

If the cookie had candy, then very few bites would have no candy.

$$\Pr(\text{no-candy bite} \mid \text{candy cookie}) = \frac{1}{3}$$

The probability of a no-candy bite, given a candy cookie, is 1/3.



If the cookie had no candy, then every bite would have no candy.

$$\Pr(\text{no-candy bite} \mid \text{no-candy cookie}) = 1$$

The probability of a no-candy bite, given a no-candy cookie, is 1.

CS 1671/2071

Human Language Technologies

Session 3: Linear algebra, probability review

Michael Miller Yoder

January 15, 2025

Overview: Linear algebra and probability review

1. Course logistics
2. JupyterHub setup and preprocessing activity
3. Probability review
4. Linear algebra review

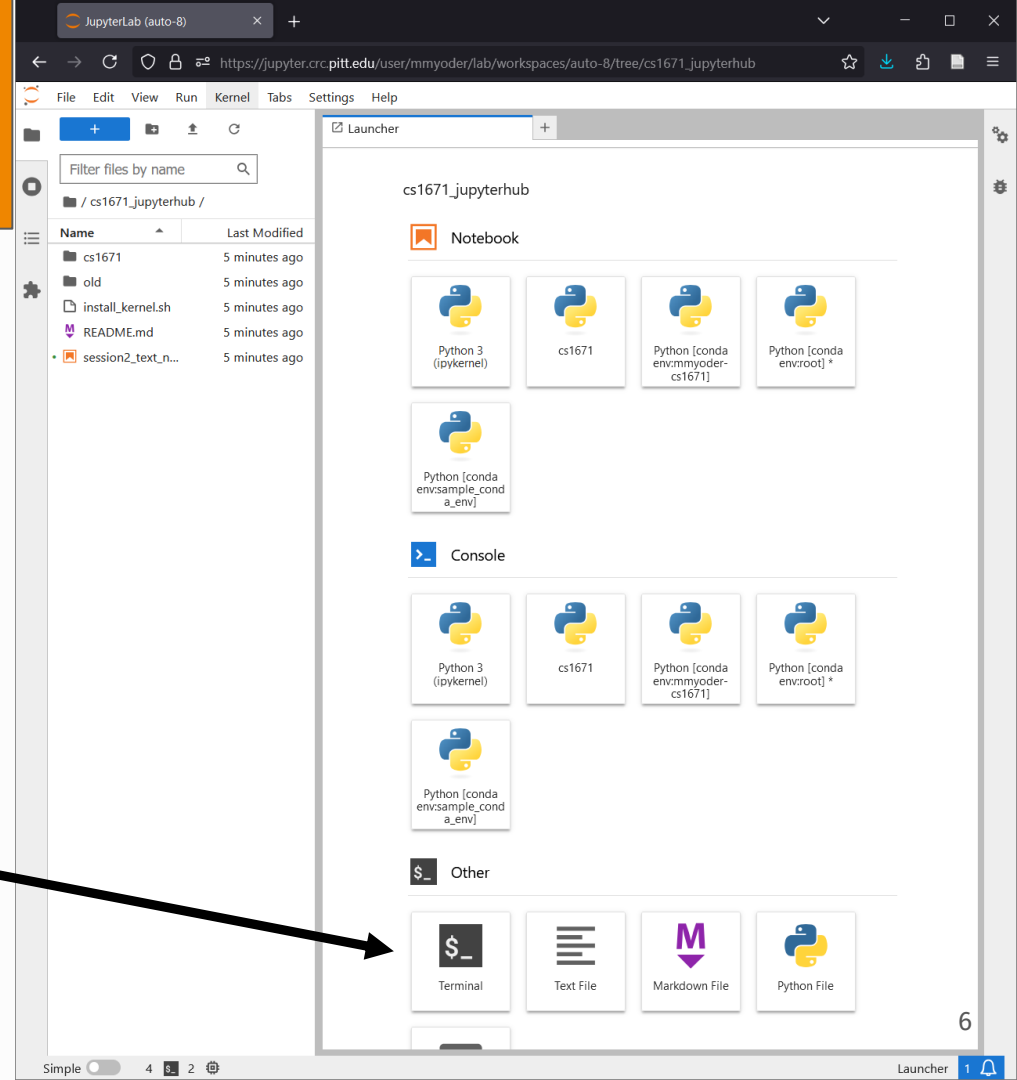
Course logistics

- No class next Mon for MLK Day
- Next class is next Wed Jan 22
- [Homework 1](#) is **due next Thu Jan 23**

