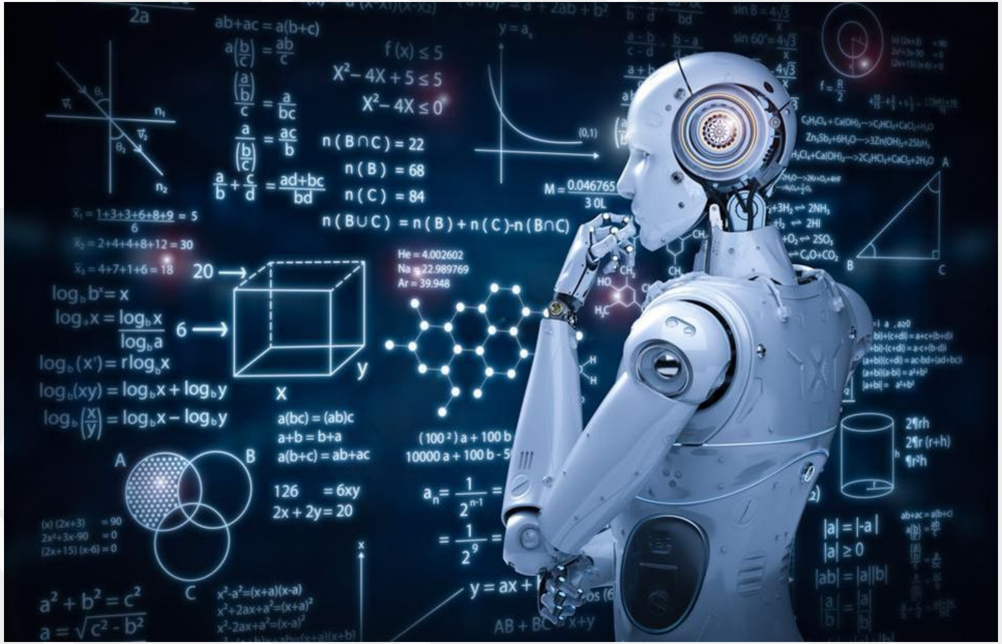
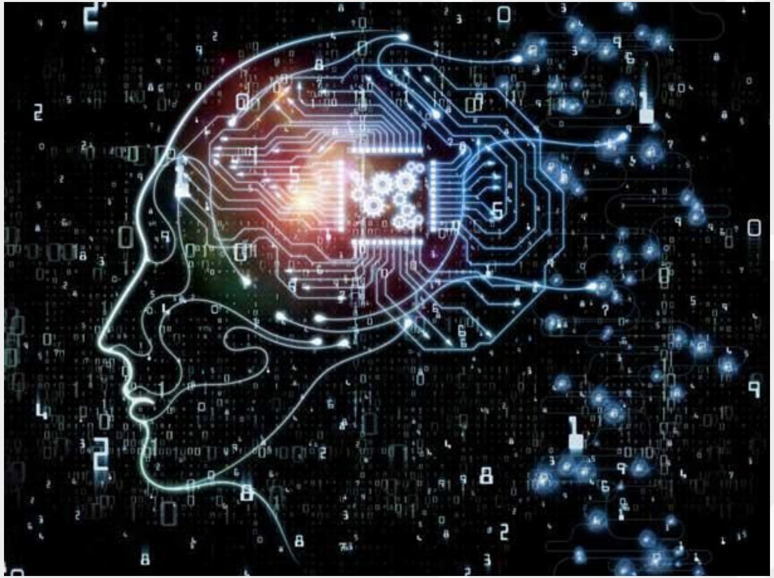
Harvard College Open Data Project

Roadmap

1. What is mAcHiNe LeArNiNg?
2. Supervised Learning
3. Unsupervised Learning

Machine Learning - Expectations





HODP

Machine Learning - Reality

Fitting Neural Networks

$$Z_m = \sigma(\alpha_{0m} + \alpha_m^T X),\ m = 1, \ldots, M,$$
$$T_k = \beta_{0k} + \beta_k^T Z,\ k = 1, \ldots, K,$$
$$f_k(X) = g_k(T),\ k = 1, \ldots, K,$$

Derivatives:

$$\frac{\partial R_i}{\partial \beta_{km}} = -2(y_{ik} - f_k(x_i))g_k'(\beta_k^T z_i)z_{mi},$$
$$\frac{\partial R_i}{\partial \alpha_{m\ell}} = -\sum_{k=1}^{K} 2(y_{ik} - f_k(x_i))g_k'(\beta_k^T z_i)\beta_{km}\sigma'(\alpha_m^T x_i)x_{i\ell}.$$

