Privacy and Data Mining

Seth Long

April 16, 2015

Overview

- Data Mining: Extracting useful information from data
- Machine Learing: Learn a way to make decisions
- Clustering: Group related items
- Needle in haystack problems
- There is often more than one way to look at a problem

Types and Security

- Duration: Session, Persistent, etc.
- First-party vs. Third-party cookies (ads)
- Secure cookies (for https sessions)
- Web browser cookie restrictions (usually not the default):
  - Reject all (kind of a headache)
  - Reject third-party cookies
  - Clear cookies on close

Sessions

- Cookie maintains session identifier
- Stolen cookies are a security vulnerability

Facebook

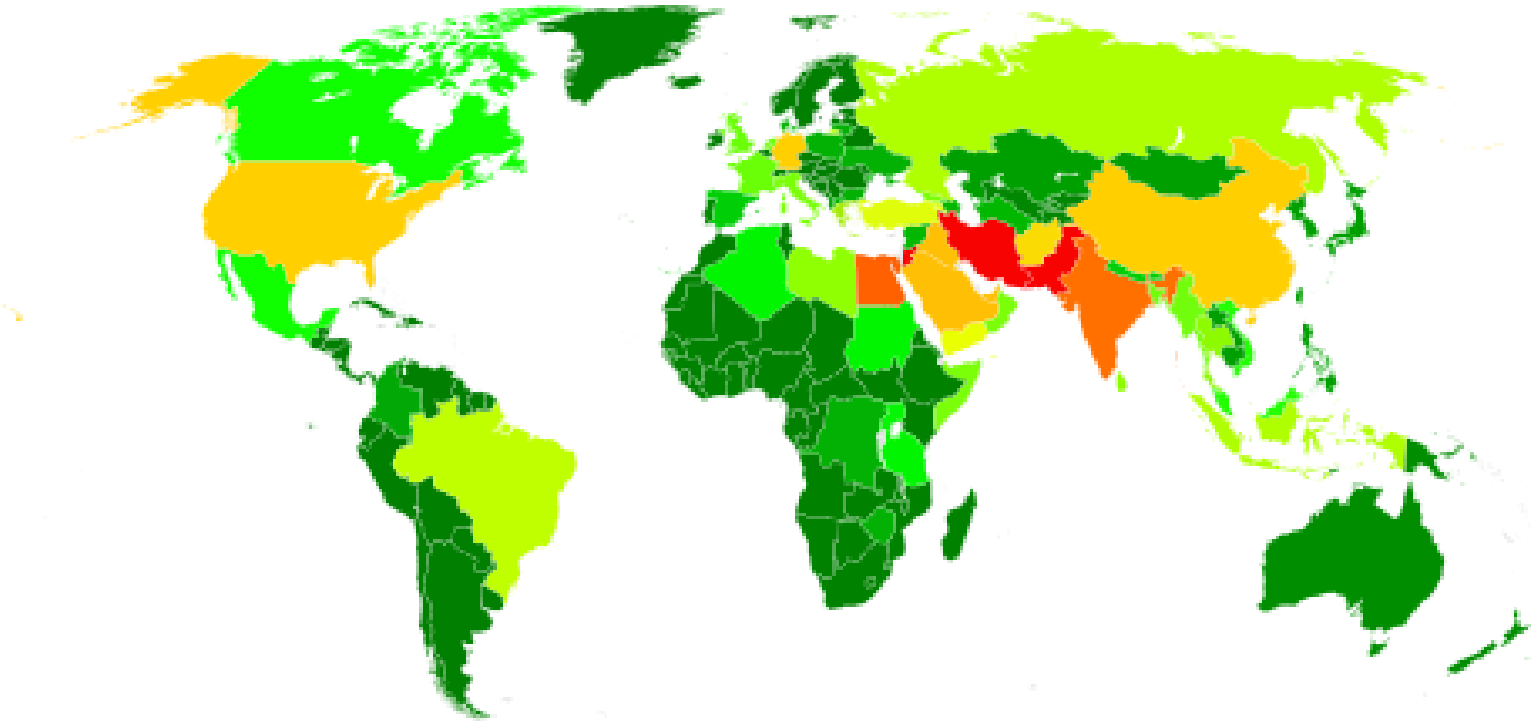
- Web sites often have Facebook content for comments, "like this page", etc.
- Facebook apps have access to a wide variety of data.
- Facebook itself has access to messages, likes, etc.

