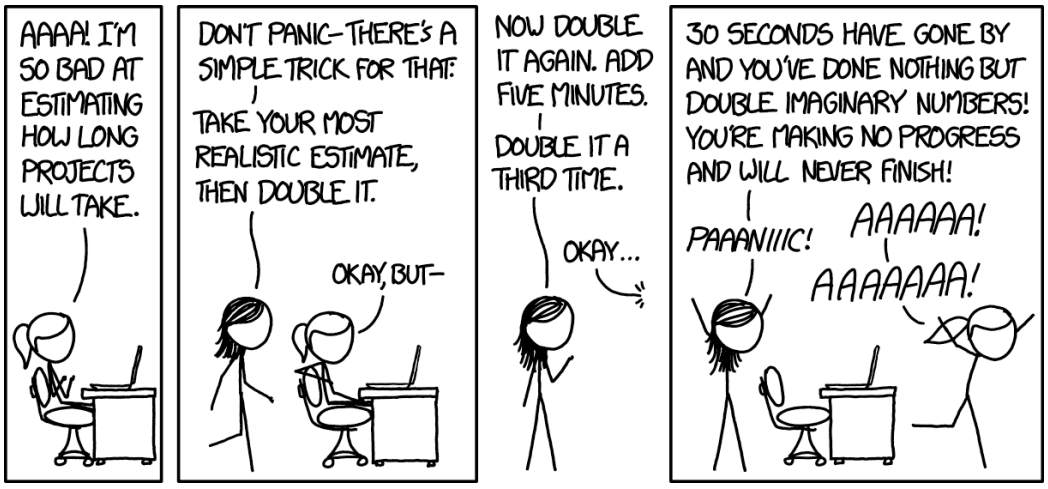
CSCI 3202

Lecture 32

November 10, 2025



xkcd by Randall Munroe. https://xkcd.com

Announcements

- HW #9 will be released on Friday
- Project due next week on Monday, November 17
  - I will post a location on Canvas to turn in your project
- Interview grading starting on Wednesday, November 19
  - We will have a sign up for times
  - 15 minute interviews
  - Interviews are done on Zoom
  - Share your IDE
  - Run your project and answer questions
- We will do interview grading after fall break as well
- You may resubmit Worked Out Problems (Q13-Q20) only from the midterm for regrading. The points you receive will replace your original score on the midterm. If you do not submit anything, there will be no change in your score. You have until Friday November 21 to submit. No late submissions will be accepted
- Before calculating your final grade, I will add 5 points to your midterm exam score due to the length of the exam

- These points will only count towards your midterm score. They will not count as extra credit toward your course grade

Quiz #9

- Average is 87%
- Review

Machine Learning

- Machine learning is a field of computer science that focuses on building systems that can **learn from data** and **make decisions** or predictions based on that data, without being explicitly programmed for every specific task.
- In traditional programming, you write a set of instructions for the computer to follow. In machine learning, instead of telling the machine exactly what to do step by step, you provide it with data and let it figure out patterns, structures, or decision rules. The machine "learns" from the data and improves its performance over time

Types of Machine Learning

- Machine learning can be broadly classified into **three categories**:
  - **Supervised Learning**
  - **Unsupervised Learning**
  - **Reinforcement Learning**

1. Supervised Learning

Supervised learning is the most common type of machine learning and is used when we have labeled data. In supervised learning, the algorithm learns from a **training set** where both the input data (features) and the correct outputs (labels) are provided.

Key Steps in Supervised Learning:

1. **Training**: The model is trained on labeled data, learning the relationship between input features and the corresponding output.
2. **Prediction**: Once trained, the model can predict the output for new, unseen data based on what it learned during training.
3. **Evaluation**: The model's performance is evaluated using metrics like accuracy, precision, recall, or F1-score.

Example Problem: Predicting whether an email is spam or not. The model is trained on a dataset of emails (input) and whether each email is labeled as "spam" or "not spam" (output).

Common Algorithms in Supervised Learning:

- **Linear Regression** (for predicting continuous values)
- **Logistic Regression** (for binary classification)
- **Decision Trees**
- **Support Vector Machines (SVM)**
- **Neural Networks**

2. Unsupervised Learning

Unsupervised learning, in contrast, is used when the data does **not have labels**. The goal here is to identify hidden patterns or structures in the data without specific output values to guide the learning process.

Key Steps in Unsupervised Learning:

1. **Data exploration**: The model looks at the data and tries to find inherent structures, clusters, or associations.
2. **Pattern discovery**: The algorithm uncovers similarities or groupings within the data, such as identifying clusters of similar items.

Example Problem: Segmenting customers into different groups based on purchasing behavior. The model isn't told in advance what the groups should be, but it tries to discover distinct clusters of customer profiles.

Common Algorithms in Unsupervised Learning:

- **K-Means Clustering**
- **Hierarchical Clustering**
- **Principal Component Analysis (PCA)**
- **Association Rule Learning**

Decision Trees

A **decision tree** is a flowchart-like structure used to make decisions or predictions. Each internal node represents a "test" or "decision" based on an attribute (feature), and each branch represents an outcome of that test. The leaves of the tree contain the predictions (class labels for classification or values for regression).

How Decision Trees Work:

- Starting from the root node, the algorithm splits the data based on the feature that provides the best separation (e.g., minimizing variance or maximizing information gain).
- This process continues recursively, creating subtrees, until the data is perfectly separated, or some stopping criterion is reached (like a maximum depth or minimum number of samples per leaf).

Example: Suppose we want to predict whether a person buys a product based on their age and income level. The decision tree might first split the data by age, then further split by income, leading to a prediction about whether the person will buy the product.

Pros of Decision Trees:

- Easy to understand and visualize.
- Can handle both numerical and categorical data.
- Non-linear, so can capture complex patterns.

Cons of Decision Trees:

- Prone to overfitting if the tree is too deep.
- Sensitive to small variations in the data.

Key Differences Between Supervised and Unsupervised Learning:

Feature	Supervised Learning	Unsupervised Learning
Data	Requires labeled data (input + output)	Uses unlabeled data (input only)
Goal	Predict or classify data based on labels	Discover hidden patterns or structures in the data
Example Algorithms	Decision Trees, Logistic Regression, SVM	K-Means, PCA, Hierarchical Clustering
Output	Prediction of output labels	Groupings, clusters, or associations

Illustrations

- Decision Trees.pdf

Upcoming

- Machine Learning
- Decision Trees