Discussion 7

COGS 108 Summer Session 2023

IA: Antara Sengupta
TAs: Hari Yadavalli, Rounak Sen, Abhishek Tanpure
Instructor: C. Alex Simpkins Ph.D.

Agenda for today

- Upcoming due dates/announcements
- Brief ML content review
- D7 Lab Notebook

****Today's discussion can be really important if you want to incorporate ML into your project but aren't super familiar with it!**

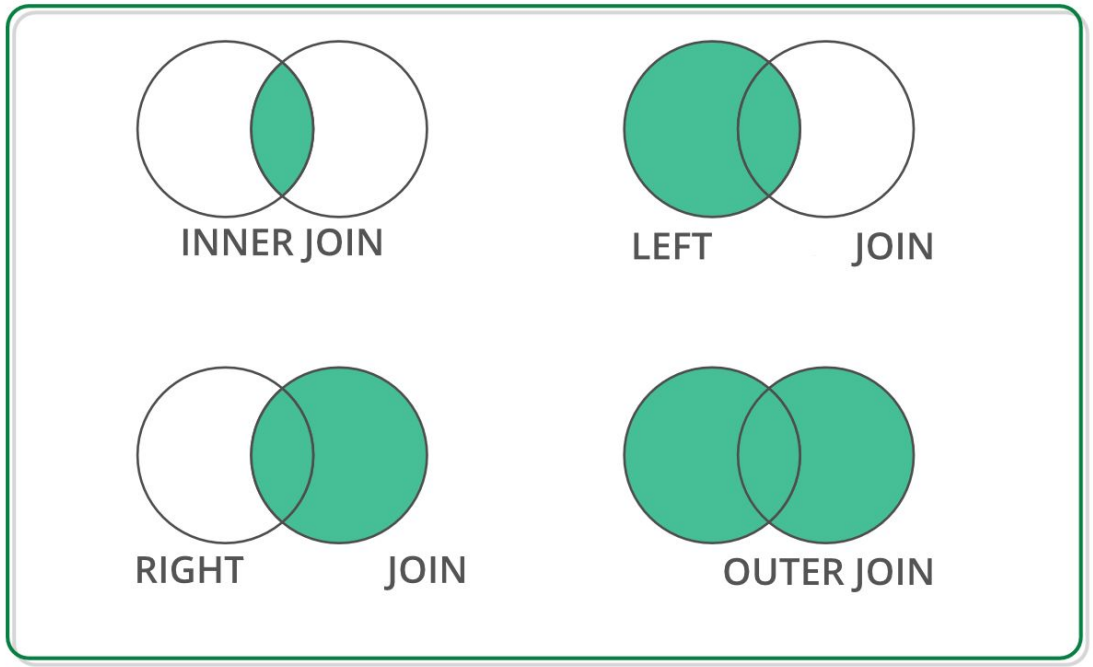
Due Dates & Announcements

- Data checkpoints graded - issues on GitHub
- Due dates for D5, D6, A3 and EDA Checkpoint extended to tonight, **July 31st 11:59pm**
- Quiz 3 Due Tomorrow
- Checkpoint collection form for extra credit submission tomorrow
- Final Project & Video Submission this weekend
- Other due dates TBA soon!
- As we wrap up the quarter, do not hesitate to reach out for help/with questions if any concepts seem fuzzy, or if you need project feedback

