CS 260: Foundations of Data Science

Prof. Thao Nguyen

Fall 2024

HAVERFORD COLLEGE

Materials by Sara Mathieson

Outline for today

- Midterm 2 Review
  - PCA
  - Naïve Bayes
  - Logistic regression and cross entropy

Outline for today

- Midterm 2 Review
  - PCA
  - Naïve Bayes
  - Logistic regression and cross entropy

From the study guide

6. Data Visualization

- Best ways of visualizing discrete vs. continuous data
- How to choose colors; idea of sequential, diverging, or qualitative color schemes
- How to make color schemes color-blind and black/white printing friendly
- Idea of principal component analysis (PCA) as a way to accomplish dimensionality reduction
- Using dimensionality reduction to visualize high-dimensional data
- Details of the PCA algorithm (except computing eigenvalues and eigenvectors)
- Runtime of PCA
- Genealogical interpretation of PCA plots for genetic data

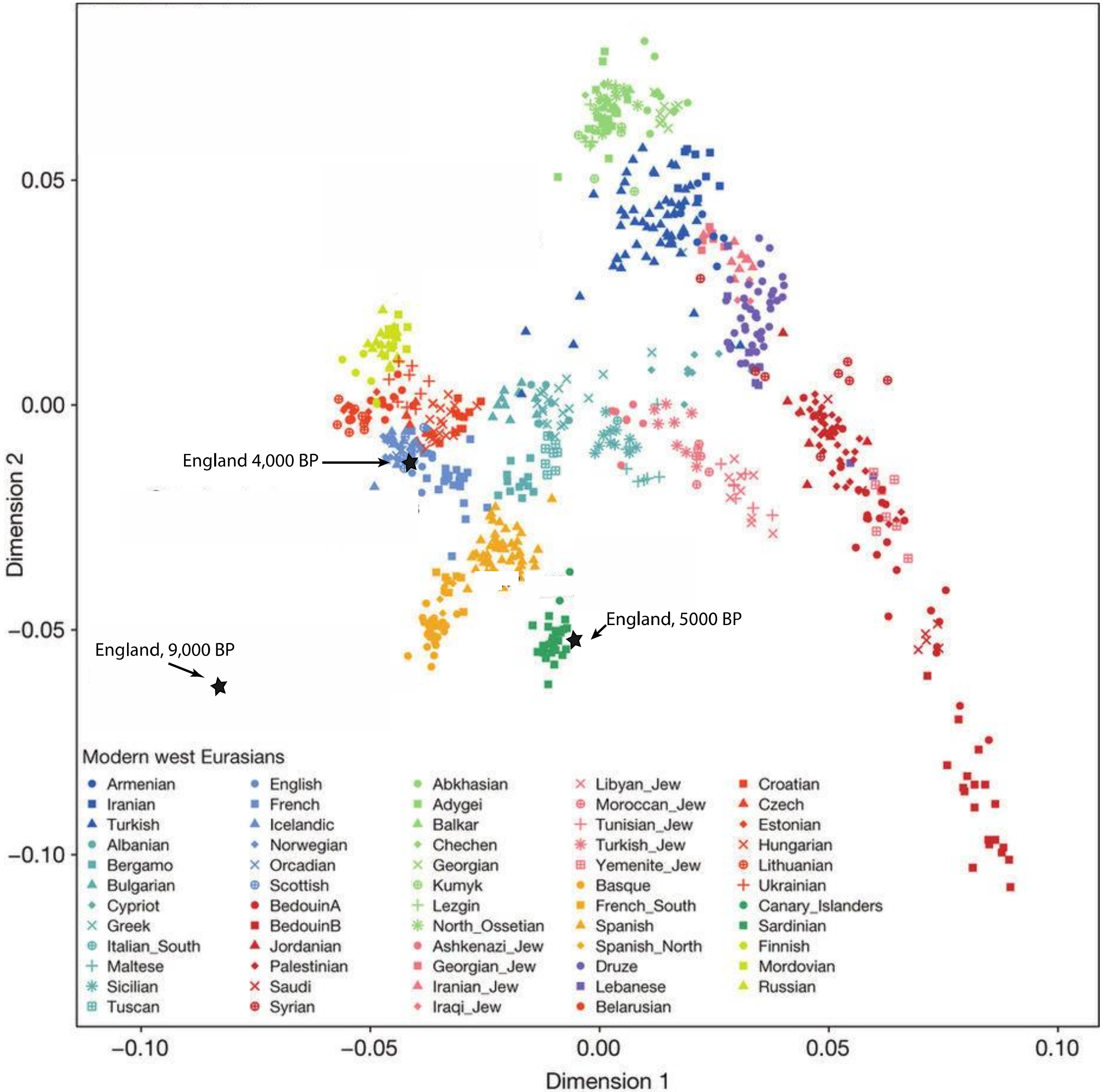
Principal Component Analysis (PCA)