Descent along the gradient:

$$\beta_{km}^{(r+1)} = \beta_{km}^{(r)} - \gamma_r \sum_{i=1}^{N} \frac{\partial R_i}{\partial \beta_{km}^{(r)}},$$
$$\alpha_{m\ell}^{(r+1)} = \alpha_{m\ell}^{(r)} - \gamma_r \sum_{i=1}^{N} \frac{\partial R_i}{\partial \alpha_{m\ell}^{(r)}},$$

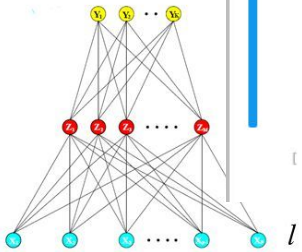


FIGURE 11.2. Schematic of a single hidden layer, feed-forward neural network.

γ_r :earning rate

i: observation index

first.ipynb

Code — Python 3

```
[1]: import tensorflow as tf
     import numpy as np
     from tensorflow import keras
```

```
---------------------------------------------------------------------------
ModuleNotFoundError                       Traceback (most recent call last)
<ipython-input-1-ca4d52c1c349> in <module>
----> 1 import tensorflow as tf
      2 import numpy as np
      3 from tensorflow import keras

ModuleNotFoundError: No module named 'tensorflow'
```

[ ]:

What is Machine Learning?

MIT 6.036: “Machine learning tries to design and understand computer programs that learn from experience for the purpose of prediction or control.”

CS 181: “Machine learning is the application of statistical, mathematical, and numerical techniques to derive some form of knowledge from data.”

Seth: “It’s math, CS, and statistics that people have given a fancy name.”

What can ML do?

- Summarize data
- Visualize data
- Create groups from data
- Predict using data

What Machine Learning ISN'T?

- Magic
- All deep learning
- Useful without first understanding your dataset (GIGO)
- A Replacement for Critical Thought

Speaking of which...

It’s YOUR job to use ML ethically

HODP

Supervised and Unsupervised Learning

2 Basic ML Settings

1. Supervised learning: Given a set of features X, <u>predict</u> an output Y
    a. Ex. Regression, Classification, etc.
2. Unsupervised learning: Given a set of features X, <u>discover patterns</u>
    a. Ex. Clustering, PCA, SVD, etc.

There's also a third method called **reinforcement learning** where our models learn to make sequential decisions in an environment to maximize overall rewards... but we have an hour and it's hard.