BRIEF CONTENT REVIEW

Pandas Merge

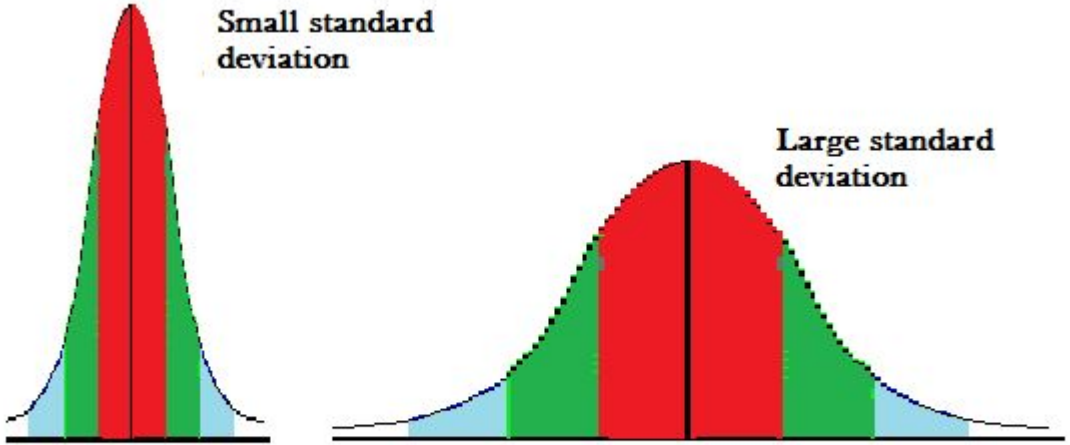
- In this class, we use inner, outer, left, and right
- Syntax for a left merge will look like: pd.merge(df1,df2, how = ‘left’)



Quick Terminology Refresher

- **Standard Deviation**: measure of data spread, showing how much individual data points deviate from the average. Smaller values indicate data points are closer to the mean, while larger values mean greater spread.
- **Variance**: measure of data dispersion, representing the average of squared differences between data points and the mean. It provides insights into how data is spread out around the mean.

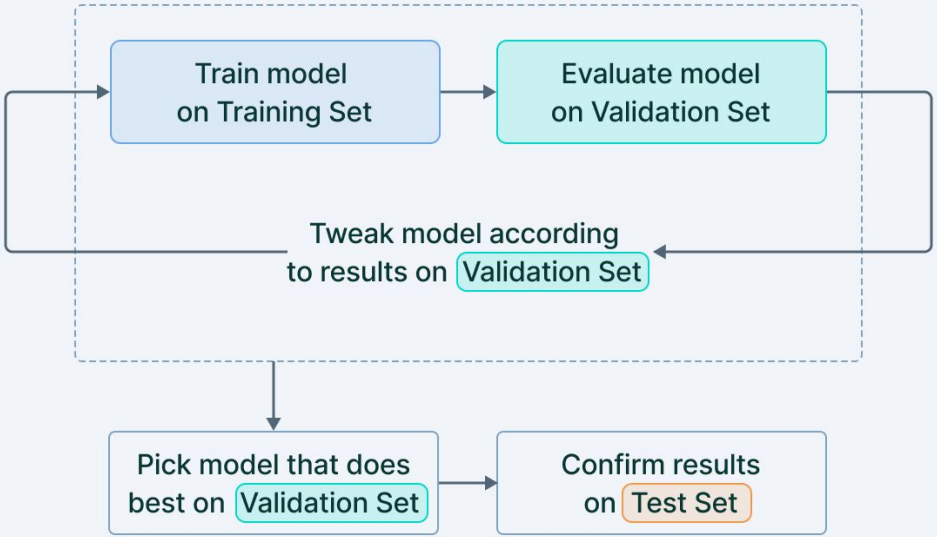
Standard Deviation Formula	
Population	**Sample**
$\sigma = \sqrt{\frac{\sum(X-\mu)^2}{N}}$	$s = \sqrt{\frac{\sum(X-\bar{x})^2}{n-1}}$
X – The Value in the data distribution *μ – The population Mean* *N – Total Number of Observations*	*X – The Value in the data distribution* *$\bar{x}$ – The Sample Mean* *n - Total Number of Observations*


