DATA605

Week9 and Week10 Lecture

What are Distributions / PDFs?

One way to look at it is that a distribution is a model of how certain random variables are produced.

Last week we reviewed the formal definition: a function that maps a value a random variable can take, and output the likelihood of seeing that value

Common Distributions

- Uniform
- Bernoulli
- Exponential
- Poisson (not in the notes)
- Normal Distribution

We'll review each one, including whether or not its discrete or continuous, the PDF / PMF, and situations when they are useful

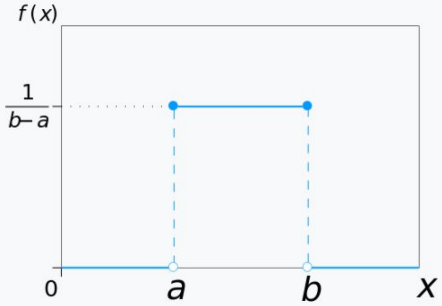
Uniform Distribution

This distribution models random variables that have equally likely outcomes for the values it can take on. Snippet on the right is from wikipedia, showing you the PDF of a random variable that takes on values in [a, b]. The PDF is flat since all the probabilities / likelihoods for those values are the same.

The discrete case is simply that all discrete values in the domain of the random variable have 1 / n probability, where n = number of discrete values

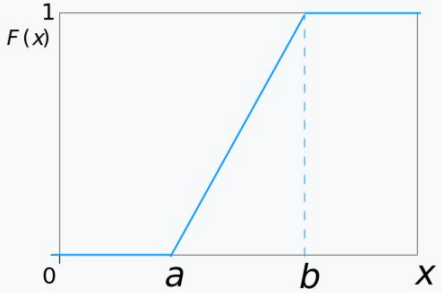
Uniform

Probability density function



Using maximum convention

Cumulative distribution function



Notation	$\mathcal{U}(a,b)$ or $\operatorname{unif}(a,b)$
Parameters	$-\infty < a < b < \infty$
Support	$x \in [a,b]$
PDF	$\begin{cases} \frac{1}{b-a} & \text{for } x \in [a,b] \\ 0 & \text{otherwise} \end{cases}$

Bernoulli Distribution

This distribution models a random variable that can take on only two values. Think of a coin flip, or a bit. The PMF is defined by a parameter p, which is the probability for one of the values allowed. Implicitly, the other value has a probability of 1 - p (since all probability values must sum to 1).

Bernoulli

Parameters	$0 < p < 1, p \in \mathbb{R}$
Support	$k \in \{0, 1\}$
pmf	$\begin{cases} q = (1-p) & \text{for } k = 0 \\ p & \text{for } k = 1 \end{cases}$

Poisson Distribution (Not in the Notes)

This distribution is used to determine the probability that a number of events that occur in a fixed interval of time.

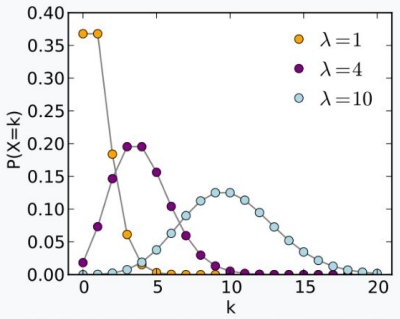
Specifically, these events must be from a ‘poisson process’, which generates events such that they occur with a known average rate (lambda) and independently of the time since the last event.

This is discrete, since we are talking about integers (there could be 0, 1, 2, etc events in a fixed time interval).

Example: For instance, an individual keeping track of the amount of mail they receive each day may notice that they receive an average number of 4 letters per day.

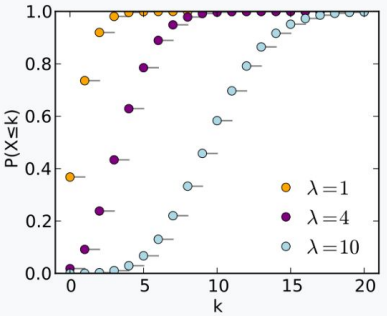
Poisson

Probability mass function



The horizontal axis is the index *k*, the number of occurrences. *λ* is the expected number of occurrences. The vertical axis is the probability of *k* occurrences given *λ*. The function is defined only at integer values of *k*. The connecting lines are only guides for the eye.

Cumulative distribution function



The horizontal axis is the index *k*, the number of occurrences. The CDF is discontinuous at the integers of *k* and flat everywhere else because a variable that is Poisson distributed takes on only integer values.