JupyterHub setup and activity

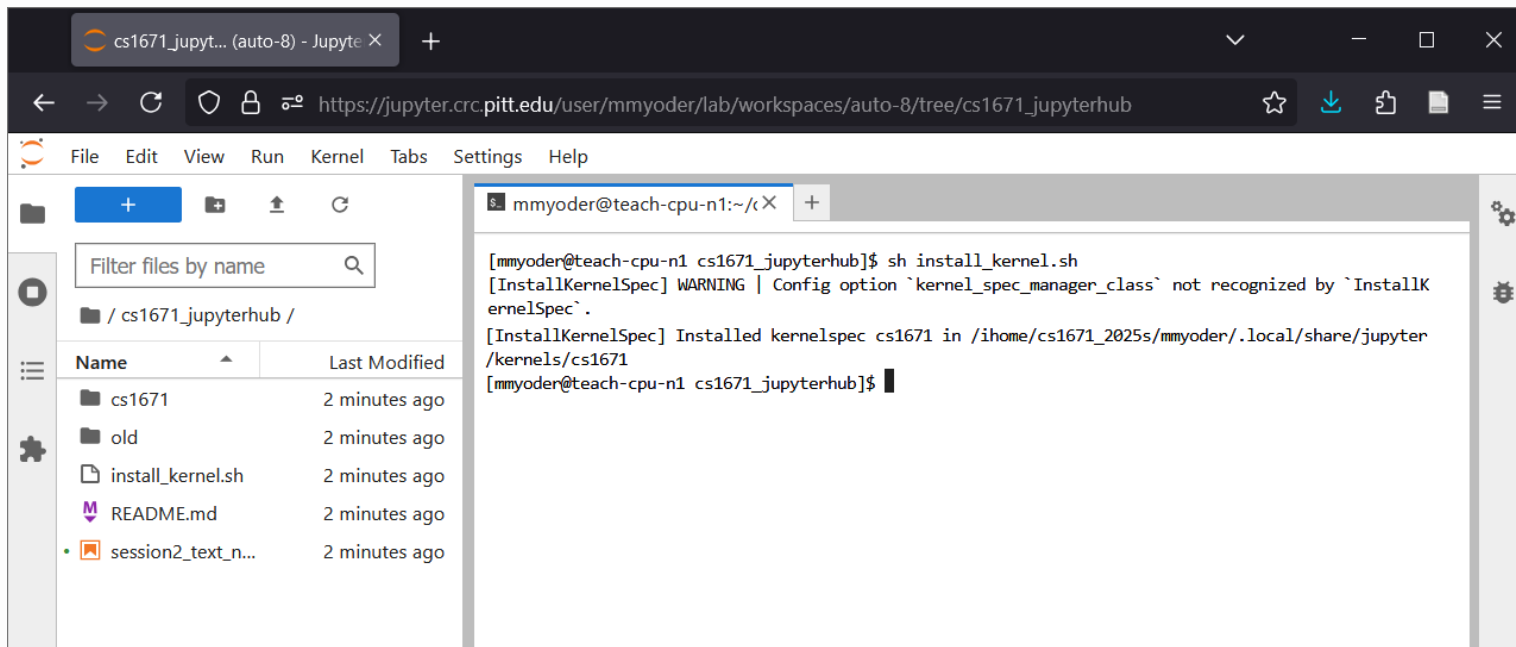
Set up Python virtual environment

1. Go to this [nbgitpuller link](#)
 - Log in with your Pitt username
 - Start a server with **Teach – 6 cores, 3 hours**
 - This should pull a folder (cs1671_jupyterhub) into your JupyterLab
2. Open a terminal



Set up Python virtual environment

In a terminal, run
`sh install_kernel.sh`



The screenshot shows a JupyterLab interface. The browser address bar displays `https://jupyter.crc.pitt.edu/user/mmyoder/lab/workspaces/auto-8/tree/cs1671_jupyterhub`. The left sidebar shows a file explorer for the directory `/cs1671_jupyterhub/` with the following files and folders:

| Name | Last Modified |
|--------------------|---------------|
| cs1671 | 2 minutes ago |
| old | 2 minutes ago |
| install_kernel.sh | 2 minutes ago |
| README.md | 2 minutes ago |
| session2_text_n... | 2 minutes ago |