- Transforms *p*-dimensional data so that the new first dimension explains as much of the variation as possible, the new second explains as much of the remaining variation as possible, and so on

- PCA is a linear transformation

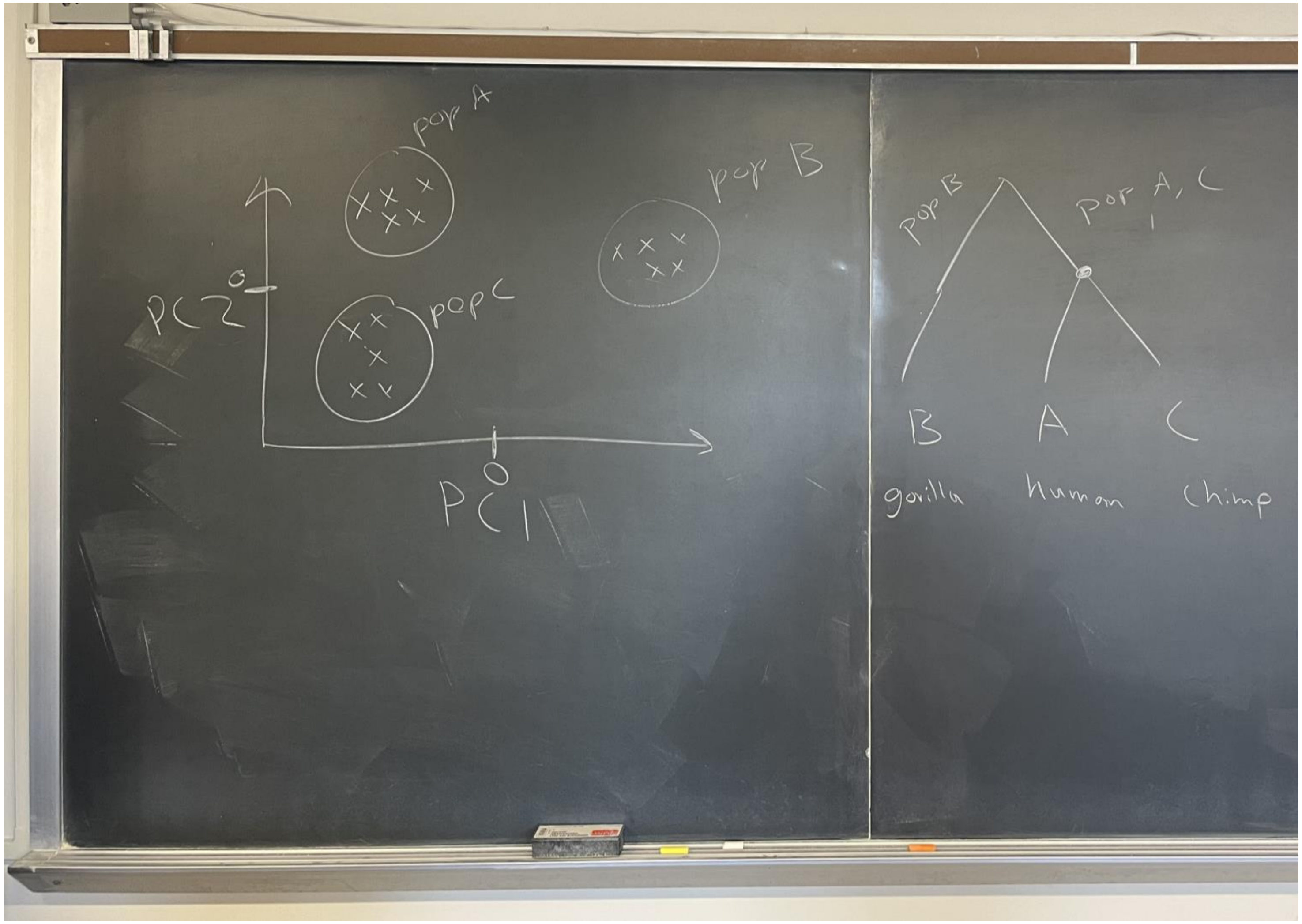
- Typically, we look at the first few dimensions of the transformed data as a means of dimensionality reduction and visualization

- PCA is often used for:
  - Data visualization
  - Infer qualitative relationships between groups



