CS 5043:
Advanced Machine Learning

Andrew H. Fagg
Symbiotic Computing Laboratory

What is Machine Learning?

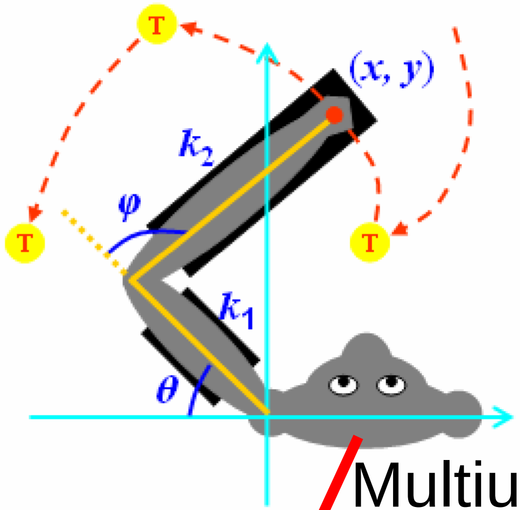
What is Machine Learning?

- Fundamentally: using data to automatically construct a model
- The model must be predictive!
  - I.E.: to be useful, it must produce meaningful output given novel situations.

Brain-Machine Interfaces

Command
prosthetic arm

Estimate of
intended
movement



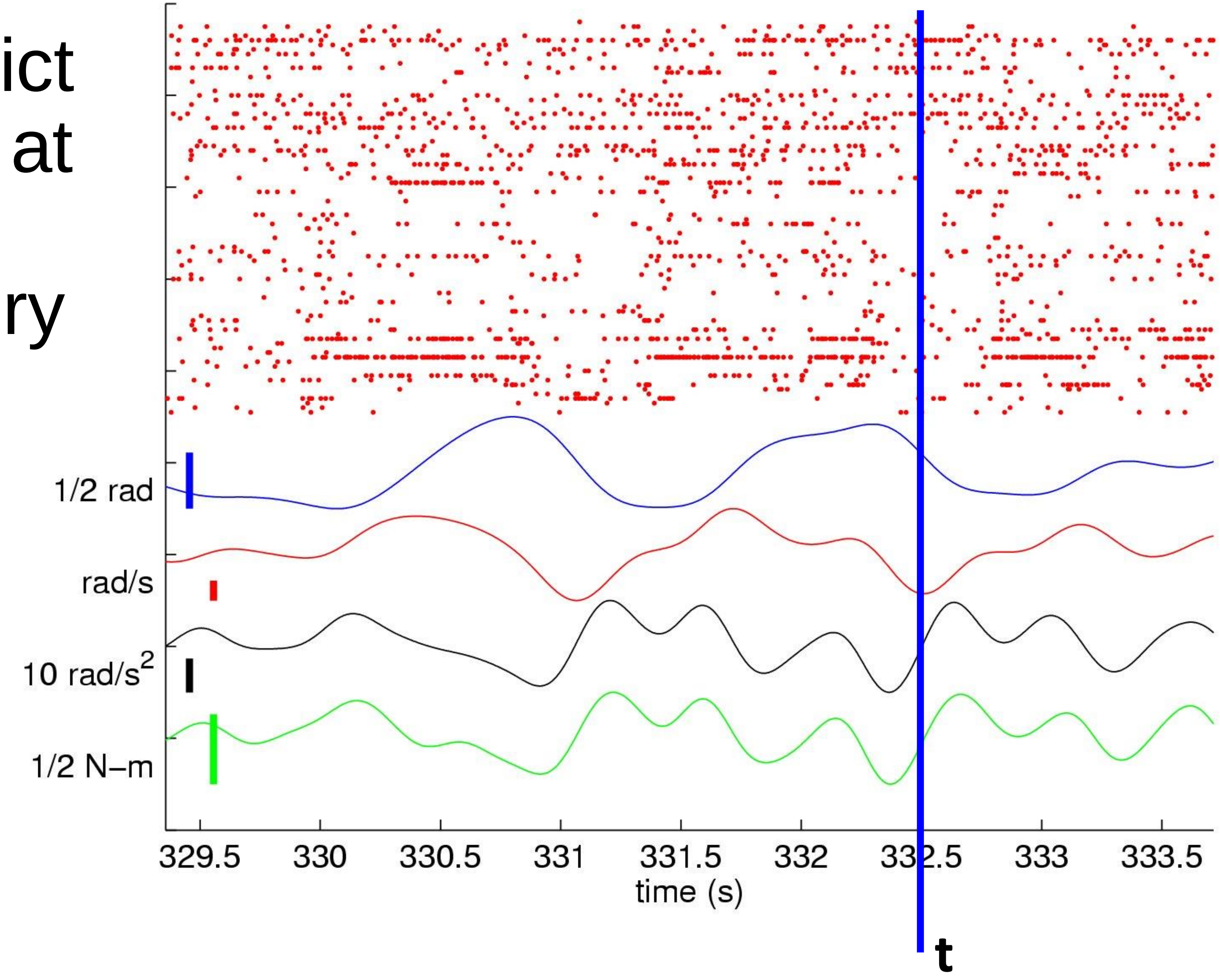
Multiunit
recording

Predictive
model

In collaboration with Nicholas G. Hatsopoulos and Lee E. Miller

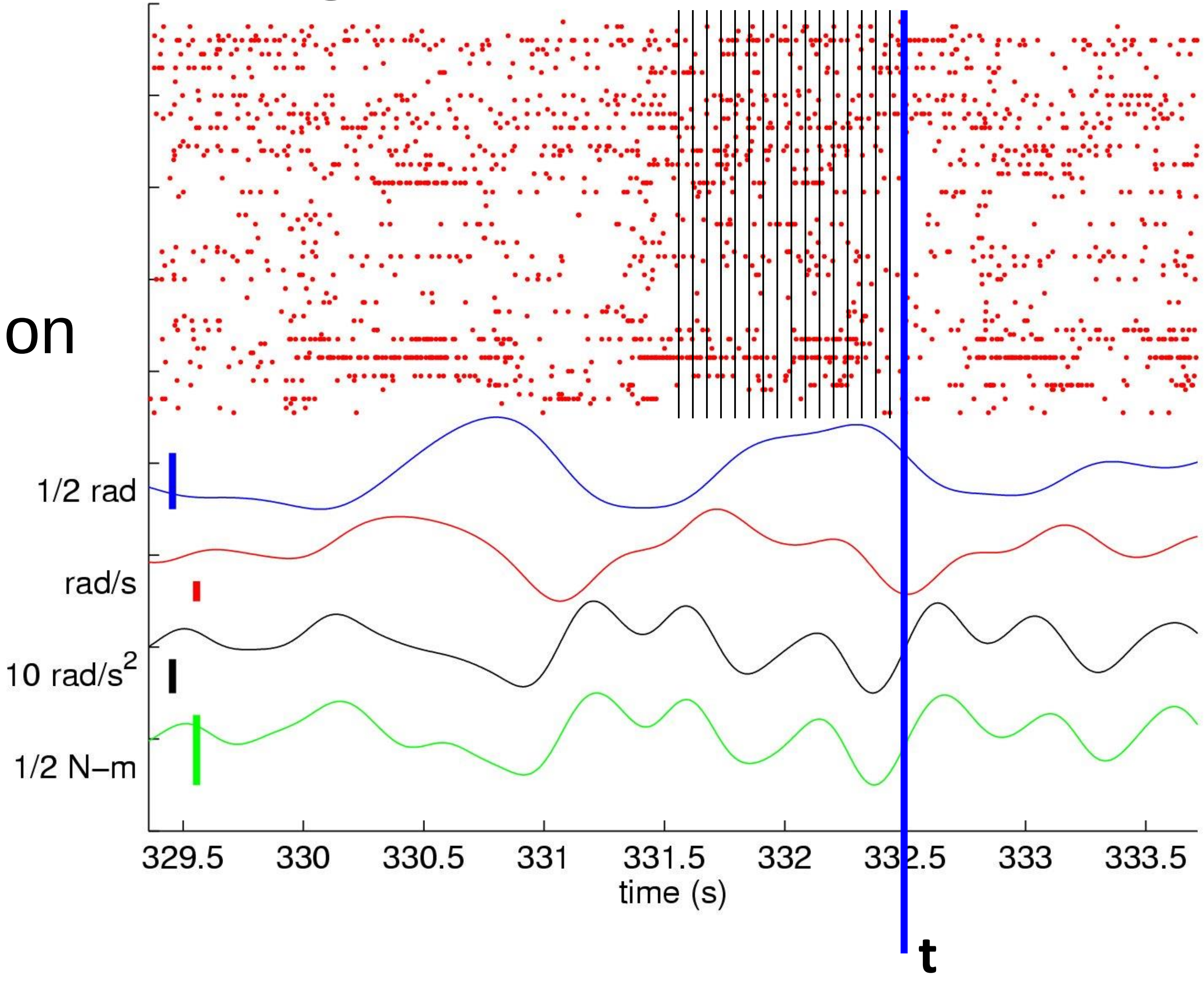
Decoding Arm State

- Want to predict arm motion at time t given recent history of spiking behavior



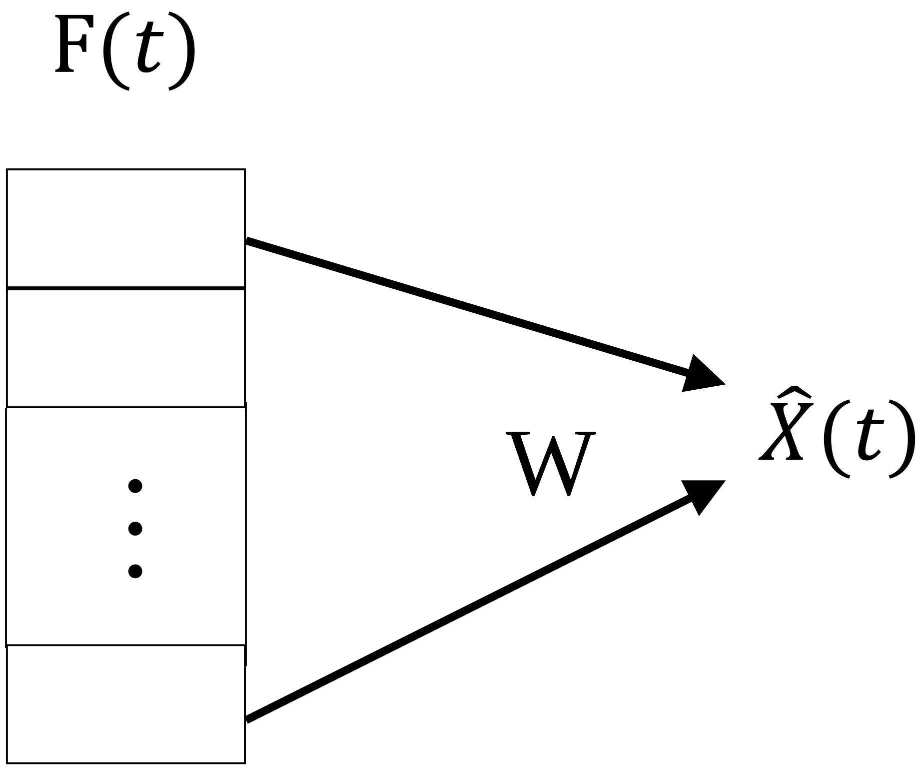
Decoding Arm State

50ms bins: 20 descriptors of neural activation for each cell



Wiener Filter

Each feature (F_i) is a count of spikes by a neuron for a 50 ms bin



$$\hat{X} = g_W\big(F(t)\big) = W^T F(t)$$