Small standard deviation
Large standard deviation

ML Basics: Training, Validation & Testing Set

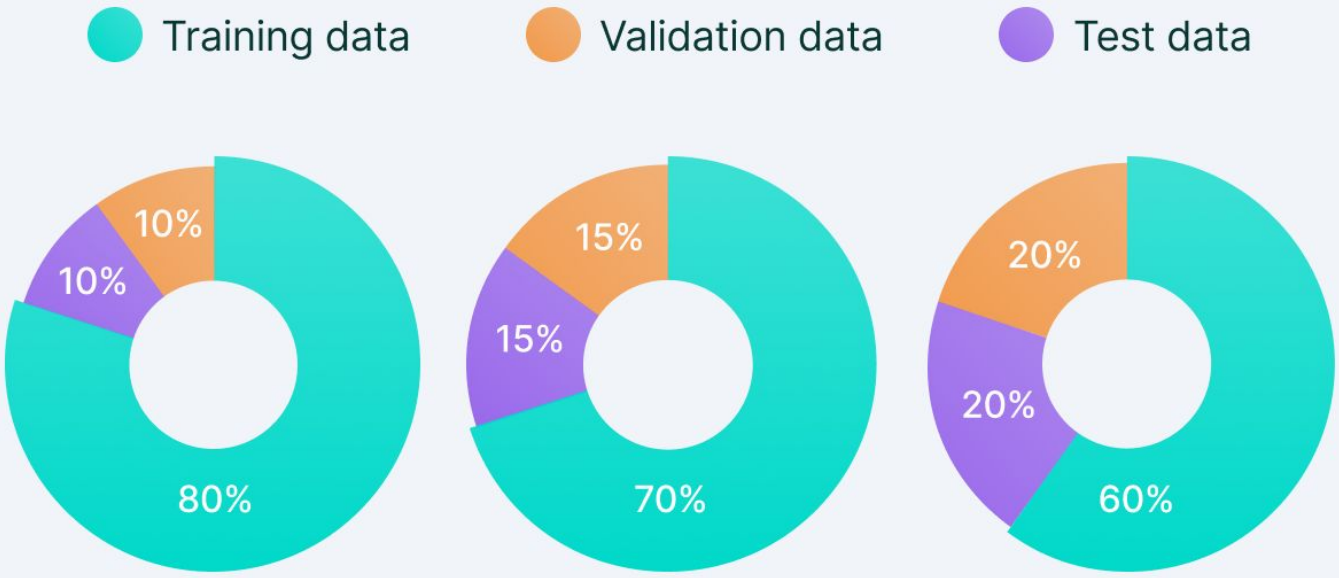
- In ML, we break our data into training, testing and validation (sometimes) sets
- **Training set**: The data that we give/provide to our model to learn from
    - Our model gets “trained” on this data, and uses it to pick up on trends and patterns so that when it is given new data, it can learn from it
- **Testing set**: the data we set aside so that we can test our model on it and see how accurately our model is able to perform, make predictions
- We will cover more general & simple ML training/testing in today’s notebook + this class, but you will learn more about validation in other COGS ML courses (or other departments)