England 4,000 BP
England, 5000 BP
England, 9,000 BP
0.05
0.00
−0.05
−0.10
Dimension 2
−0.10
−0.05
0.00
0.05
0.10
Dimension 1
Modern west Eurasians
Armenian
Iranian
Turkish
Albanian
Bergamo
Bulgarian
Cypriot
Greek
Italian_South
Maltese
Sicilian
Tuscan
English
French
Icelandic
Norwegian
Orcadian
Scottish
BedouinA
BedouinB
Jordanian
Palestinian
Saudi
Syrian
Abkhasian
Adygei
Balkar
Chechen
Georgian
Kumyk
Lezgin
North_Ossetian
Ashkenazi_Jew
Georgian_Jew
Iranian_Jew
Iraqi_Jew
Libyan_Jew
Moroccan_Jew
Tunisian_Jew
Turkish_Jew
Yemenite_Jew
Basque
French_South
Spanish
Spanish_North
Druze
Lebanese
Belarusian
Croatian
Czech
Estonian
Hungarian
Lithuanian
Ukrainian
Canary_Islanders
Sardinian
Finnish
Mordovian
Russian

PCA "classic" genetics example



Outline for today

- Midterm 2 Review
  - PCA
  - Naïve Bayes
  - Logistic regression and cross entropy

From the study guide

2. Naive Bayes

- Bayes rule in data science: identify and explain the evidence, prior, posterior, likelihood.
- Derivation of the Naive Bayes model for $p(y = k|\vec{x})$ (via the Naive Bayes assumption).
- How do we estimate the probabilities of a Naive Bayes model?
- Laplace counts (motivation, application details)
- How can we predict the label of a new example after fitting a Naive Bayes model?
- What types of features/label do we currently require for Naive Bayes?
- How Naive Bayes can be implemented using dictionaries in Python

Bayes' Theorem

- P(A,B) = P(A|B)P(B)
- P(A,B) = P(B|A)P(A)

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Independence

- Independence: $P(A,B) = P(A)P(B)$
- Conditional independence: $P(A|B,C) = P(A|C)$

not always true!

Naïve Bayes
assumption

Naïve Bayes Model

$$p(y = k|\boldsymbol{x}) \propto p(y = k) \prod_{j=1}^{p} p(x_j|y = k).$$

Naïve Bayes Prediction

$$\hat{y} = \underset{k \in \{1,2,\cdots,K\}}{\arg\max} \; p(y = k) \prod_{j=1}^{p} p(x_j|y = k).$$

Estimating prior: p(y=k)

$$\theta_k = \frac{N_k + 1}{n + K}$$

Estimating likelihood: p(x$_j$=v | y=k)

$$\theta_{k,j,v} = \frac{N_{k,j,v} + 1}{N_k + |f_j|}$$

Outline for today

- Midterm 2 Review
  - PCA
  - Naïve Bayes
  - Logistic regression and cross entropy

From the study guide

5. Logistic Regression

- Motivation for logistic regression; our model is a logistic function that takes in $\vec{w} \cdot \vec{x}$
- Logistic regression creates a *linear* decision boundary (visualize for $p = 1$).
- In logistic regression our cost is the negative log likelihood (don't need to derive)
- Intuition/visualization of the cost function (and relationship to cross entropy)
- Idea of SGD for logistic regression, relationship to linear regression