Motivating Examples

To Code

--> workshopNotebook.ipynb

HODP

Supervised Learning

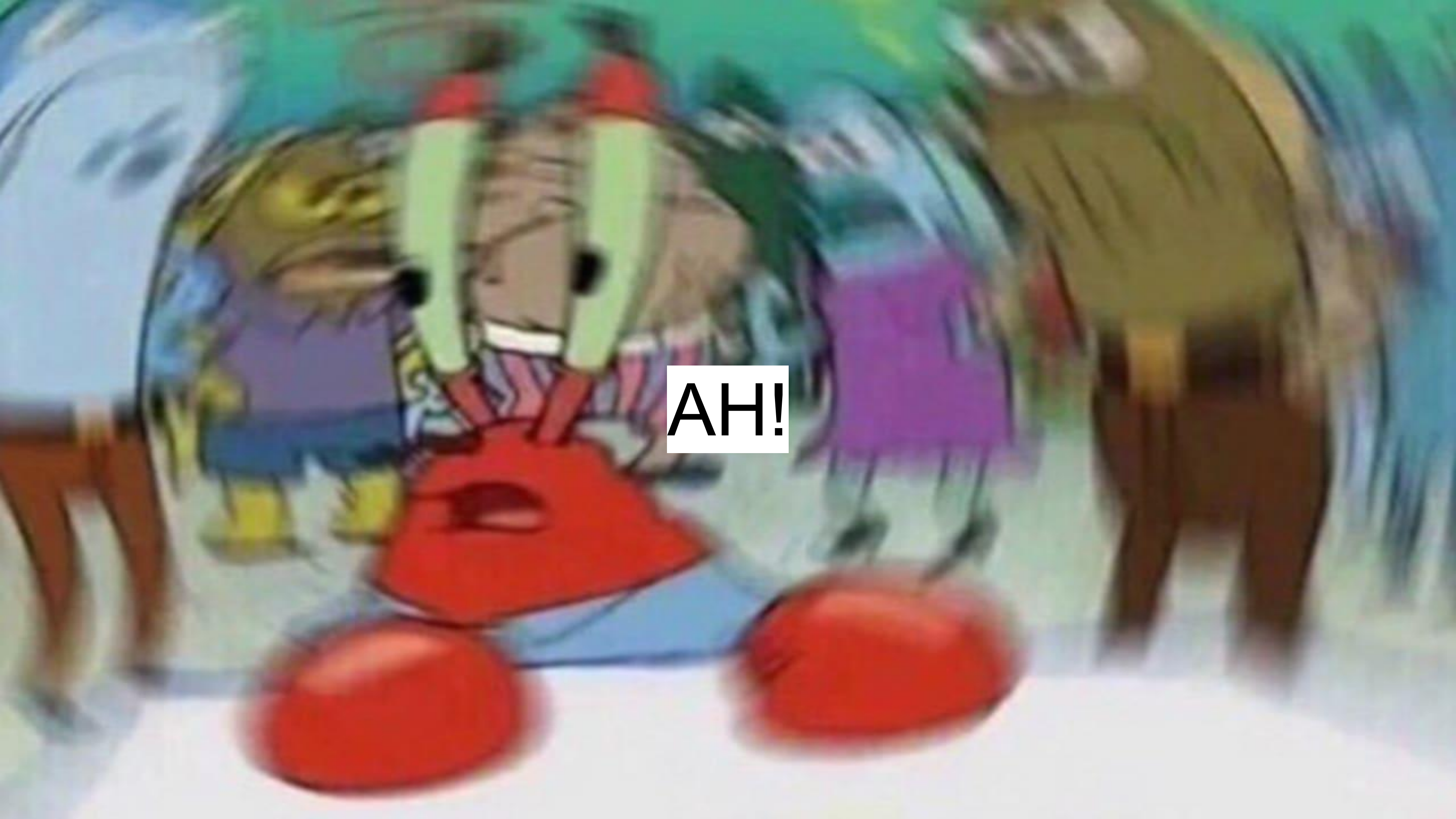
HODP

Supervised Learning

- Given a training dataset with features X and labels Y, learn about the relationship between X and Y. Then, make predictions on the test dataset.
- There are 2 main flavors of supervised learning that depend on the response variable Y.
  - Regression: Y is continuous
  - Classification: Y is categorical
- For now, let's focus on classification: Can we use demographic features to predict whether or not Akimel O'otham Native Americans have diabetes?

ML Models for Classification

- Classic Statistical Models
  - Logistic Regression (more specifically, multinomial logistic regression)
  - Linear Discriminant Analysis
- Non-parametric models
  - k-NN Classification
  - Decision Tree Classifiers
- Ensemble Models
  - Random Forest classifiers
  - Boosting Classifiers
  - Stacking any of the above Methods
- And more!
  - Feed-forward Neural Networks, Bayesian methods, etc.


AH!

Before you start fitting models...

- There are pros and cons to each model that depend on a number of factors.
- Here are some questions to answer **before fitting your models:**
  - **Goals and Interpretability:** What is the goal of your analysis? To what extent do you want to understand your prediction model? Is it okay to have accurate predictions even if you don't know how your model arrived at those predictions?
  - **Data Generating process:** What is the data generating process of your predictors? Of your response? Can you expect test data to be similar to your training data?
  - **Feature Engineering:** Should I transform features? Is my model robust to different transformations?
  - **Constraints:** How much data do you have? How much time do you have? How much computing power do you have?
- Answer these questions first! These outline the goals and constraints of your analysis.

