11. Statistics: distributions and spread LING 471

1

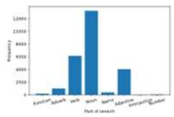
Learning outcomes

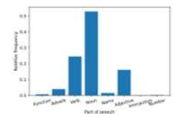
- Describe a bar plot and a histogram
- Describe probability density
- Describe the Gaussian/normal distribution
- Write code that evaluates the probability density of a Gaussian distribution
- Write code that finds the mean and standard deviation from data
- Implement a linear discriminant analysis classifier by calculating conditional probability with Gaussian distributions

2

Probability mass, density, and distributions

3

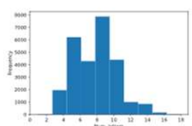


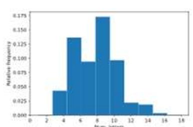


Bar plots

- Last week, we talked about frequency of occurrence and relative frequency
- We can visualize **category** frequencies with **bar plots**
- Either raw or relative frequency can be visualized
- When using relative frequency, we are approximating a **probability mass function**

4

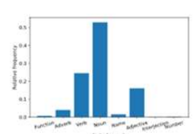


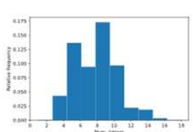


Histograms

- We can also visualize the frequency of non-categorical data with a histogram
- Must first create discrete bins (basically, numerical categories)
- When using relative frequencies, we are approximating the **probability density function**

5





Probability mass vs. density

- Probability **mass** is for **categorical** outcomes
  - Part of speech, coin flips, rolling dice, etc.
- Probability **density** is for **continuous** outcomes
  - Word/review length (sort of), word/utterance duration, height, etc.
- Notice the different spacing between bars for bar plots and histograms

6

Probability distributions

- Recall that we said the probabilities of mutually exclusive events need to sum to 1
- For a particular sample space, we need to **distribute** those probabilities (or relative frequencies)
  - I.e., make a **probability distribution**
- Not all possible distributions will distribute the probability evenly or **uniformly**
  - See: bar plots and histograms on previous slides!

7

Notes on probability distributions

- Centuries ago, mathematicians noticed similar shapes kept appearing when analyzing data
- These commonly occurring shapes were named and formalized
  - Examples: uniform, Gaussian/normal (named for Gauss), Bernoulli/binomial (named for Bernoulli)
- However, not all distributions resemble known functions
  - We will often approximate them with known functions, though

8

Continuous variables and distributions

9

Continuous random variables

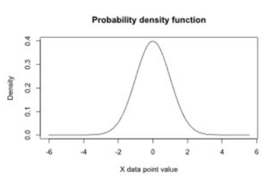
- Some random variables are **continuous**
  - Bascially, your values should theoretically be able to "reach" infinity or have infinite possible values (even if you can't realistically observe some of them)
    - E.g., age: 25.31415 years old
    - E.g., a word could, theoretically, have infinite letters
      - And number of letters is a proxy for how long it takes to say a word
  - Contrast with a discrete variable like a coin flip (either H or T, no in between, no infinite possible values)

ADAPTED FROM OLGA ZAMARAEVA 10

10

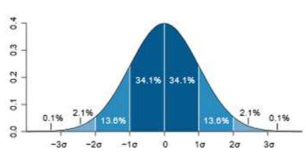
Probability density functions (PDFs)

- A continuous random variable has a probability **density**
- Area under curve must sum to 1
- Each specific point is a density, not a probability!
  - Probability requires calculating the area under a region (integrating)
- Most common PDF use for continuous variables is **normal/Gaussian distribution**



ADAPTED FROM OLGA ZAMARAEVA 11

11



The normal/Gaussian distribution

- A lot of data in the world is normally distributed
  - Or can reasonably be approximated as such
- Has two **parameters**

1. Mean (signified μ or m)
   - Indicates where the middle of the distribution is; most typical/likely value
2. Standard deviation (signified σ or s)
   - Indicates how much **spread** or **variation** there is in the distribution

