CS 68: Bioinformatics

Prof. Sara Mathieson
Spring 2018
Swarthmore College
Outline: April 25

• Lab 6 notes
• Finish Genome-Wide Association Studies (GWAS)
• Begin: machine learning for biology

Notes:
• Hand back project proposals today
• Office hours TODAY 1-3pm
• Midterm 2 in-lab on Thursday (make/bring cheat-sheet)
Lab 6 Notes

- $n =$ number of samples/sequences
- $m =$ number of sites
- Runtime of naïve algorithm: $O(nm^2)$
  - Need to consider all pairs of sites => $O(m^2)$
  - Containment/disjoint linear in $n$ by using a dictionary
- Runtime of Gusfield’s algorithm: $O(nm)$
  - Each step (radix sort, transform rows, build trie) considers each entry in the matrix $(n \times m)$
- Naïve is NOT exponentially faster than Gusfield! It is **quadratic** in $m$
- Recombination is the reason we don’t expect a perfect phylogeny when considering many sites for samples from the same species
Blue lines represent recombination rate.
697 independent SNPs significantly associated with height – Wood et al. 2014
Together explain about 15% of the phenotypic variance

ZBTB38: Zinc Finger And BTB Domain Containing 38

GDF5: Growth differentiation factor 5
32 independent SNPs explain 1.45% of the variance in BMI – Speliotes et al. 2010
Type 2 Diabetes GWAS

**TCF7L2**: Transcription factor 7-like 2  
Risk allele increases T2D risk ~40%

McCarthy et al 2013

63 independent loci explain 5.7% of the variance – Morris et al. 2012

Slide: modified from Iain Mathieson
Schizophrenia GWAS

Major Histocompatibility Complex - a region with many genes that produce cell surface proteins, important for immunity

Other associations:
- Glutaminergic neurons
- Calcium channels
- Synaptic plasticity
- Dopamine receptor DRD2

108 independent loci explain 3.4% of the variance – Ripke et al. 2014

Slide: modified from Iain Mathieson
The case of the missing heritability

When scientists opened up the human genome, they expected to find the genetic components of common traits and diseases. But they were nowhere to be seen. Brendan Maher shines a light on six places where the missing loot could be stashed away.
The bigger the sample size, the more variants you find.

Simons & Sella 2018

Slide: modified from Iain Mathieson
Missing Heritability?

“Missing heritability” is not really missing

Mostly just hidden in very small effects that GWAS are not big enough to detect

May be some hidden in epistatic effects or gene-environment interactions

Heritability estimates might be a bit too high

Slide: modified from Iain Mathieson
Almost all GWAS are carried out in European-Ancestry populations

**PERSISTENT BIAS**

Over the past seven years, the proportion of participants in genome-wide association studies (GWAS) that are of Asian ancestry has increased. Groups of other ancestries continue to be very poorly represented.

- **2009**
  - 373 studies
  - 1.7 million samples
  - 96% European ancestry
  - 4% Non-European ancestry

- **2016**
  - 2,511 studies
  - 35 million samples
  - 81% European ancestry
  - 19% Non-European ancestry

**BREAKDOWN**

Proportion of non-European ancestry samples

- Asian ancestry
- African ancestry
- Mixed ancestry
- Hispanic & Latin American ancestry
- Pacific Islander
- Arab & Middle Eastern
- Native Peoples

Terms for ethnicity are those used in the GWAS Catalog. Some have changed between 2009 and 2016 as sampling has increased. Samples of European origin have the most specific descriptions of population ancestry.

Slide: modified from Iain Mathieson

Popejoy & Fullerton 2016
European GWAS results do not translate to non-European ancestry populations

Ware et al 2018

Slide: modified from Iain Mathieson
How successful have GWAS been?

Twelve years.

Thousands of studies

Tens of thousands of researchers

Tens of millions of patient-participants

Billions (?) of dollars
How successful have GWAS been?

GWAS

- Extremely successful!

Find associations with traits and diseases

- Not very successful at all

Understand function of those associations i.e. “find genes”

- Hasn’t really happened

Develop drugs

- Hasn’t happened at all

Profit

New Opportunities

- Find connections between traits
- Predict genetic risk
- Understand complex trait evolution

Slide: modified from Iain Mathieson
Summary

Genome-wide association studies:

Map common/low frequency variants associated with traits/disease

The bigger the sample size (more people) the smaller the effects you can detect

Do not tell us anything about function

Need to be extremely careful about population structure and multiple testing
Machine Learning in Biology
What is machine learning?