The terminal window on the right shows the following output:

```
mmyoder@teach-cpu-n1:~/ /cs1671_jupyterhub$ sh install_kernel.sh
[InstallKernelSpec] WARNING | Config option `kernel_spec_manager_class` not recognized by `InstallKernelSpec`.
[InstallKernelSpec] Installed kernelspec cs1671 in /ihome/cs1671_2025s/mmyoder/.local/share/jupyter/kernels/cs1671
mmyoder@teach-cpu-n1 cs1671_jupyterhub$
```

Open Jupyter Notebook

1. Double-click `session2_text_normalization.ipynb` on the left panel to open the notebook
2. From the top menu, click **Kernel > Change Kernel...**
3. Select `cs1671` as your kernel
4. Run the first code cell under **Test kernel and environment** that imports `pandas` and `nltk`

The screenshot shows the Jupyter Notebook interface. The left sidebar displays a file browser with a table of files:

| Name | Last Modified |
|--------------------|---------------|
| cs1671 | 8 minutes ago |
| id | 8 minutes ago |
| install_kernel.sh | 8 minutes ago |
| README.md | 8 minutes ago |
| session2_text_n... | 8 minutes ago |

The main notebook area shows a code cell under the heading "Test kernel and environment" with the following code:

```
[ ]: import pandas as pd
import nltk
```

The "Kernel" menu is open, and the "Select Kernel" dialog box is displayed. The dialog box shows the following options:

- Select kernel for: "session2_text_normalization.ipynb"
- cs1671 (selected)
- Start Preferred Kernel**
- cs1671
- Python [conda env:mmyoder-cs1671]
- Python [conda env:root] *
- Python [conda env:sample_conda_env]
- Python 3 (ipykernel)
- Use No Kernel**
- No Kernel
- Use Kernel from Preferred Session**
- Use Kernel from Other Session**
- session2_text_normalization.ipynb
- Untitled.ipynb

The status bar at the bottom indicates the current kernel is "cs1671_kernel" and the mode is "Command".

Preprocessing Airbnb listings

Implementation

- Remove undesired text with regular expressions
- Lowercase
- Remove stopwords
- Tokenize with the NLTK package
- Stem the tokens with NLTK