Phones

- "Family Tracker"
- The company can locate phones
- Apps may (depending on settings) access location services
- Facebook, again

The NSA

- 75% of Internet traffic
- Much other data



Hiding

- Random mac, unsecured network
  - LCSC
  - Your neighbor
  - WEP counts as unsecured...
  - Stealing cookies (Hotel Guest Network)
  - Hotspots: The modern pay phone.
  - Webcams at hotspots?
- Proxy servers
  - You do have to trust the proxy server

Machine Learning

- Useful when the solution to a problem is not known.
- Example: Brain Scan Classification
- Example: Identifying a particular person
- Or when the solution changes or is variable
- Example: Speech Recognition
- Example: Spam filters
- Recommendations: Many possible solutions
- Final Example: Stock Trading
- Another Scenario: Clustering

Problems from Large-Scale Data Collection

- Which advertisement for E-mail / Web Browsing / etc
- Product recommendations, Amazon and such
- Find threats in a large network
- More?

Terms

- Examples
- Labels
- Training Set
- Test Set
- Noise
- Overfitting

Many Problem Formulations

- Supervised Learning
- Clustering
- Link prediction
- Anomoly Detection

Bayesian Learning

$$P(C|X) = \frac{P(X|C)P(C)}{P(X)}$$

- Use Bayes Rule for classification
- Commonly used for spam filters, among many other things
- Bayes rule example: Medical Screening (99% accurate, $\frac{1}{1000}$ prevalence)
- X is the features of the example, which are not independent! Must be discrete.
- Calculating $P(X|C)$ is infeasible.
- Naive assumption: All features are independent! That is:

$$p(x_1, x_2, x_3, \ldots, x_D) = \prod_{i=1}^{D} P(x_i|C)$$

Parallel Programming

- Problem 1: Data may not fit in memory
- Problem 2: Data may not fit on hard drives either
- Problem 3: Computational time may be too long
- Embarassingly Parallel: Table example, GNA
- Waste vs. Overhead (Clusters)
- Special computers (Cray XMT)
- Cluster management demo

