Tools & Models for Data Science

Introduction to Modeling 1

Chris Jermaine & Risa Myers

Rice University

What is a Model?

- Many definitions!
- Traditional statistical definition:
  - A set of assumptions regarding the (stochastic) process that generated the data
  - Classical statistical approach:
    - Assume some stochastic process [1] generated the data
    - We want to figure out how the model generated the data
- More modern definition:
  - A mathematical object that enables an analyst to use data to understand the past and present, and make predictions about the future

[1] A stochastic process is a random process that changes over time i.e., a mathematical object usually defined as a collection of random variables

Why Do We Model?

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

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 about future events
  - Example: Supervised learning

Modeling Process

- This what data scientists do every day
- In modeling, four big tasks
  - 1. Choosing the model—choose family, complexity, hyperparameters
  - 2. Learning the model—“fit” model to data by adjusting parameters
  - 3. Validating the model—make sure model matches data
  - 4. Applying the model—use the model to explain past/present make predictions on future
- Often, 1 thru 3 repeated iteratively until model matches data
- Will focus on all four in upcoming weeks!

1. Choosing the model

- Select the distribution or distribution family [2]: e.g. Exponential family
- Choosing the hyperparameters [3]
  - Can be informative (e.g. biasing a parameter to be close to 0)
  - Can be noninformative (e.g. allowing values to be selected uniformly over a range)
- Note that hyperparameters are external to the model and aren't based on the data
- Model parameters are estimated from / learned from the data

[2] Probability distributions are not a single distribution or function, but are a family of distributions because they have different shape parameters that allow them to have a variety of different forms/shapes.

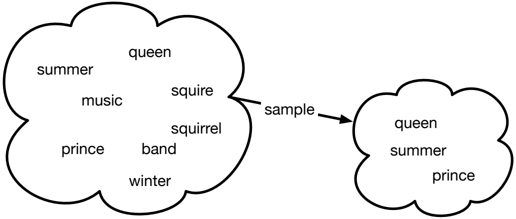
[3] A hyperparameter is a parameter of a prior distribution used in a model

2. Learning the model

- Use the existing dataset to figure out the model parameters
- Approach can be dependent on the quantity of data you have
- Example
  - Choose an appropriate loss function
  - Minimize or maximize the loss function to optimize the parameters

Course Scope

- Models can be biased based on the data you choose
- Data evolves over time
- These are really important issues
- . . . that we will NOT cover in this course



3. Validating the model

- Assume you have “learned” a model
- Want to figure out if the model is useful or not
- Common problem is Overfitting
- Approach can be dependent on the quantity of data you have

4. Applying the model

- Use the model on new data
- This is what you report & use

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"
  - Flip a coin
  - H T H H H H
  - P(Heads) = $\frac{5}{6}$
- Probability is used less in modern models
  - Deep learning

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 11AM
  - 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

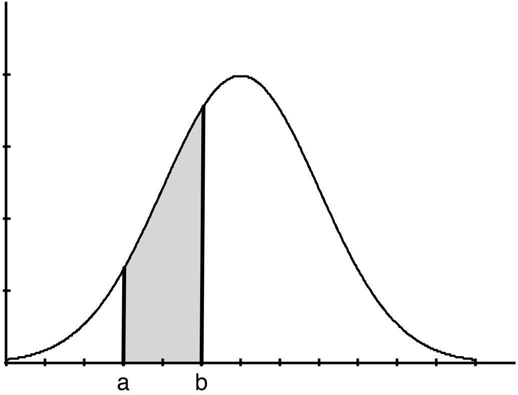
- Probability density gets around this problem of having to deal with very small absolute probabilities
  - Measures the relative likelihood of an event—not absolute
- Probability A2 takes 5 hours is nonsensical
- But...
  - Probability density at ‘A2 takes 5 hours’ is $5X$’ A2 takes 1 hour
  - Sensical!