Saving your work

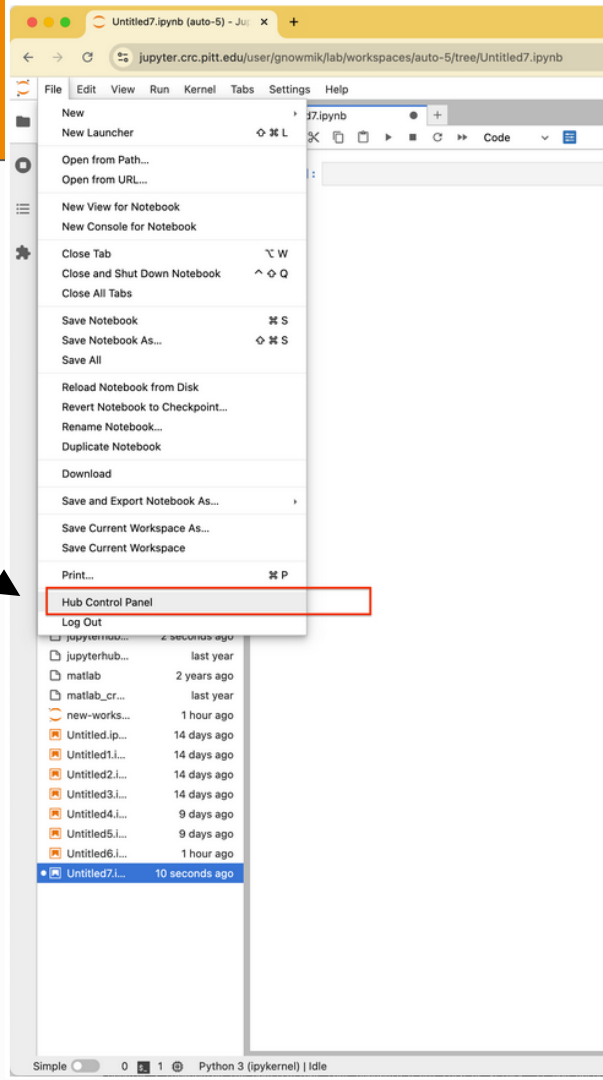
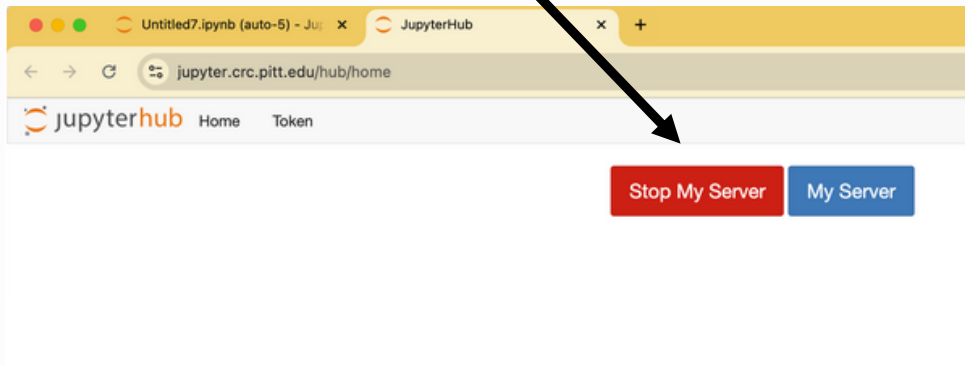
The screenshot shows a Jupyter Notebook interface in a web browser. The browser's address bar displays the URL: `jupyter.crc.pitt.edu/user/gnowmik/lab/workspaces/auto-5/tree/Untitled7.ipynb`. The notebook's title bar shows `Untitled7.ipynb` and the kernel is identified as `Python 3 (ipykernel)`. The `File` menu is open, and the `Save All` option is highlighted with a red rectangular box. Other menu items include `New`, `Open from Path...`, `Save Notebook`, `Download`, and `Log Out`. A file browser sidebar on the left shows a list of files and folders, with `Untitled7.i...` selected and marked as `now`.

| File Name | Last Modified |
|----------------|---------------|
| jupyterhub... | last year |
| matlab | 2 years ago |
| matlab_cr... | last year |
| new-works... | 1 hour ago |
| Untitled.jp... | 14 days ago |
| Untitled1.i... | 14 days ago |
| Untitled2.i... | 14 days ago |
| Untitled3.i... | 14 days ago |
| Untitled4.i... | 9 days ago |
| Untitled5.i... | 9 days ago |
| Untitled6.i... | 1 hour ago |
| Untitled7.i... | now |