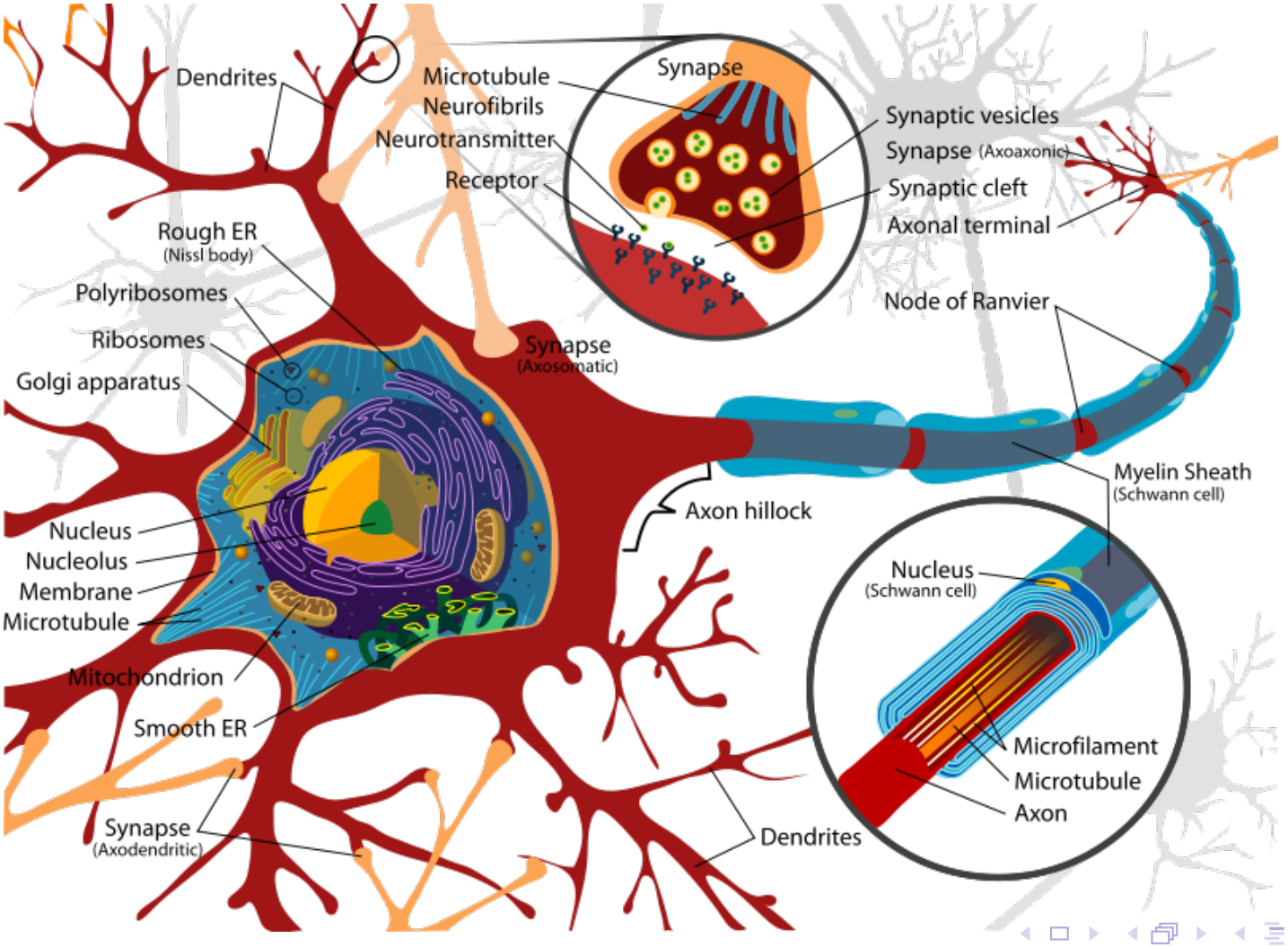
Decision Trees

- Graph of solution: Loans
- These are for supervised learning
- Entropy: measues impurity
- Specifically, Entropy $= -P_+ log_2 P_+ - P_i log_2 P_-$
- Information Gain: Loss of entropy
- Choose root based on *information gain*
- Overfitting

Weather Example

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Biological Inspiration



Real Neural Networks

- Biological inspiration in CS, horses, bees (China + tracking)
- Neurons connected with Axons
- 10^{11} in human
- Areas have distinct functions
- Example: See and swat fly
- Association Cortex
- Back to details: synaptic cleft (receptors, vesicles, transmitters)
- Re-uptake and breakdown
- Action Potential (ions)

Real Neural Netoworks Continued

- Major Neurotransmitters:
  - Amino acids: glutamate, aspartate, D-serine, aminobutyric acid (GABA), glycine
  - Monoamines and other biogenic amines: dopamine (DA), norepinephrine (noradrenaline; NE, NA), epinephrine (adrenaline), histamine, serotonin (SE, 5-HT)
  - Peptides: somatostatin, substance P, opioid peptides
  - Others: acetylcholine (ACh), adenosine, anandamide, nitric oxide, etc.
- Effect of neurotransmitters: Reward (dopamine), mood (serotonin)
- Neurotoxicity
- Tolerance (and homeopathic principle)
- Ions and mylination

Artificial Version

- Real brain is more complex (cats, singularity)
- Each perceptron takes inputs
- Weighted sum is taken over all inputs
- Back-propegation for training
- Fixed layers
- MultilayerPerceptron in Weka

Support Vector Machines

- Regular feature space may be inadequate
- Solution: Use high dimensional feature space!
- Kernel trick: Perform vector distance measure in lower dimensional space
- Kernel can vary (RBF is popular)
- Concept is simple, math is not