Probability Density Function

- A PDF is a function that computes the relative likelihood of an event
- Most famous: normal PDF

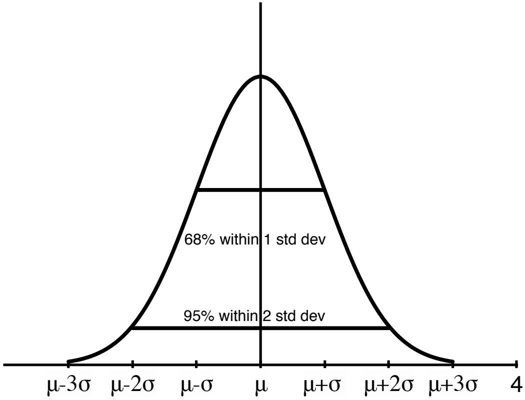
$$f_{\text{Normal}}(x|\mu,\sigma) = \sigma^{-1}(2\pi)^{-\frac{1}{2}}e^{-\frac{1}{2}(x-\mu)^2\sigma^{-2}}$$

- 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



The Normal/Gaussian Distribution

- Is continuous
- Arguably the most popular statistical distribution
- Many data in real life follow this distribution
- Models processes that can be viewed as the sum of multiple processes
- The math is nice: $e^a * e^b = e^{a+b}$
- Is super important because of the Central Limit Theorem
  - Under certain conditions the histogram of the normalized sum of independent random variables will follow a Normal distribution



- Parameters
  - μ = the mean value
  - σ^2 = the variance

Choosing a Model

- There is a well known aphorism:
  - “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 (too much) inaccuracy!!

Example: Predicting Grade in Class

- A student has completed 5/10 assignments
  - Want to predict grade in class
- First, choose a model
  - Ex: assume $X_i \sim \mathrm{Normal}(\mu, \sigma)$
  - i is the identity of the assignment
  - Note: X_i is a random variable controlling a score
  - $f_{X_i}(x)$ gives relative likelihood X_i takes value x
  - (or the probability if X_i is discrete!)
  - So $f_{X_i}(x) = f_{\mathrm{Normal}}(x|\mu, \sigma)$

i	Score
1	89
2	92
3	78
4	94
5	88
6	-
7	-
8	-
9	-
10	-
Avg	?

Should We be Assuming Scores are Normal?

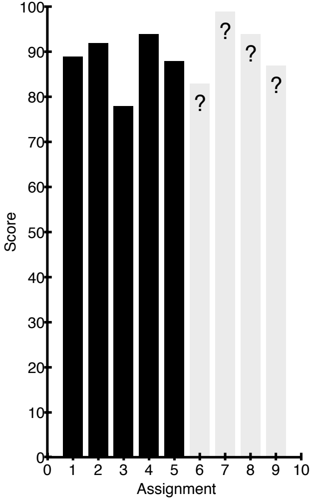
- Probably not, for a single student
- Scores are probably relatively similar for a single student
- Sometimes life happens, and a student does poorly on an assignment
- So, reality might be a more right skewed curve
- Also, scores are usually discrete
- But it's typically easier to use continuous distributions in practice

Random Variables

- X_i is a Random Variable (RV)
- It is normally distributed with some mean and variance, μ and σ
- e.g. X_2 denotes the RV that controls the student's score on assignment 2
- A RV is basically a machine:
  1. Press a button
  2. A stochastic process spits out an outcome
- The distribution of the RV controls which stochastic process is inside the machine

Learning the Model

- Scores so far: $\{89, 92, 78, 94, 88\}$
  - Estimate mean $\mu = 88.2$, $\sigma^2 = 30.56$
  - ? Where did we get these values?
- Thus, $X_i \sim \text{Normal}(\mu, \sigma^2) \sim \text{Normal}(88.2, 30.56)$
- And so $(\sum_{i=6\ldots10} X_i) \sim \text{Normal}(88.2 \times 5, (30.56 \times 5))$
- This is an example of the "Method of moments" estimator
  - 1st: Mean
  - 2nd: Variance
  - ...
  - ? What assumptions have we made?



Our Assumptions

- The data are independent
- Probably not true in this case
- If a student does well so far, the student is likely to do well the rest of the semester
- If a student is doing poorly, the student may give up and do even worse
- We could take this into account (add covariances, etc.), but not in this course

Validating the Model

- So little data, won't do it here
    - In general, requires checking whether $\text{Normal}(88.2, 30.56)$ actually describes data
    - Often involves holding back test and validation sets
    - More on this later
    - Let's just assume our model is valid...