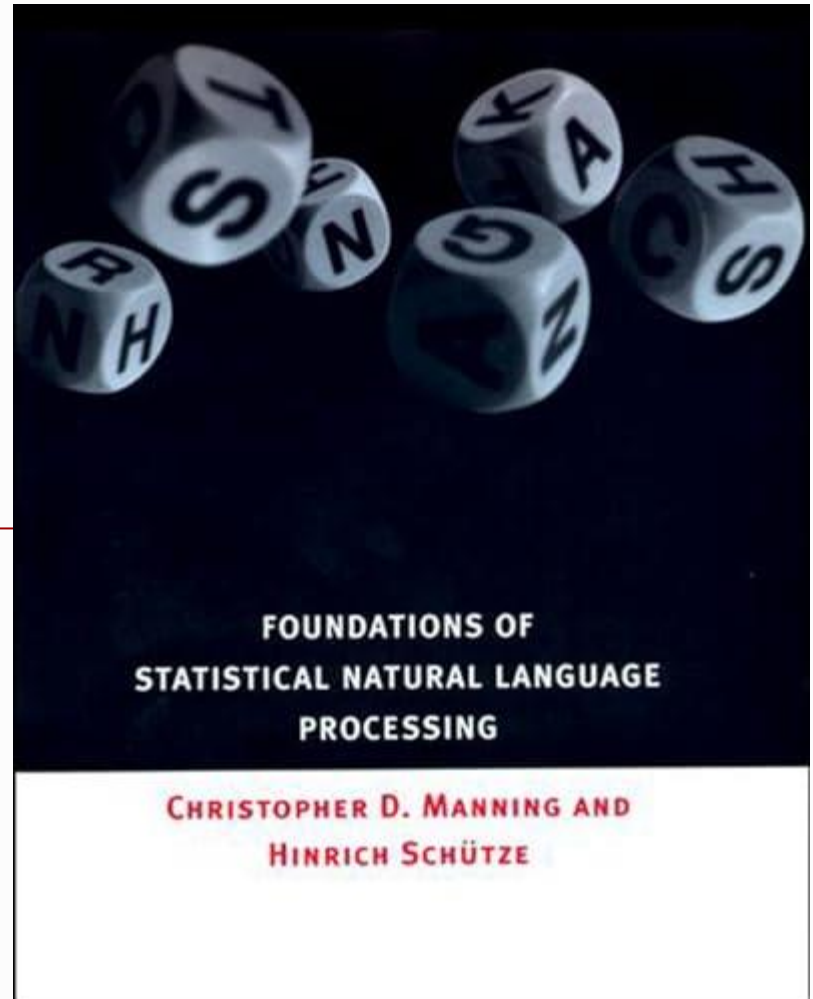
Ending your session

Be sure to save your work before ending the session

1. Select **File > Hub Control Panel**
2. Click **Stop My Server**



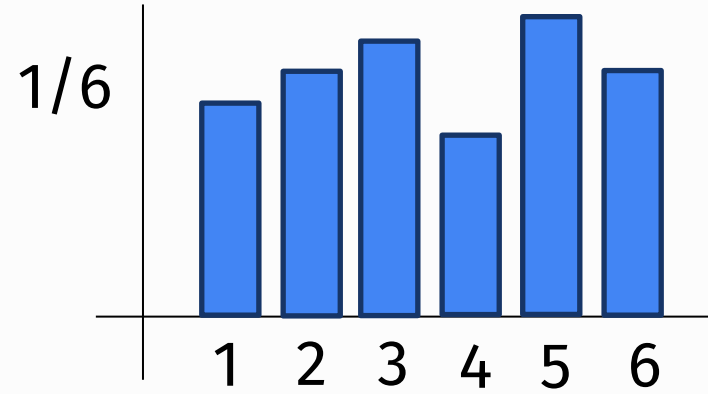
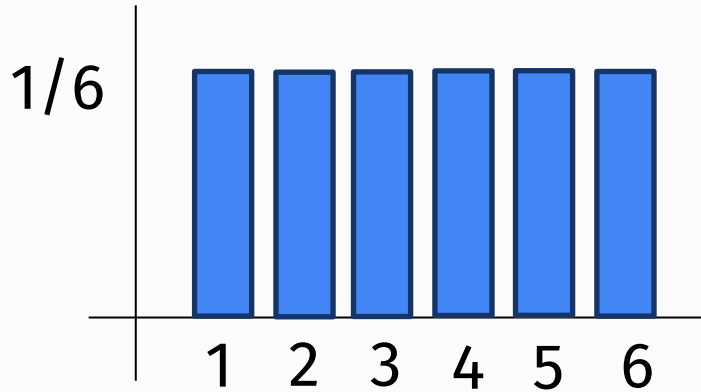
Probability review



Probability

- Probability of an event a occurring
- $P(a)$
 - For example, a could be a die showing a 2 out of $\{1, 2, 3, 4, 5, 6\}$
- Estimate $P(a)$ as $\frac{\text{count}(a)}{\text{count}(\text{all events})}$
 - Relative frequency or maximum likelihood estimate (MLE)

Probability distributions



Random variables

- **Random variable:** a mapping from a domain of possible outcomes in a sample space to a range of measurable space, such as counts
 - Typically the “result of an experiment”
 - For example, flipping a coin multiple times (possible outcomes {H, T}) and recording the result as 0 for tails and 1 for heads
- Distribution of a random variable X
 - $P(X)$ is a probability distribution over all possible values in the sample space. Probability mass function
 - $P(X = x)$ is the probability that the random variable X has the value x
 - $P(X = \text{heads})$, where X is the random variable of a coin flip

