data No Insult	8	19

- What is kappa?

  A: 0.645

  B: 0.502

  C: 0.700

  D: 0.398

Quiz

	Detector Academic Suspension	**Detector No Academic Suspension**
Data Suspension	1	2
Data No Suspension	4	141

- What is kappa?

  A: 0.240

  B: 0.947

  C: 0.959

  D: 0.007

Next lecture

- ROC curves
- A'
- Precision
- Recall