ADAPTED FROM OLGA ZAMARAEVA 12

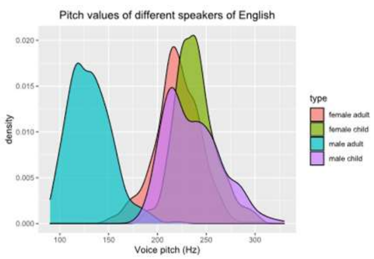
12

Normal distribution notes

- Greek letters used for population, Roman letters for sample
- Sometimes, variance (var) is used instead of standard deviation
  - $var = \sigma^2$
- Etymology of "normal"
  - Gauss used it to refer to orthogonality (right angles); we might talk a bit about this next week
  - Today, "normal distribution" doesn't really break down into "normal" + "distribution"
  - Rather, "normal" is used as a label (with no other meaning)

ADAPTED FROM OLGA ZAMARAEVA 13

13



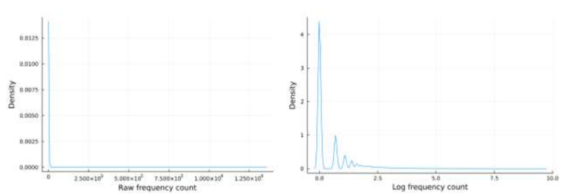
Language data and the Gaussian distribution

- Is language data reasonably normal?
  - It depends…
- Some of it is!
  - See pitch plot here
- Some of it isn't!
  - See frequency data

ADAPTED FROM OLGA ZAMARAEVA 14

14

Probability density of frequency counts in *Ulysses*



Clearly not Gaussian…

ADAPTED FROM OLGA ZAMARAEVA 15

15

Normal distribution features

- Mean is calculated as the average of all possible outcomes
- Mean is most likely value
- About 67% of data is within 1 standard deviation from the mean
- About 95% of data is within 2 standard deviations from the mean
- How to calculate Gaussian probability density
  - $pdf(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\pi-\mu}{\sigma}\right)^2}$
- If you write a function to do this, you'll want to split this up a bit
  - Usually, you want to use someone else's code to do this…

ADAPTED FROM OLGA ZAMARAEVA 16

16

Applying Gaussians to language data (phonetics)

- Let's try applying Gaussian distributions to some language data
- We're going to try to classify certain acoustic measurements as belonging to one vowel or another
- To do this, we are going to manually implement a **linear discriminant analysis classifier**
  - This will probably feel overwhelming and dense
  - We are going to work slowly!

ADAPTED FROM OLGA ZAMARAEVA 17

17

Linear discriminant analysis (LDA) classification

- Uses a mix of probability density and frequency counts to estimate conditional probabilities of the different categories
  - Quantifying the probability of a data point X being in a particular category k
- Remember from last week? (I didn't…)
  - $P(Y = k|X) = \frac{P(X \cap Y)}{P(X)}$
  - note $P(X \cap Y)$ is an alternative for our $P(X \; and \; Y)$
- We're going to do some hand-waving and just say
  - $P(X \cap \; Y) = P(X|Y = k) * P(Y = k)$
  - Where $P(X|Y = k)$ is the **density** from the Gaussian distribution, and $P(Y = k)$is the **relative frequency** of the different classes

ADAPTED FROM OLGA ZAMARAEVA 18

18

Full LDA classification formula/algorithm

1. Assume we have *K* categories
2. Estimate means for each category
3. Estimate the standard deviations for each category, and then take the mean of those values
4. For each data point *X*
   a) For each category *k* in our *K* categories:
      i. Calculate $P(Y = k|X)$ for *k*
   b) Compare each calculate probability
   c) Classify *X* as belonging to the highest probability category

19

Full $P(Y = k|X)$

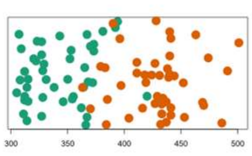
FORMULA

$$P(Y = k|X) = \frac{P(X|Y = k) * P(Y = k)}{\sum_l P(X|Y = l) * P(Y = l)}$$

NOTES

- The normal distribution appears twice: once in the numerator, and once in the denominator
- In the denominator, we are just taking a weighted average of the different densities
  - I promise this will be simple once we implement it!