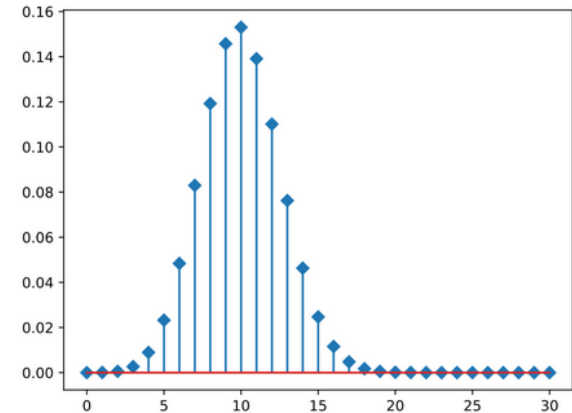


Figure 7.1: $P(k \text{ heads})$ in 30 tosses, success prob $1/3$.

Joint probability

- Probability of 2 events both occurring

$$P(A \cap B)$$

$$P(A, B)$$

- When rolling 2 dice, what's the probability of getting two 5s?

Let D_1 be dice 1, D_2 be dice 2. These events are independent, so:

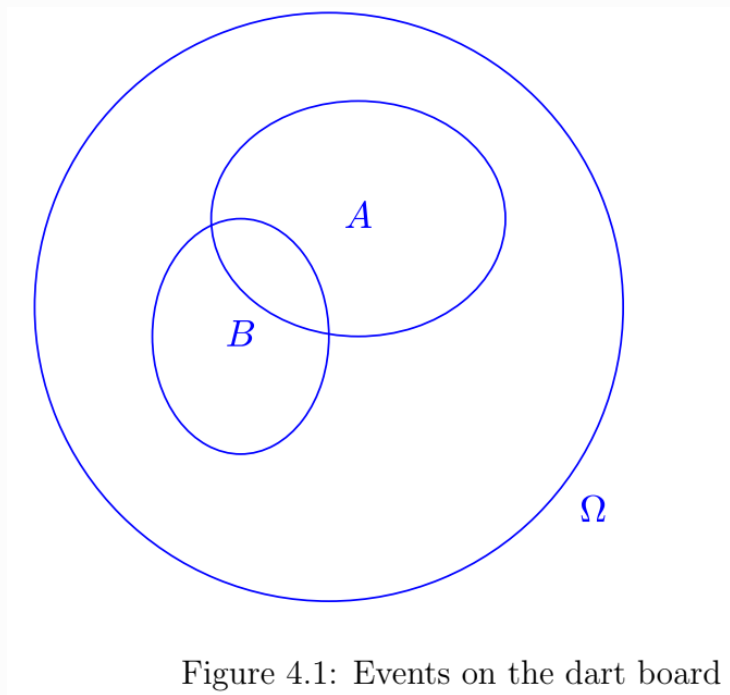
$$P(D_1 = 5, D_2 = 5) = P(D_1 = 5) \cdot P(D_2 = 5)$$

$$\frac{1}{6} \cdot \frac{1}{6} = \frac{1}{36} \text{ since there are 36 different possible combinations}$$

Conditional probability

- Probability distributions sometimes change if you know another event has occurred or not occurred
- **Conditional** probability of an event a occurring **given that another event, b , has already occurred**
 - $P(a|b)$
- Assume
 - X is the outcome of rolling a die once
 - F is the event $X = 6$
 - E is the event $X > 4$
- Die is rolled and we are told that E has occurred
- What is $P(F|E)$?

Conditional probability



- Assume a very bad dart thrower (maybe Michael)