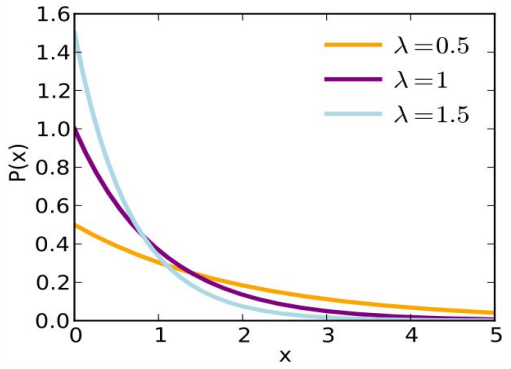
Parameters	$\lambda > 0$ (real)
Support	$k \in \mathbb{N} \cup 0;$
pmf	$\frac{\lambda^k e^{-\lambda}}{k!}$

Exponential Distribution

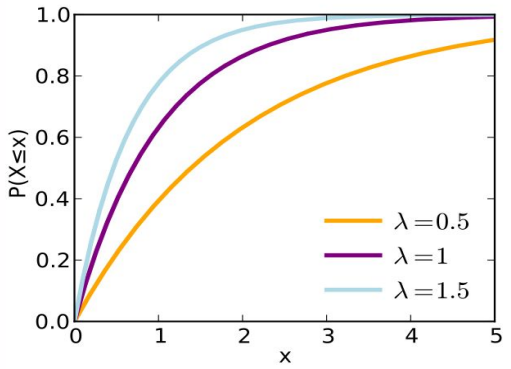
This probability distribution describes the time between events in a Poisson process. Before in the Poisson distribution, we cares about the number of events; here we care about time in between events occurring from the same process.

Exponential

Probability density function



Cumulative distribution function



Parameters	$\lambda > 0$ rate, or inverse scale
Support	$x \in [0, \infty)$
PDF	$\lambda e^{-\lambda x}$
CDF	$1 - e^{-\lambda x}$

Normal / Guassian Distribution

Many of you have probably seen this before. This distribution is a model of a random variable that varies around a central mean (hence the two parameters that define the PDF, the mean and the variance)

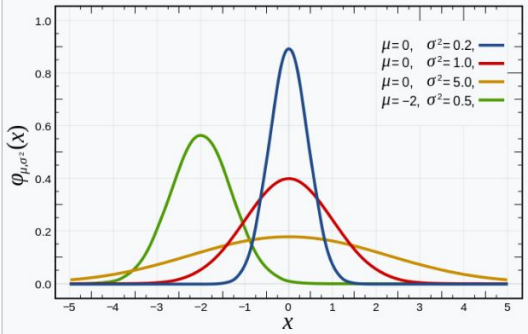
When dealing with random variables dealing with measurements from nature, normal distributions happen often. Important also due to Central Limit Theorem. Another useful property is that sometimes assuming a variable is normal makes certain math ‘tractable’.

Sometimes used to represent the distribution of a variable where the distribution isn’t really known yet.

Check here for a deeper discussion.

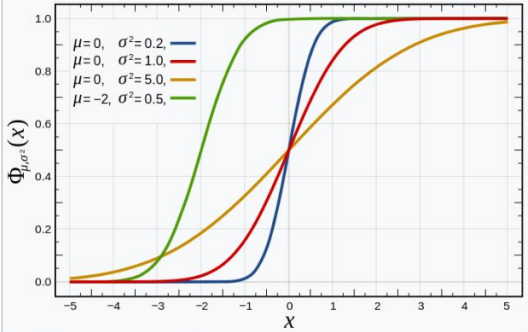
Normal distribution

Probability density function



The red curve is the *standard normal distribution*

Cumulative distribution function



Notation	$\mathcal{N}(\mu, \sigma^2)$
Parameters	$\mu \in \mathbf{R}$ — mean (location) $\sigma^2 > 0$ — variance (squared scale)
Support	$x \in \mathbf{R}$
PDF	$\frac{1}{\sqrt{2\sigma^2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

Central Limit Theorem

“In precise terms, given certain conditions, the arithmetic mean of a sufficiently large number of iterates of independent random variables, each with a well-defined (finite) expected value and finite variance, will be approximately normally distributed, regardless of the underlying distribution” -- https://en.wikipedia.org/wiki/Central_limit_theorem

Terminology here is confusing. Sometimes we use the word ‘sample’ to mean one value that came from a random value. However, let us call that an observation. In the **CLT**, a sample therefore is made up of n observations of the random variable. Given a sample of n observations, we can find the sample mean.