Based on these answers, choose some models

- Example: Akimel O'otham Native Americans (Pima) dataset
    - I have a small dataset with 332 observations total and 265 training observations. I need models that do well with **small/medium-sized datasets**. This eliminates Neural networks/deep learning, and suggests that **k-NN classifiers** are a good place to start.
    - Since I'm dealing with medical data, I want to be able to interpret my results to stakeholders. Classical statistical methods like **logistic regression** seem like a good place to start.
    - I want to be able to detect non-linear relationships and interactions between my predictors. **decision tree-based models** are great for that!
- This requires an understanding of the pros and cons of models - READ, LEARN, RESEARCH

HODP

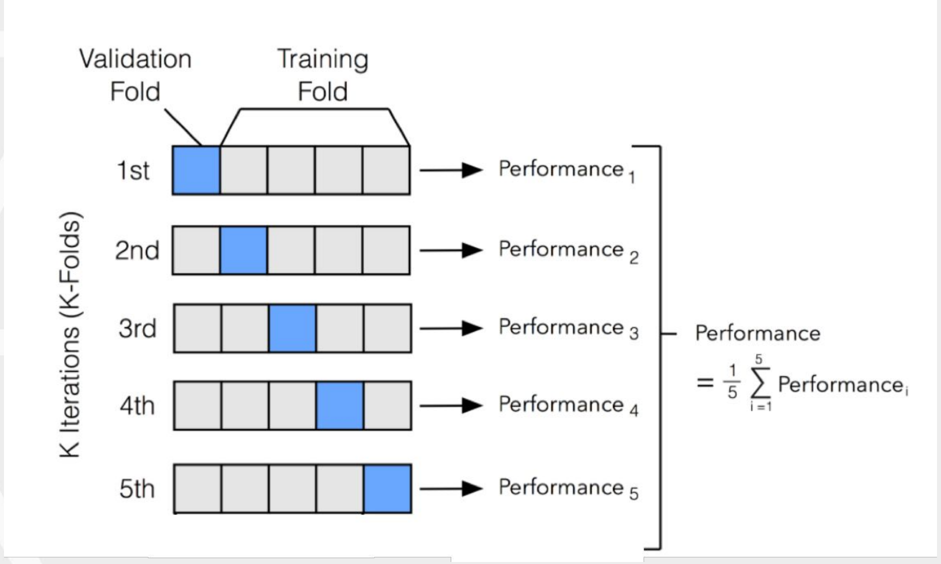
We’ve got candidate models – More Choices!

- There are lots of important choices that you can make when selecting models
- Loss/Cost/Objective functions
  - **Regression:** Mean squared error, Mean Absolute error
  - **Classification:** Cross entropy, classification error, 0-1 loss, etc.
  - **Regularization**: LASSO, Ridge, Elastic Net, etc.
- Feature Engineering
  - Normalize/Standardize your predictors or your response
  - How do you encode categorical predictors?
  - Are your models robust to transformations
- Optimization Algorithms
  - Analytical solutions, gradient descent, stochastic gradient descent, Newton-Raphson, etc.
- Evaluation metrics
  - Mean Squared Error, R^2, F1 score, classification error

HODP

Too Much Choice! What do we do?

- Use K-fold cross-validation to pick the best set of hyperparameters
- Then, compare model performance on the test set.



To Code

--> workshopNotebook.ipynb

Unsupervised Learning

Unsupervised Learning

- Given a dataset with features X and **no labels**, <u>discover patterns</u> in the data.
- Types of Unsupervised Learning include:
  - Clustering: Want to find groups
    - Ex. K-means, Gaussian Mixture models, DBSCAN, etc.
  - Dimension Reduction: Want to make X simpler
    - Ex. PCA, SVD, Autoencoders
- For now, let's focus on clustering: Given these features, can we figure out how many different types of flowers are in our dataset? Once we get a number of flowers, can we group observations into flower types using our predictors?


Model Selection again...

Model Selection

- Example: Iris dataset
    - I want to find groups based off of the features in my dataset. I should probably try clustering. **K-Means clustering** is always a good starting point for cluster analysis.
    - I also would like to quantify my uncertainty about how sure I should be that a point belongs to a given cluster. This leads me to **Gaussian Mixture Models.**
- Again, this requires an understanding of the pros and cons of models.

HODP

To Code

--> workshopNotebook.ipynb

HODP

Deep Learning

Deep Learning Models

- Deep Neural Networks can be used for supervised or unsupervised learning
- Though Neural Networks are complicated, the general framework for model selection/hyperparameter tuning is the same!
- Deep learning models tend to be extremely hard to interpret.
  - In practice, I like to use deep learning for things like data annotation and data augmentation.
  - It's harder to use deep learning in high stakes applications, but companies/consumers are getting increasingly more comfortable with it.
- P.S. Neural Networks are also not magic - lots of linear algebra and vector calculus though!

HODP

THANKS!

Any questions?

- sethbilliau@college.harvard.edu
- harvardopendataproject@gmail.com

HODP