$$\mathbf{P}(A) = \frac{\mathbf{area}(A)}{\mathbf{area}(\Omega)}$$

Conditional probability

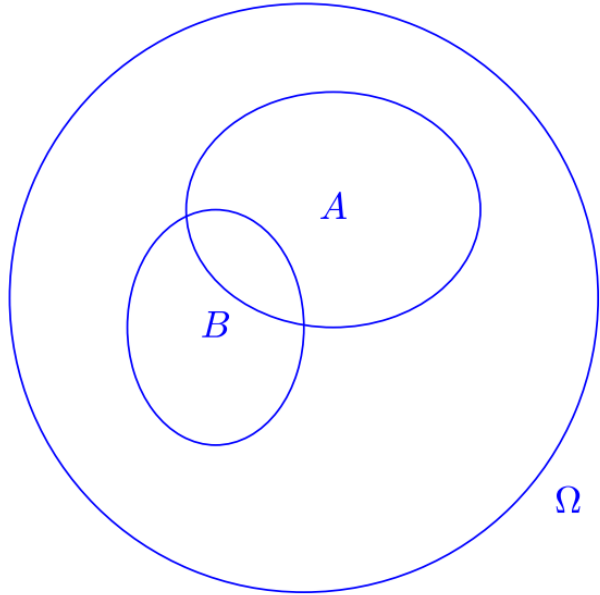


Figure 4.1: Events on the dart board

- You don't see the throw, but somebody tells you that the dart landed in B (so B occurred)
- What is the formula for $P(A|B)$?

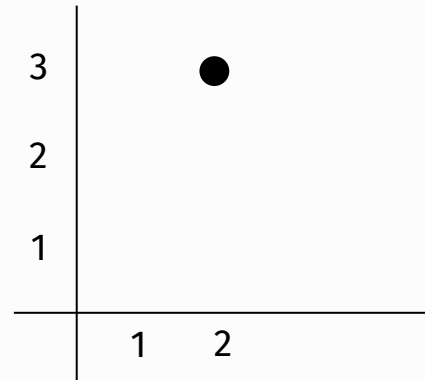
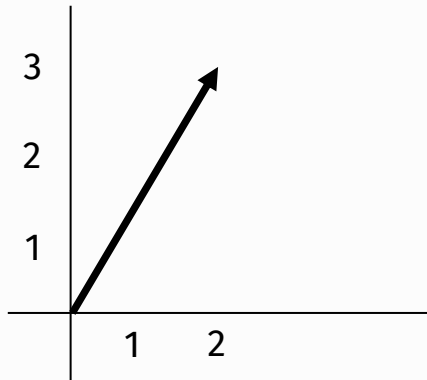
Linear algebra review

Vectors

An array of numbers with D dimensions

[2 3]

Can be represented as a point in D -dimensional space



Dot product: vector \cdot vector

Sum of the products of each vector dimension

$$\mathbf{V} \cdot \mathbf{W}$$

Diagram illustrating the dot product of two vectors \mathbf{V} and \mathbf{W} . Vector \mathbf{V} is represented by a horizontal blue bar containing the components V_1 , V_2 , \dots , and V_N . Vector \mathbf{W} is represented by a vertical green bar containing the components W_1 , W_2 , \vdots , and W_N . A dot \cdot is placed between the two vectors.

$$\mathbf{V} \cdot \mathbf{W} = \sum_{i=1}^N V_i W_i = V_1 W_1 + V_2 W_2 + \dots + V_N W_N$$

Matrices

A matrix is an array of numbers

$$\begin{bmatrix} 6 & 4 & 24 \\ 1 & -9 & 8 \end{bmatrix}$$

Two rows, three columns.

It's Easy to Multiply a Matrix by a Scalar

$$2 \cdot \begin{bmatrix} 5 & 2 \\ 3 & 1 \end{bmatrix} = \begin{bmatrix} 2 \cdot 5 & 2 \cdot 2 \\ 2 \cdot 3 & 2 \cdot 1 \end{bmatrix} = \begin{bmatrix} 10 & 4 \\ 6 & 2 \end{bmatrix}$$

Dot product: vector \cdot matrix

$$\begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} \\ \\ \end{bmatrix}$$

Dot product: matrix · matrix

Let a_1 and a_2 be the row vectors of matrix A and b_1 and b_2 be the column vectors of a matrix B. Find $C = AB$

$$\begin{bmatrix} 1 & 7 \\ 2 & 4 \end{bmatrix} \cdot \begin{bmatrix} 3 & 3 \\ 5 & 2 \end{bmatrix} = \begin{bmatrix} a_1 \cdot b_1 & a_1 \cdot b_2 \\ a_2 \cdot b_1 & a_2 \cdot b_2 \end{bmatrix} = \begin{bmatrix} 38 & 17 \\ 26 & 14 \end{bmatrix}$$

A must have the same number of rows as B has columns.

Questions?

No class next Mon for MLK Day.

Will see you again on Wed.

Take a look at HW1