CLT talks about the distribution of the sample means. The mean value of each such sample set will be normally distributed around the mean of the **original distribution**. No matter what the underlying distribution of the random variable is, when we take samples of n observations

$$\text{The } \mathbf{sample\ mean}\ \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \text{ is approximately normal } N(\mu, \sigma^2/n).$$

From: http://www.math.illinois.edu/~ajh/408/408clt.pdf

CLT versus **Law of Large Numbers**

Law of Large Numbers is a property about a single sample of n observations: “Law of Large Numbers states that if you take a sufficiently large sample (n increases), then the sample mean approaches the Expected Value of the distribution”

CLT is a property about a multiple samples, in that the mean of the sample means follows a normal distribution with the same mean as the original distribution.

Review Interactive Website

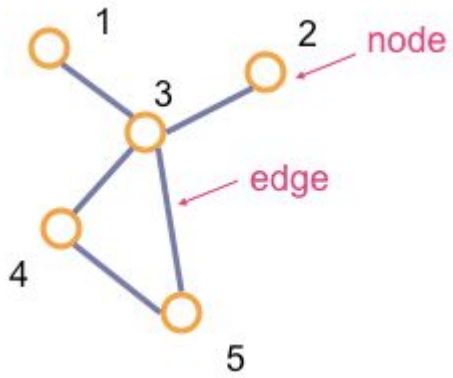
http://students.brown.edu/seeing-theory/distributions/index.html#third

The above link has a CLM visualization

Review Python Code for Sampling

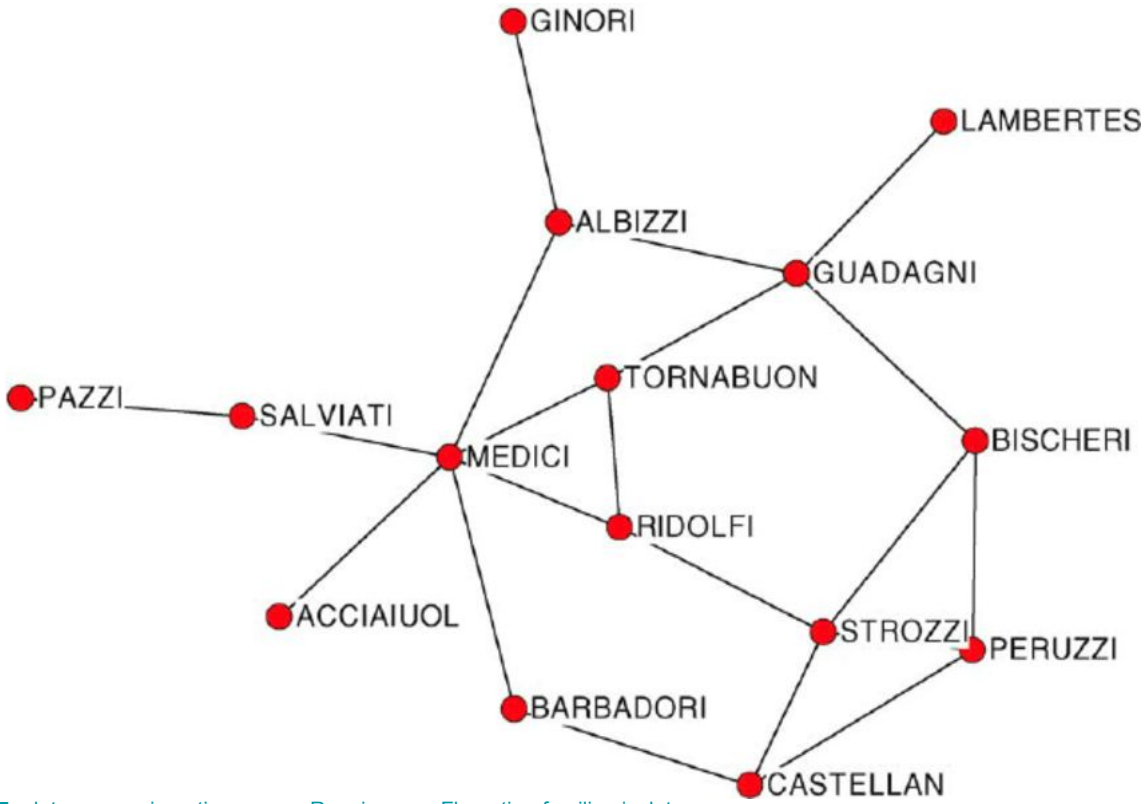
Graphs

A network or graph is a mathematical structure made up of nodes that connect to one another via edges. This is useful for social media analysis, where each node is a person and edges represent friendship / following relationships