Column vector encoding spike counts for N cells at T taps up to time t

Data Types

Data Types

- Continuous
  - Probabilities
- Categorical (enumerated)
  - Binary
- Structured
  - Bind together arbitrary primitive data types
    - A class in an Object-Oriented sense does this
    - Heterogeneous objects are possible, too
  - Vectors, matrices

Data Types

- In some cases, we want to be able to acknowledge the fact that some data values are unknown or uncertain

Classes of Models

Classes of Models

Defined by the data type of the output (and possibly the input). Very broadly:

- Categorical output: classifier models
- Continuous output: regression-type models

Classes of Machine Learning Problems

Classes of Machine Learning Problems

Supervised learning

- Training set contains only input / output (labels) pairs
- Outputs could be continuous, probabilistic or categorical

Classes of Machine Learning Problems

Semi-Supervised learning

- Part of the training set contains input / output pairs
- The rest of the training set contains only inputs

- Using all of the data can yield a better model than if we only used the labeled data

Classes of Machine Learning Problems

Unsupervised learning

- The training set contains only inputs
- Fundamental question: what is the structure of these inputs?
  - Most common case: algorithm assigns categorical labels to each sample
  - But we can also ask continuous questions. For example: how different are two samples?

Classes of Machine Learning Problems

Reinforcement learning

- The training set contains inputs and an evaluation (reinforcement) of the output that is generated
- Most common case: a single evaluation can be a function of the sequence of outputs that is generated
  - How much time did it take to solve a problem?
  - How much energy did you use while solving the problem?
- Learning algorithm: for a given input, what is the output that maximizes the reinforcement over time?

