6.S085 Statistics for Research Projects **IAP 2014**

Lecture 8: January 30

Lecturer: Ramesh Sridharan and George Chen *Notes by: William Li*

Correction from last time: Kruskal-Wallis non-parametric ANOVA

Noise distribution needs to be the same across data points (need not be Gaussian)

H_0: median of all groups is the same

8.1 Machine Learning and Predictive Analytics

Code Blue Example

Recall: how many predicted "positive" / how many there were positive

False positive: how many predicted "positive" incorrectly

8.2 Supervised Learning vs. Unsupervised Learning

Supervised:

- data is labeled
- want to predict label for new point
- Example: email spam classification
  - Training data: emails labeled (manually as spam/notspam)
  - Test data: new emails (want to classify as spam/notspam)

Unsupervised:

- data is not labeled
- want to find structure in the data

Notes:

- Features matter a lot!
- When we can't predict well, generally it's because we don't have good features
- Good features give huge performance gains vs. specifics of what ML algorithm is used

8.3 Supervised Learning

Two approaches: one is probabilistic, the other is not

8.3.1 Probabilistic

Specify a probabilistic model that we assume data is actually generated from

Naive Bayes (probabilistic model) for email spam classification

Idea: specify how an email's features are are generated.

1. Flip a biased coin: If H, new email is spam. If T, new email is notspam
2. Generate each feature value independently, dependending on whether email is spam/ham

Notice model has parameters!

5 Numbers

1. Prob(email is spam)
2. Prob("asks for password" is True |spam), Prob("mentions Nigerian prince" |spam)
3. Prob("asks for password" is True |notspam), Prob("mentions Nigerian prince" |notspam)

Training: Estimate these parameters

Testing: Prob(spam|new email)

Use Bayes' Rule:

$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$

New email: (True, True, ?)

Want:

$P(\text{SPAM}|(\text{ TRUE, TRUE, ? }))$

$= \frac{P((\text{TRUE, TRUE, ?})|\text{SPAM})P(\text{SPAM})}{P((\text{TRUE,TRUE,?})|\text{SPAM})P(\text{SPAM})+P(|\text{HAM})P(\text{HAM})}$