Centrality / Importance Measures

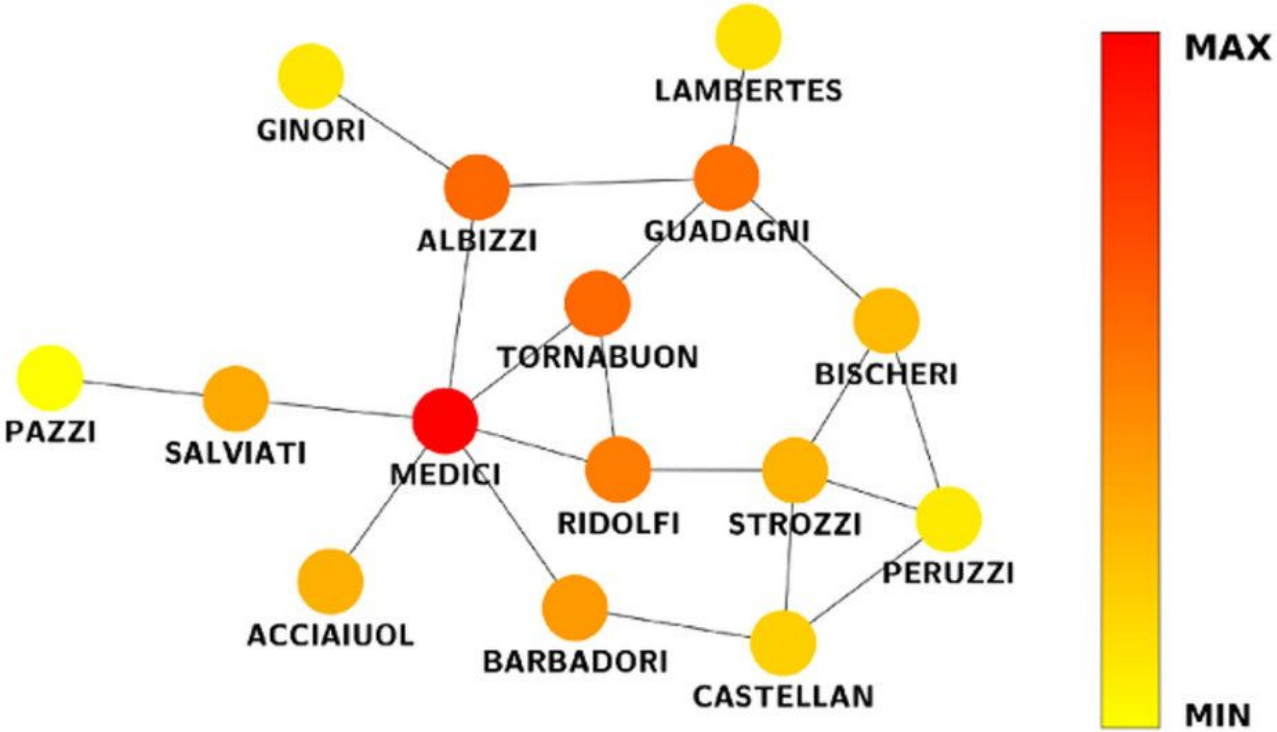
This is an example social network, where each node is an important family in Florence around the 15th century. We'd like to have some metric that measures how 'important' a node is. We call these 'centrality measures'.



https://www.researchgate.net/figure/222538101_fig4_Fig-4-Padgett%27s-data-on-marriage-ties-among-Renaissance-Florentine-families-isolate

Centrality / Importance Measures

Here is an image based on the same network, but colored by a centrality measure we call PageRank. Medici is very central. Different metrics will give different values, but are generally trying to pinpoint nodes that are important based on how they link to others



Centrality / Importance Measures

Degree Centrality: Importance measured by simply the degree of the node. A node has more edges in-bound to it, the more important it is

Closeness Centrality: Importance based on a node's average length of the shortest path between the node and all other nodes in the graph

Betweenness Centrality: Importance based on the number of times a node acts as a bridge along the shortest path between two other nodes.