Training data/validation/test



V7 Labs

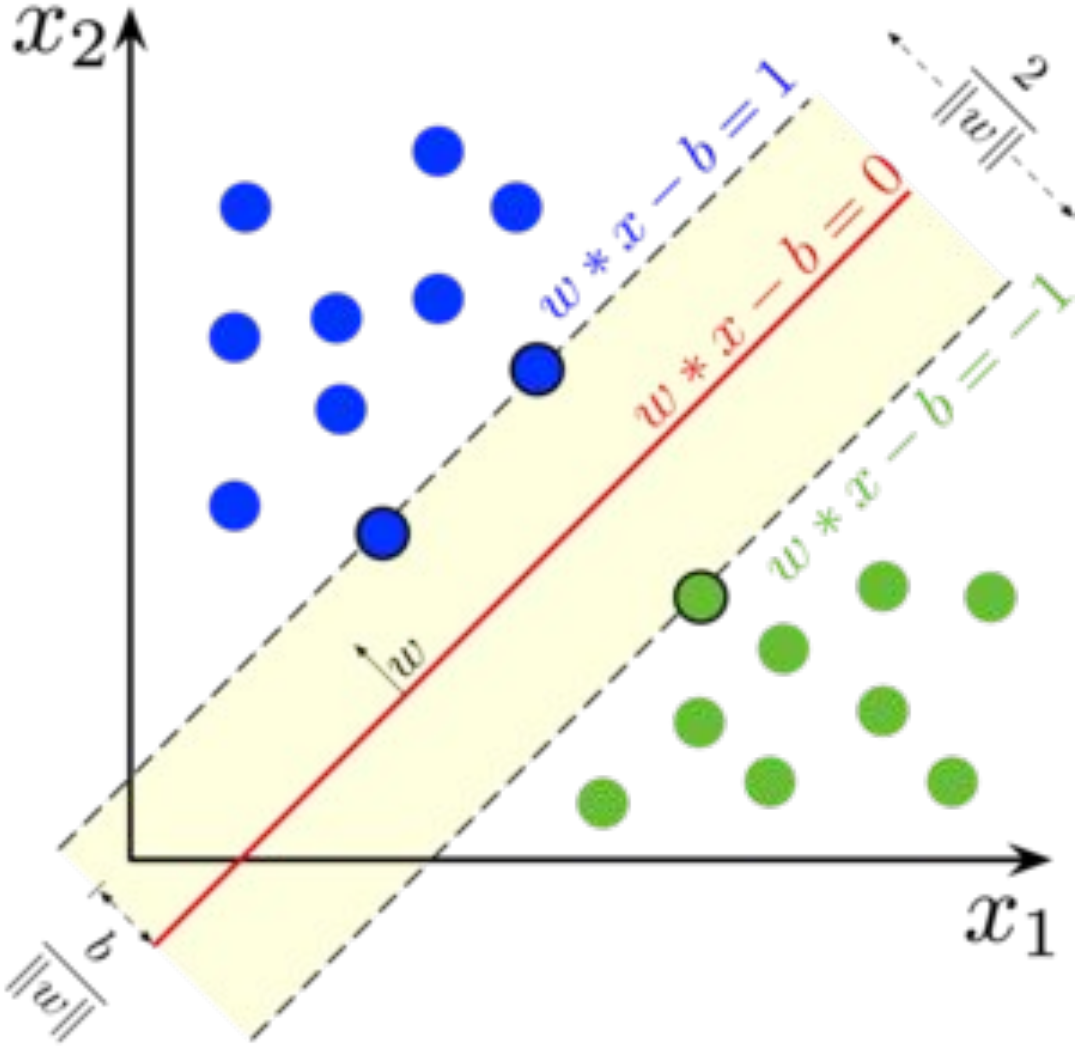
Data Training Needs



V7 Labs

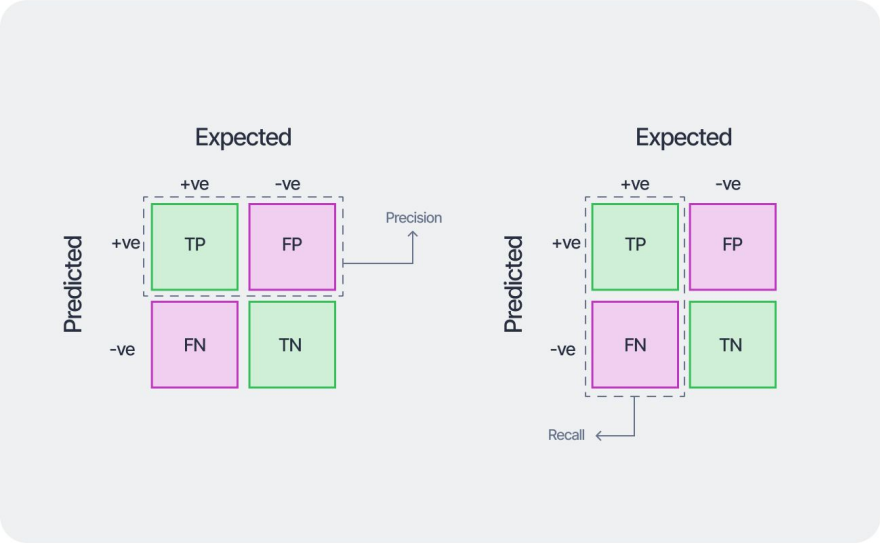
Support Vector Machine Classifier

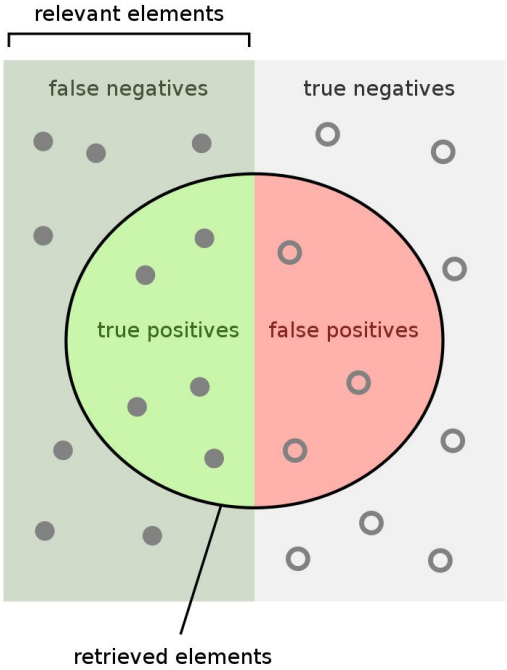
- https://scikit-learn.org/stable/modules/svm.html
- SVM: works by selecting the hyperplane that maximizes the margin between the two classes, which is the distance between the closest data points from each class to the hyperplane