A child can see one giraffe and then be able to identify giraffes in many different contexts.

Images: Wikipedia, San Diego Zoo, National Geographic, CNN.com
Can we train a computer to do the same thing?
Can we train a computer to do the same thing?

How can we *distinguish* between similar objects?
What is machine learning?

• One flavor of machine learning is *classification*

• Goal: separate examples into (many) different *classes*
Why do we care?

• Email filtering (spam vs. not-spam)

From: cheapsales@buystufffromme.com
To: ang@cs.stanford.edu
Subject: Buy now!
Deal of the week! Buy now!
Rolex watches - $100
Medicine (any kind) - $50
Also low cost M0rgages available.

From: Alfred Ng
To: ang@cs.stanford.edu
Subject: Christmas dates?
Hey Andrew,
Was talking to Mom about plans
for Xmas. When do you get off
work. Meet Dec 22?
Alf

• Handwriting recognition (digits in a check)
Why do we care?

Self-driving cars are in our present and future

AlphaGo: plays humans never thought of

Images: Scientific American
Why do we care?

• Tumor detection (benign vs. malignant)
ML and “Big Data”

• As datasets become larger and more complex, humans can no longer make sense of them without machines

• Machine learning is in all of our lives and understanding it will be increasingly valuable
Machine learning terminology

• *Training*: usually involves the program processing many *examples* (from different classes) where we know the “answer” or *label*, and learning how to separate them

• *Testing*: program classifies new examples
Machine learning terminology

• *Supervised learning*: a human (usually) has hand-labeled the training examples, so it’s easier for the computer to learn differences

• *Unsupervised learning*: data is unlabeled (no class information)
Machine Learning Methods
Regression

**Training data:** vectors $\mathbf{x}$ (independent variable) and $\mathbf{y}$ (dependent variable)
Regression

**Training data:** vectors $x$ (independent variable) and $y$ (dependent variable)

**Testing goal:** given a new $x$ value, can we predict $y$?
Logistic regression for classification
Logistic regression for classification

![Graph showing the probability of passing an exam versus hours of studying.](Images: Wikipedia)
Support Vector Machines (SVM)

Idea: for 2 (or more classes), try to create the “best” boundary between them

New examples can be classified based on which side of the hyperplane they fall on
Clustering (unsupervised learning)

Choose two random data points to be the first means
Clustering (unsupervised learning)

Color each point based on which mean is closest, then find means of resulting clusters.

Images: Polymatheia
Clustering (unsupervised learning)

Repeat the process until the means are not changing
Clustering (unsupervised learning)

Repeat the process until the means are not changing

Images: Polymathelia
Clustering (unsupervised learning)

Repeat the process until the means are not changing
Clustering (unsupervised learning)

Repeat the process until the means are not changing
HMMs form a class of machine learning methods too

• Can be *supervised* (i.e. we know the hidden state sequence for some examples, use that to infer transition/emission probabilities)

• Then estimate hidden state sequence for new unlabeled data

• Can be *unsupervised* (i.e. we don’t know the hidden state sequence and we want to learn/predict this latent variable)
Recent trends in ML

• Inspired by how neurons are connected in our brains, “deep learning” has recently become successful in many fields.
Deep learning for images

\[ \begin{align*}
X_1 & \quad \rightarrow \quad Y_1 \text{ (glasses)} \\
X_2 & \quad \rightarrow \quad Y_2 \text{ (smiling)} \\
X_3 & \quad \rightarrow \quad Y_3 \text{ (eye size)} \\
X_4 & \quad \rightarrow \quad \text{ } \\
X_5 & \quad \rightarrow \quad \text{ } \\
X_6 & \quad \rightarrow \quad \text{ } \\
\text{input data} & \quad \rightarrow \quad ? \\
\end{align*} \]
Classical neural network