'Prestige' type Centrality Measures: importance ~ importance of neighbors. By definition recursive, and hence is not straightforward to calculate. However recursive equations can be represented by eigenvector questions!

https://en.wikipedia.org/wiki/Centrality

Page Rank

PageRank is measuring importance of nodes in a network by treating nodes as web pages and edges as links between them, upon which a simulated web surfer starts on a random page and has a ‘random walk’ through the graph. This emulates how a user navigates the web. If a node appears more often in the random walks, then it has a higher importance that others based on this metric.

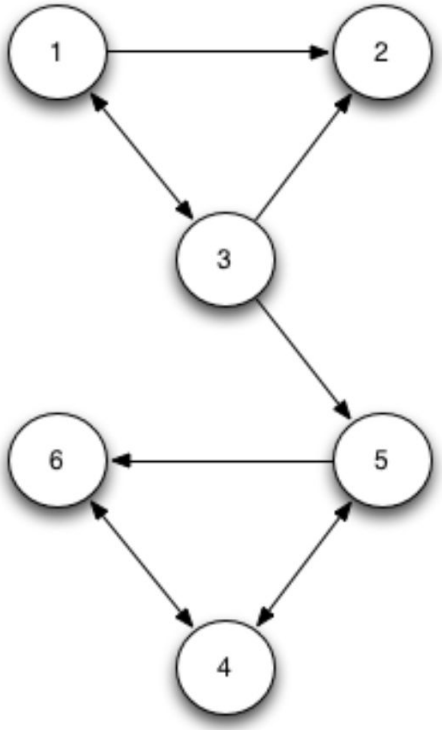
Review visualization: http://bl.ocks.org/emeeks/f448eef177b5fe94b1c0

Page Rank Model

Each node in the graph is P_i and its PageRank is denoted $r(P_i)$. The number of outlinks from a node P_i is denoted $| P_i |$. The recursive definition of PageRank is:

$$r(P_i) = \sum_{P_j \in B_{P_i}} \frac{r(P_j)}{|P_j|}$$

PageRank Transition Matrix



$$\mathbf{A} = \begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{3} & 0 & 0 & \frac{1}{3} & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix}$$

Every non-zero value is 1 / | P_i |. The rows are nodes that link to another node. So A_{ij} is node i connecting to node j. Note that this can be interpreted as the probability of clicking one of the links in page i at random

PageRank Algo

We start off with a vector whose components are the page ranks of vectors, called r. We initialize each component to 1 / number of nodes, which means we start off with each node having a equal chance for a user to start their random walk.

Multiplying by the transition matrix simulates the user taking one random step. We will keep multiplying until we notice that r is no longer changing that much (or essentially $r_{n+1} = r_n = A * r_n$. Notice how this is an eigenvalue problem!

PageRank Issues

If the network has subgraphs that are disconnected, we get page rank orderings per group, which is not useful since we want to compare importance across these subgroups. What to do?

We add a thing called decay, which makes the transition matrix well-behaved and have rows that sum to 1. Each row represents the probability of transferring from one URL to another URL.

Notice: we can think of this as a way to extend our random walk to include cases where the user doesn’t follow a link but simply goes to another URL at random.

PageRank with Decay

Given the A transition matrix before, we form B, the decay version of A (where d = 0.85 is an empirical number chosen):

$$\mathbf{B} = 0.85 \times \mathbf{A} + \frac{0.15}{n}$$

By forcing every row to sum to 1, we can use a Linear Algebra Theorem to prove that the eigenvalue of 1 does exist.

PageRank Visualizations

http://it.toolbox.com/blogs/lim/d3js-for-pagerank-visualization-56160 is what this visualization should look like, but has since been taken down. The Internet Archive has a copy but the nodes do not display sadly, but I think still helps visualize this process: https://web-beta.archive.org/web/20160326095623/https://bebffd479efdabe8c274b02b19ae9140ad412589.googledrive.com/host/0B2GQktu-wcTiaWw5OFVqT1k3bDA/

Please note: in our version of PageRank, there is no need for a ‘dangling node matrix’. What the page calls H, is their transition matrix, or what we called A in the notes.

Graph Databases

In the NoSQL world, graph databases are starting to become common. Some future classes may introduce them, but some of the most popular are:

- Neo4j
- OrientDB
- Ontotext GraphDB
- ArrangoDB