Getting Ready to Apply the Model

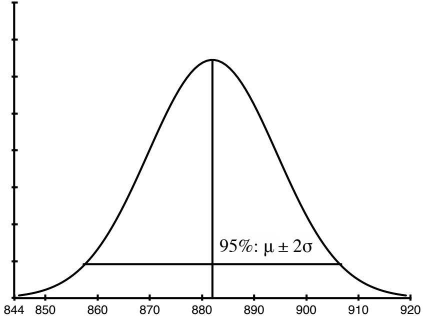
- Scores so far: $\{89, 92, 78, 94, 88\}$
  - Estimate mean $\mu = 88.2$, $\sigma^2 = 30.56$
  - Thus, $X_i \sim \text{Normal}(88.2, 30.56)$
  - And so $\left(\sum_{i=6...10} X_i\right) \sim \text{Normal}(88.2 \times 5, (30.56 \times 5))$

Applying the Model

- We have a mean of 88.2 on the first 5 scores
- We expect a mean of 88.2 on the next 5 scores
- This gives us a total of $88.2 * 10 = 882$ for the expected sum on the mean of all the scores
- 95% confidence on sum: $882 \pm 2 \times 12.36 = 882 \pm 24.7$

? Where does the $\pm 2 \times 12.36$ come from? =

- Hence, 95% confidence on grade is 88.2 ± 2.47



The Sniff Test

- 95% confidence interval on grade is 88.2 ± 2.47
- ? What do we mean by 95% confidence?

The Sniff Test

- 95% confidence interval on grade is 88.2 ± 2.47
- What do we mean by 95% confidence?
  - If repeated samples were taken out of a given population, and you calculated a 95% confidence interval for each sample, that means that 95% of these intervals would contain the population mean
  - Remember that CIs are the probability of the parameter lying in the interval BEFORE you calculate this interval (not that the parameter has a 95% chance of being in a given interval you've calculated with 95% CI)

The Sniff Test

- 95% confidence interval on grade is 88.2 ± 2.47

? Does this seem reasonable?

The Sniff Test

- 95% confidence interval on grade is 88.2 ± 2.47
- Does this seem reasonable?
- The standard deviation seems low
- Low standard deviation on existing scores implies small range in the future

? Where does the smallness come from?

The Sniff Test

- 95% confidence on grade is 88.2 ± 2.47
- Does this seem reasonable?
- The standard deviation seems low
- Low standard deviation on existing scores implies small range in the future
- Where does the smallness come from?
- Our standard deviation is based on only 5 data points
- We could have a bad estimation for the moments of distribution because we have such little data

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

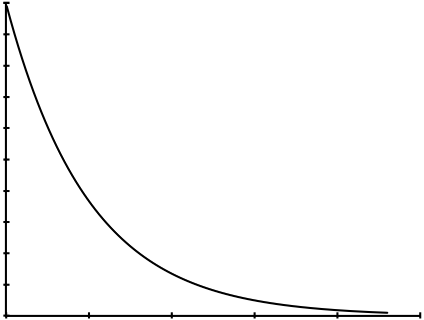
Choosing a Model

- 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 $X_i \sim \text{Exponential}(\lambda)$
  - Exponential PDF:

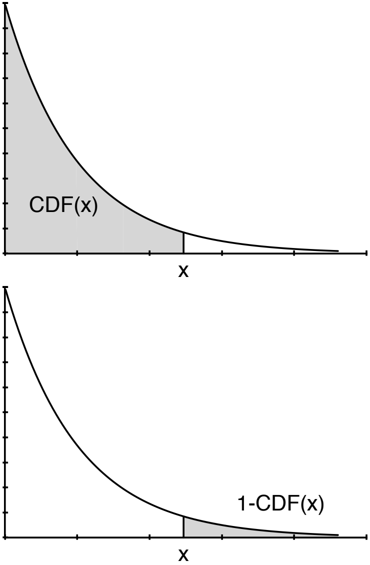
$$f_{Exp}(x|\lambda) = \lambda e^{-\lambda x}$$