x_2
$w * x - b = 1$
$w * x - b = 0$
$w * x - b = -1$
$\frac{2}{\|w\|}$
w
$\frac{b}{\|w\|}$
x_1

Precision, Recall, F1-Score, Confusion Matrix

- Used to measure accuracy of classification models





How many retrieved items are relevant?

Precision =

How many relevant items are retrieved?

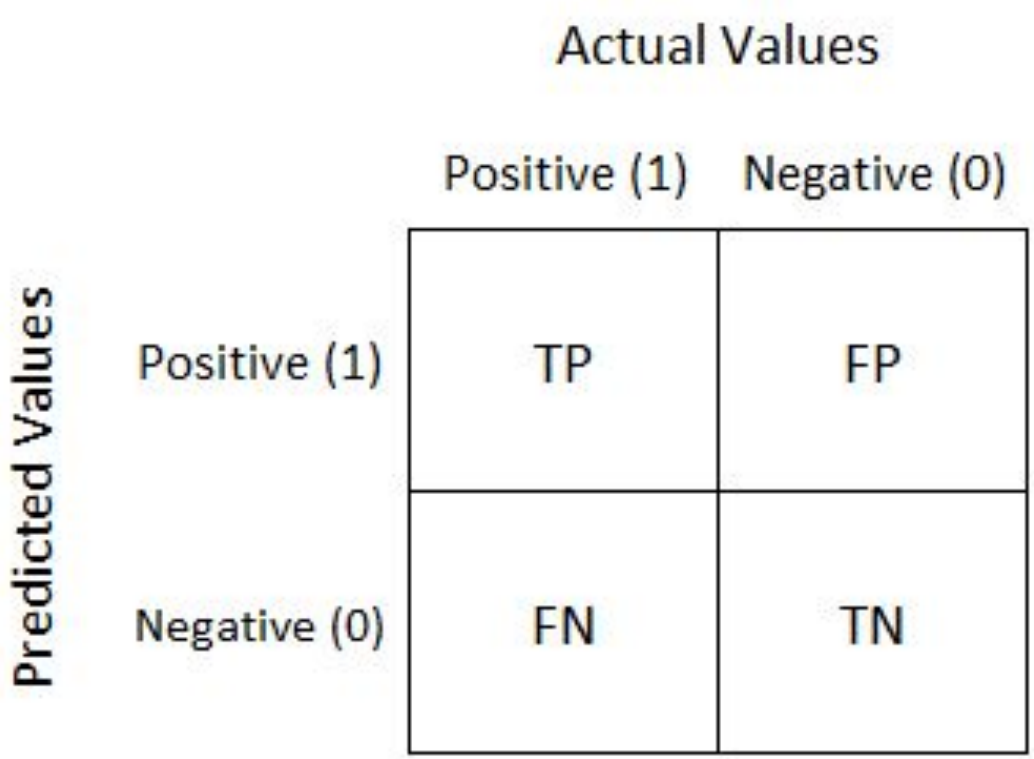
Recall =

$$(10.1)\quad Accuracy = \frac{T_p + T_n}{T_p + T_n + F_p + F_n}$$

$$(10.2)\quad Precision = \frac{T_p}{T_p + F_p}$$

$$(10.3)\quad Recall = \frac{T_p}{T_p + T_n}$$

$$(10.4)\quad F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$


Actual Values
Positive (1)
Negative (0)
Predicted Values
Positive (1)
Negative (0)
TP
FP
FN
TN

SciKit Learn is your best friend

- Very detailed explanation of all concepts and code
- Keep open when you code to refer to
- https://scikit-learn.org/stable/index.html

Thank you for coming, good luck on the final project!