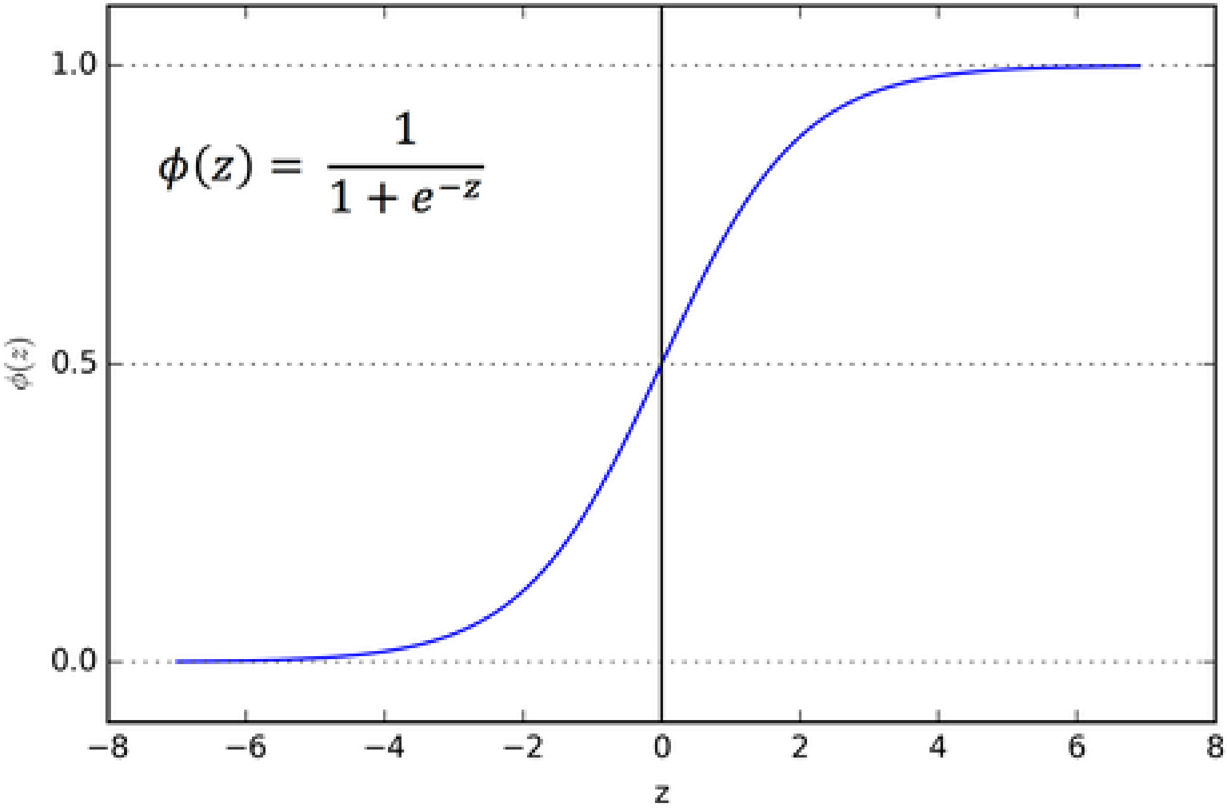
3 important pieces to SGD

- Hypothesis function (prediction)

$$\boxed{h_{\boldsymbol{w}}(\boldsymbol{x}) = p(y = 1|\boldsymbol{x}) = \frac{1}{1 + e^{-\boldsymbol{w} \cdot \boldsymbol{x}}}}$$

Logistic (sigmoid) function

Transforms a continuous real number into a range of (0, 1)



Logistic Regression

- Binary classification $y \in \{0,1\}$
- Model will be

$$h_{\vec{w}}(\vec{x}) = p(y = 1|\vec{x}) = \frac{1}{1 + e^{-\vec{w}\cdot\vec{x}}}$$

- Classification (already have $\vec{w}$)

if $\vec{w} \cdot \vec{x} \geq 0 \Rightarrow \hat{y} = 1$

$\vec{w} \cdot \vec{x} < 0 \Rightarrow \hat{y} = 0$

3 important pieces to SGD

- Hypothesis function (prediction)

$$h_{\boldsymbol{w}}(\boldsymbol{x}) = p(y = 1|\boldsymbol{x}) = \frac{1}{1 + e^{-\boldsymbol{w}\cdot\boldsymbol{x}}}$$

- Cost function (want to minimize)

$$J(\boldsymbol{w}) = -\sum_{i=1}^{n} y_i \log h_{\boldsymbol{w}}(\boldsymbol{x_i}) + (1 - y_i) \log(1 - h_{\boldsymbol{w}}(\boldsymbol{x_i}))$$

3 important pieces to SGD

- Hypothesis function (prediction)

$$\boxed{h_{\boldsymbol{w}}(\boldsymbol{x}) = p(y = 1|\boldsymbol{x}) = \frac{1}{1 + e^{-\boldsymbol{w}\cdot\boldsymbol{x}}}}$$

- Cost function (want to minimize)

$$\boxed{J(\boldsymbol{w}) = -\sum_{i=1}^{n} y_i \log h_{\boldsymbol{w}}(\boldsymbol{x_i}) + (1 - y_i)\log(1 - h_{\boldsymbol{w}}(\boldsymbol{x_i}))}$$

- Gradient of cost wrt single data point x_i

$$\boxed{\nabla J_{\boldsymbol{x}_i}(\boldsymbol{w}) = (h_{\boldsymbol{w}}(\boldsymbol{x_i}) - y_i)\boldsymbol{x_i}}$$

Stochastic Gradient Descent for Logistic Regression (binary classification)

set $\vec{w} = \vec{0}$

while cost $\mathrm{J}(\vec{w})$ is still changing:

  shuffle data points

  for i = 1,…,n:

    $$\vec{w} \leftarrow \vec{w} - \alpha \underbrace{\nabla J_{\vec{x_i}}(\vec{w})}_{\text{derivative of } \mathrm{J}(\vec{w}) \text{ wrt } x_i}$$

  store $\mathrm{J}(\vec{w})$

For each method/approach, is X continuous or discrete? What about y?

- Linear regression
- Polynomial regression
- Decision trees/stumps
- ROC curve as an evaluation metric
- Naïve bayes
- Logistic regression
- Entropy and information gain
- PCA

Think about offline!