Our Topics

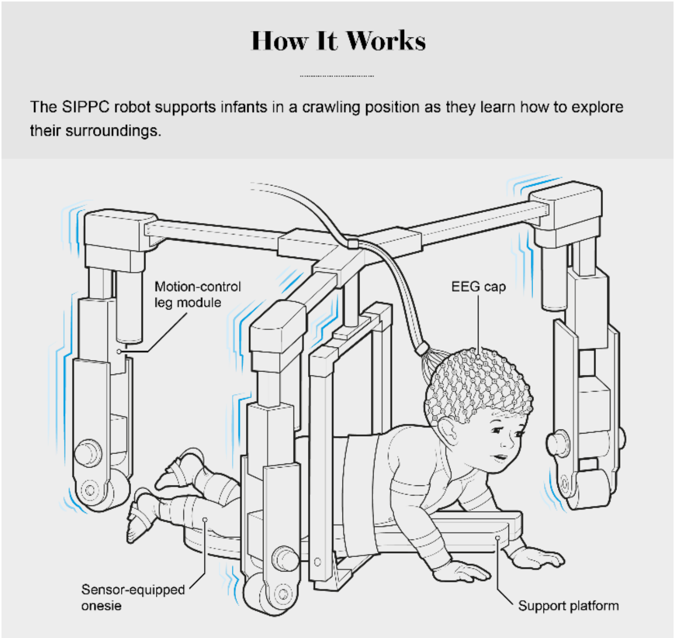
- Decision trees: ensemble methods, random forests, boosting
- Combatting overfitting with advanced regression: ridge regression, Tikhonov regression, lasso, elastic nets
- Dimensionality reduction: Kernel PCA, local linear embedding, ISOmap, multidimensional scaling
- Semi-supervised learning

Our Topics

- Deep Learning
  - Dropout and other normalization techniques
  - Convolutional neural networks
  - Recurrent neural networks
  - Autoencoders
- Evaluation in ML: metrics, cross-validation, statistics, addressing the multiple comparisons problem
- Reinforcement learning

Real-Time Activity Recognition for Assistive Robotics

OU Crawling Assistant
(Kolobe, Fagg, Miller, Ding)

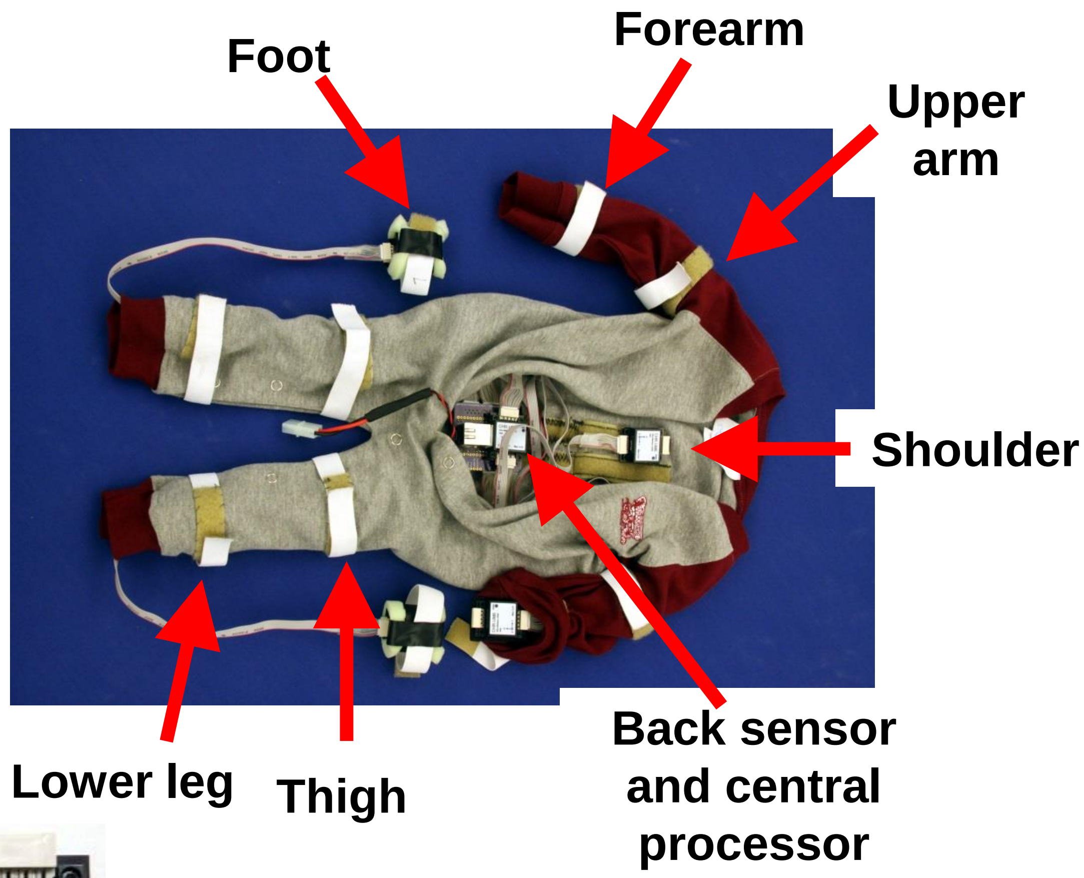


Scientific American (Oct 2016)