Deep Learning
Number of articles about deep learning over time

2006: Hinton and Salakhutdinov make a break-through in initializing deep learning networks
Break-through: unsupervised learning, autoencoder
1. Project data into a lower dimension:

\[ h_j = \sigma(W_{j}^{(1)} \cdot x) \]

\[ \sigma(z) = \frac{1}{1 + e^{-z}} \]
1. Project data into a lower dimension:

\[ h_j = \sigma (W_j^{(1)} \cdot x) \]

\[ \sigma(z) = \frac{1}{1 + e^{-z}} \]

2. From reduced features, reconstruct:

\[ x_i^* = \sigma (W_i^{(2)} \cdot h) \]

---

Sara Mathieson
Break-through: unsupervised learning, autoencoder

1. Project data into a lower dimension:

\[ h_j = \sigma(W_j^{(1)} \cdot x) \]

\[ \sigma(z) = \frac{1}{1 + e^{-z}} \]

2. From reduced features, reconstruct:

\[ x_i^* = \sigma(W_i^{(2)} \cdot h) \]

3. Minimize objective function:

\[ J_x(W) = \frac{1}{2} \| x - x^* \|^2 \]
PCA vs. Autoencoder

Original image
PCA vs. Autoencoder

Original image

PCA reconstruction
PCA vs. Autoencoder

Original image  |  PCA reconstruction  |  Autoencoder reconstruction

- Original image:
  - Top row: 3.00, 8.01, 13.87.
  - Middle row: 500, 2000, 1000.
  - Bottom row: 2, 10, 4.

- PCA reconstruction:
  - Top row: 3.00, 8.01, 13.87.
  - Middle row: 500, 2000, 1000.
  - Bottom row: 2, 10, 4.

- Autoencoder reconstruction:
  - Top row: 3.00, 8.01, 13.87.
  - Middle row: 500, 2000, 1000.
  - Bottom row: 2, 10, 4.
Transform the input data
Feature learning for hidden layer 1
Feature learning for hidden layer 1

input data → hidden layer 1 → reconstructed input

- $X_1^*$
- $X_2^*$
- $X_3^*$
- $X_4^*$
- $X_5^*$
- $X_6^*$

Deep Learning
Low-level features become the new data

hidden layer 1
Feature learning for hidden layer 2

Diagram showing the flow of information from hidden layer 1 to hidden layer 2 and then to the reconstructed input.
Feature learning for hidden layer 2
High-level features become the new data
Last layer: supervised learning
Last layer: supervised learning
"Fine-tune" the entire deep network

Deep Learning
Application of deep learning to population genetics
Motivation: demographic history of *Drosophila*

![Diagram showing demographic history](image)

- $N_3 = 5,224,100$
- $N_2 = 620$
- $N_1 = 4,975,360$

Demographic Inference Reveals African and European Admixture in the North American *Drosophila melanogaster* Population

Pablo Duchen, Daniel Živković, Stephan Hutter, Wolfgang Stephan and Stefan Laurent

*GENETICS* January 1, 2013 Vol. 193 No. 1 291-301; [https://doi.org/10.1534/genetics.112.145912](https://doi.org/10.1534/genetics.112.145912)
Joint analysis of demography and selection in population genetics: where do we stand and where could we go?

The complete genome sequence of a Neanderthal from the Altai Mountains, *Nature* (2013)
Selective sweeps can cause a loss of diversity
Training data: simulated datasets

400,000 datasets:
- 2,500 bottlenecks
- 160 regions/genome
Training data: simulated datasets

400,000 datasets:
- 2,500 bottlenecks
- 160 regions/genome

1. baseline effective population size: $N_e = 100,000$
2. $n = 100$ individuals
3. $L = 100,000$ bases per region
4. 75% of data for training and 25% for testing
Selection: four different classes

Population Sizes 1

Population Sizes 2

Population Sizes 3

- shaded green circle: de novo mutation (hard sweep)
- red triangle: balancing selection
- yellow circle: standing variation (soft sweep)

⇒ 4 selection classes
Compute statistics around selected site

region 1: within 10kb
region 2: 10-30kb
region 3: 30-50kb
Summary statistics as features

- Number of segregating sites: 3 stats
- Tajima’s $D$: 3 stats
- Folded site frequency spectrum (SFS): 150 stats
- Length distribution between segregating sites: 48 stats
- Identity-by-state (IBS) tract length distribution: 90 stats
- Linkage disequilibrium (LD) distributions: 48 stats
- Haplotype frequency statistics: 3 stats

= 345 features total
A deep learning method for population genetics