The Exponential Distribution

- Is continuous
- Has 1 parameter λ, which determines how quickly the mass drops off
- Mean: λ^{-1}
- Variance: λ^{-2}
- Is "memoryless"
    - That t time units have passed doesn't matter
    - Means if waited t units so far...
    - $f_{Exp}(x|\lambda, x \geq t) = f_{Exp}(x - t|\lambda)$
- Good for modeling time horizons (e.g. arrivals) and time between events



The Cumulative Distribution Function



- The probability that the RV will have a value $\leq$ x
- Total mass at point x is the area to the left of x
- CDF, F_X of a RV X:
  - $F_X(x) = P(X \leq x)$
  - $F_X(b) - F_X(a) = P(a < X \leq b)$

Learning the Model

- Turn in times so far at tick 100: $\{18, 22, 45, 49, 86\}$
  - Know mean of exponential is λ^{-1}
  - In our case, $44 = \lambda^{-1}$ so $\lambda \approx 0.0227$
  - Use the CDF equation: $1 - e^{-\lambda x}$

  - Recall: Want to predict how many have completed by 1 hour before due date
  - So, $x = 167 - 100$
  - CDF = $1 - e^{-\lambda x} = 1 - e^{-0.0227 * 67} \approx 0.781$
  - So, the probability of each remaining person turning in by deadline is 0.781

The Sniff Test

- If we only look at the early finishers, we are underestimating the mean
- Also, we've only looked at half the students
- Fixing these assumptions is non-trivial - we will examine it next time
- If we accept our assumptions as valid …

Applying the Model

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

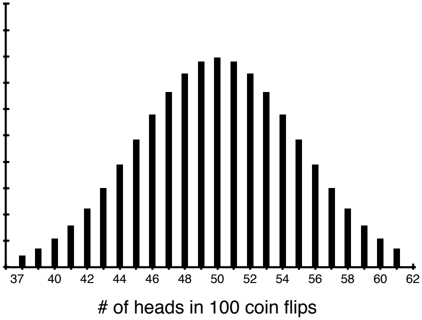
? How should we model this?

Applying the Model

- 5 people, each with 0.781 chance of turning in at deadline -1 hour
- How should we model this?
  - We have a probability and two possible outcomes (Turned in by deadline or Not turned in by deadline)
  - This looks like a good fit for the Binomial distribution

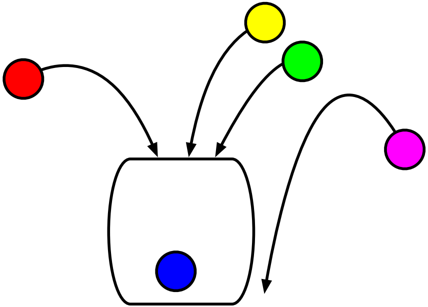
The Binomial Distribution

- Is discrete
- Has 2 parameters
  - $n =$ number of independent experiments
  - $p =$ probability of success
  - Probability Mass Function = $\binom{n}{k} p^k (1-p)^{(n-k)}$
  - Mean: np
  - Variance: $np(1-p)$
- Good for modeling Yes/No choices, n times
- Assumes trials are independent
- Degenerative form is the Bernoulli distribution, when $n = 1$



The Binomial Distribution In Our Example

- Think about tossing n balls into a trash can
- Each ball has a 0.781 probability of success
- The binomial PMF and CDF will tell me probabilities of success
- Use PMF for exact number of successes
- Use CDF and 1-CDF for greater than or less than



Applying the Model

- 5 people, each with 0.781 chance of turning in at deadline -1 hour
  - $N \sim \text{Binomial}\,(5, 0.781)$
  - N is the number turning in assignment by the deadline
  - $\Pr(N = 5) = 0.291 =$ prob all 10 turn in
  - $\Pr(N \geq 4) = 0.698 =$ prob $9+$ turn in
  - $\Pr(N \geq 3) = 0.926 =$ prob $8+$ turn in
  - $\Pr(N < 3) = 0.074 =$ prob < 8 turn in
- Note: there's a slight problem here
  - We ignored people missing when estimating λ
  - We will fix this next lecture!

Questions?

? How can we use what we learned today?

? What do we know now that we didn't know before?