20



The data we are working with

- We will be working with acoustic measurements from vowels
- Speakers are adult males
- Modeling first resonant frequency of [i] (as in *heed*) and [ɪ] (as in *hid*)
  - Correlates with how high or low your tongue is in the mouth
  - Y-axis doesn't mean anything here
- Need to choose where to draw a vertical line to best separate the vowels

21

Programming activity part 1: Implementing the numerator

- We are going to start with the top part of the fraction
- Do everything marked with "TODO_1"

1. Install the scipy and numpy package in PyCharm
2. Download the skeleton and txt file from the class GitHub
3. Fill in first todo
   a) Use numpy.mean and numpy.std to calculate means and standard deviations
      i. Make sure to use give the argument "ddof=1" to numpy.std
   b) Use scipy.stats.norm.pdf to calculate the Gaussian PDF
   c) The classes are evenly distributed, and we only have two classes, so anything like P(Y=k) or P(Y=l) is 0.5

ADAPTED FROM OLGA ZAMARAEVA 22

22

Programming activity part 2: Implementing the denominator

- Now, we're going to implement the denominator
- This is marked with "TODO_2"
- You might be able to make use of your numerator function if you recognize an equivalence…
- Uncomment the "test_denom()" line to test your implementation

ADAPTED FROM OLGA ZAMARAEVA 23

23

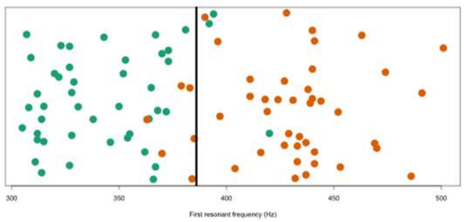
Programming activity part 3: Implementing the fraction and LDA

- Finish implementing the program
- This is marked with "TODO_3"
- Uncomment "test_lda()" to test your implementation
- When ready, uncomment "accuracy()" to see how your classifier performs
  - It should be 0.9
- Uncomment the rest of the code to determine where we should draw our separator line

ADAPTED FROM OLGA ZAMARAEVA 24

24

Where to draw our separating line? 386.03 Hz



25

Zooming out: What did we just do?

- We calculated means and standard deviations
- We computed conditional probabilities
  - Based on Gaussian distributions from the means and standard deviations
- We used the conditional probabilities to classify vowel data
  - Answering, "what is the probability that the vowel is [i] if the measurement is X?" and "what is the probability that the vowel is [ɪ]…?"

26

Zooming further out

- This was practice working on a specific kind of programming
- Technical implementation from a specification (math formula) even if you may not know entirely what everything is doing
  - This can be very hard!
  - But also not that uncommon when you are beginning to learn something
  - Often helps to decompose the problem into parts (as was done here)

27

Where are Gaussian distributions useful? And LDA?

- Gaussian distributions are everywhere… Many common statistical methods take advantage of them in some form
- Many continuous linguistic phenomena can be modeled with Gaussian distributions (and sometimes they're modeled well!)
  - Acoustic measurements, semantic vectors, psycholinguistic and neurolinguistic data, sociolinguistic variables like age, etc.
- LDA can be used any time you want to classify something
  - Not always the best choice…
  - Also commonly used to reduce a large number of variables to a smaller number of variables

ADAPTED FROM OLGA ZAMARAEVA 28

28

Reminders

- Assignment 3 due on 5/10
- Meeting in Miller Hall 301 (MLR 301) on 5/10 and 5/12

ADAPTED FROM OLGA ZAMARAEVA 29

29