Kinematic Capture Suit

IMU-based kinematic suit

- 12 sensors mounted in suit
- Real-time reconstruction of body posture
- Recognition of crawling-like actions



Southerland (2012)

SIPPC Crawling Assistant

6-Axis
Load Cell

Wide-Angle
Cameras

Vertical
Lifts

EEG
Head
Net

Infant
Support

Covered Omni-
Wheels

Infant-Robot Interaction

Three modes of interaction:

- **Force control**: robot velocity is linearly related to ground reaction forces
- **Power steering**: small ground reaction forces produce a substantial robot movement
- **Gesture-based control**: recognized crawling-like movements produce robot movement

Machine Learning Questions

- Predict robot motion from kinematic data
- Predict visual attention from kinematic and robot data
- Predict limb motion from EEG data
- Predict visual attention from EEG data
- …

Other Challenges

Other Challenges

- Noisy data
- Unknown / invalid data values
- Data does not provide a complete “view” of the world
- Shift in the statistics of the data over time (non-stationarity)
- Small data sets
- Measuring model performance

To “Solve” a Model Building Problem We Must Answer:

- What is the nature of the data that we have?
- How much data do we have?
- What is the prediction problem?
- How do we measure performance of a model?
- How to select an appropriate model and learning algorithm?
- How to choose parameters?
- How to convince ourselves (and others) that we have a useful model?

What I am assuming about you…

- Programming skills
- Able to jump into Python, including the “Object-Orientedness” of it
- Know or can learn unix command-line tools

Resources

- Course web page:
  `http://www.cs.ou.edu/~fagg/classes/mlfds`
- Text: Aurélien Géron (2017) *Hands-On Machine Learning with Scikit-Learn and TensorFlow (Concepts, Tools, and Techniques to Build Intelligent Systems),* O'Reilly Media
- May also be useful: Nikhil Buduma (2017) *Fundamentals of Deep Learning*, O'Reilly Media
- Web resources: documentation, tutorials, papers (linked from the schedule or announced on Canvas)

Computing Environment

Setting up a ML environment (especially one based on TensorFlow) can be a bear …

- We are providing a pre-configured compute cluster on Amazon Web Services (AWS)
- Key tools: Python, Scikit-learn, TensorFlow (Deep Learning), Jupyter (Interactive Development Environment)
- Other software: editors (emacs, vi, gedit)
- Will also house our common data sets

Computing Environment

AWS machines are nice, but cost us based on the resources we use

- Right now, we have one AWS machine configured (1 processor; 2 GB of memory; 32 GB of swap).
- This machine will increase in resources as needed
- Additional machines will be added as needed
  - They will share a common user accounting system and file system
  - We will experiment with different machine configurations (some GPUs and AWS “Spot Instances”)

Configure your Cluster Account

- Install SSH on your local machine
- Generate a public key; email this and your desired username to me
- Once your account is created: SSH to the machine and configure Jupyter
  - The Jupyter server runs on the cluster
  - The front end runs in the browser of your local machine
  - A ***SSH tunnel*** allows the two to communicate securely

Homework Assignments

- First half of the semester
- Explore different ML methods and data sets

Projects

- Last half of the semester
- Topic / data set are your choice, but must be approved
- Several in-class presentations
- Final paper

Grading

- In-class participation: 10%
- Homework: 40%
- Project work: 50%

Proper Academic Conduct

- Homework assignments are to be done on your own
  - No communication of solutions in any form

- Projects: I am still deciding whether these will be group or individual projects

For Next Time

- Chapters 1 and 2 from the book
- We will discuss Pandas and start in on Scikit-learn

Advanced Machine Learning for Data Science

Software Tools

- Python
- Scikit-Learn
- XG-Boost
- TensorFlow

Reading

- Hands-On Machine Learning with Scikit-Learn & TensorFlow (O'Reilly)
- XG-Boost
- Online tutorials