COMP 330: Intro to Modeling 1

Chris Jermaine and Kia Teymourian

Rice University

What is a Model?

Many definitions!

Traditional statistical definition:

- A set of assumptions regarding the (stochastic) process that generated the data

More modern definition:

- An algorithm that can be used to generate an artifact explaining the data

What's the difference?

Why Do We Model?

- Real data are big, complex, difficult to understand
- A model is (hopefully!) compact, simple, comprehensible

Why Do We Model?

- Real data are big, complex, difficult to understand
- A model is (hopefully!) compact, simple, comprehensible

Just as important:

- Models can often be used to make predictions re future events

Statistical Modeling

Many (not all!) models rely on the idea of probability

- "the extent to which an event is likely to occur, measured by the ratio of the favorable cases to the whole number of cases possible"

Statistical Modeling

Many (not all!) models rely on the idea of probability

- "the extent to which an event is likely to occur, measured by the ratio of the favorable cases to the whole number of cases possible"

What about infinitely many possible events?

Then probability tends to zero

- Ex: the chance I jump exactly 3 feet
- Ex: the chance class ends at exactly 11A
- Ex: the chance it takes 5 hours to complete A2

Statistical Modeling

Many (not all!) models rely on the idea of probability

- "the extent to which an event is likely to occur, measured by the ratio of the favorable cases to the whole number of cases possible"

What about infinitely many possible events?

Then probability tends to zero

- Ex: the chance I jump exactly 3 feet
- Ex: the chance class ends at exactly 11A
- Ex: the chance it takes 5 hours to complete A2

Motivation for the idea of probability density

Probability Density

Probability density gets around this problem

- Measures the relative likelihood of an event—not absolute

Probability A2 takes 5 hours—nonsensical

But...

- Likelihood A2 takes 5 hours is $5X$ it takes 1 hour
- Sensical!

Probability Density Function

A PDF is a function that computes the relative probability of an event

Most famous: normal PDF

$$f_{\text{Normal}}(x|\mu,\sigma) = \sigma^{-1}(2\pi)^{-\frac{1}{2}}e^{\frac{1}{2}(x-\mu)^2\sigma^{-2}}$$

- A PDF can be used to calculate the probability of a range of events
- $\int_a^b f(x)dx$ is the probability we see a value in range a to b

Choosing a Model

I borrow an (incorrect but useful) proverb from Luay:

- "All models are wrong, but some are useful"

Remember:

- "A model is (hopefully!) compact, simple, comprehensible"
- We choose models to reduce, simplify, comprehend data
- Hopefully, without incurring inaccuracy!!

Example: Predicting Grade In Class

A student has completed 5/10 assignments

- Want to predict grade in class

First, devise a model

- Ex: assume $X_i \sim \textbf{Normal}(\mu, \sigma)$ (Why normal?)
- i is the identity of the assignment
- Note: X_i is a random variable controlling grade
- $f_{X_i}(x)$ gives relative likelihood X_i takes value x
- (or probability if X_i is discrete!)
- So $f_{X_i}(x) = f_{\text{Normal}}(x|\mu, \sigma)$

Example: Predicting Grade In Class

Scores so far: $\{89, 92, 78, 94, 88\}$

- Estimate mean $\mu = 88.2$, $\sigma^2 = 30.56$
- Thus $\sum_{i=6\ldots10} X_i \sim \textbf{Normal}(88.2 \times 5, (30.56 \times 5)^{0.5})$
- 95% confidence on sum: $882 \pm 2 \times 12.36$
- Or, 95% confidence on average: 88.2 ± 2.47

Another Example: Assignment Turn In

5/10 students have completed the assignment

168 hours (one week) to complete the assignment

- Want to predict how many have completed by 1 hour before due date
- X_i: number of hours after assignment student i turns in
- Assume $X_i \sim \textbf{Exponential}(\lambda)$

Another Example: Assignment Turn In

5/10 students have completed the assignment

168 hours (one week) to complete the assignment

- Want to predict how many have completed by 1 hour before due date
- X_i: number of hours after assignment student i turns in
- Assume $X_i \sim$ **Exponential**(λ)
- Exponential:

$$f_{Exp}(x|\lambda) = \lambda e^{-\lambda x}$$

- Memoryless!
- Means if waited t unites so far...
- $f_{Exp}(x|\lambda, x \geq t) = f_{Exp}(x - t|\lambda)$

Another Example: Assignment Turn In

Times so far at tick 100: $\{18, 22, 45, 49, 86\}$

- Know mean of exponential is λ^{-1}
- In our case, $41 = \lambda^{-1}$ so $\lambda \approx 0.0227$
- Look up CDF: $1 - e^{-\lambda x}$
- Is 0.878 at $167 - 100$
- So prob of each remaining person turning in by deadline is 0.781
- What about number of people?

Another Example: Assignment Turn In

5 people, each with 0.781 chance of turning in at deadline -1

How to model?

- $N \sim$ **Binomial (0.878, 5)**
- N is the number turning in
- $\Pr(N = 5) = 0.291 =$ prob all 10 turn in
- $\Pr(N = 4) = 0.698 =$ prob 9+ turn in
- $\Pr(N = 3) = 0.926 =$ prob 8+ turn in
- $\Pr(N < 3) = 0.074 =$ prob $<$ 8 turn in

Closing Remarks

In modeling, three big tasks

- Choosing the model
- Learning the model
- Applying the model

Will focus on all three in upcoming weeks!

Questions?