

CS 260 – Foundations of Data Science

Lecture 21– Midterm 2 Review

11/19/2026

Adam Poliak

BRYN MAWR COLLEGE

Outline for today

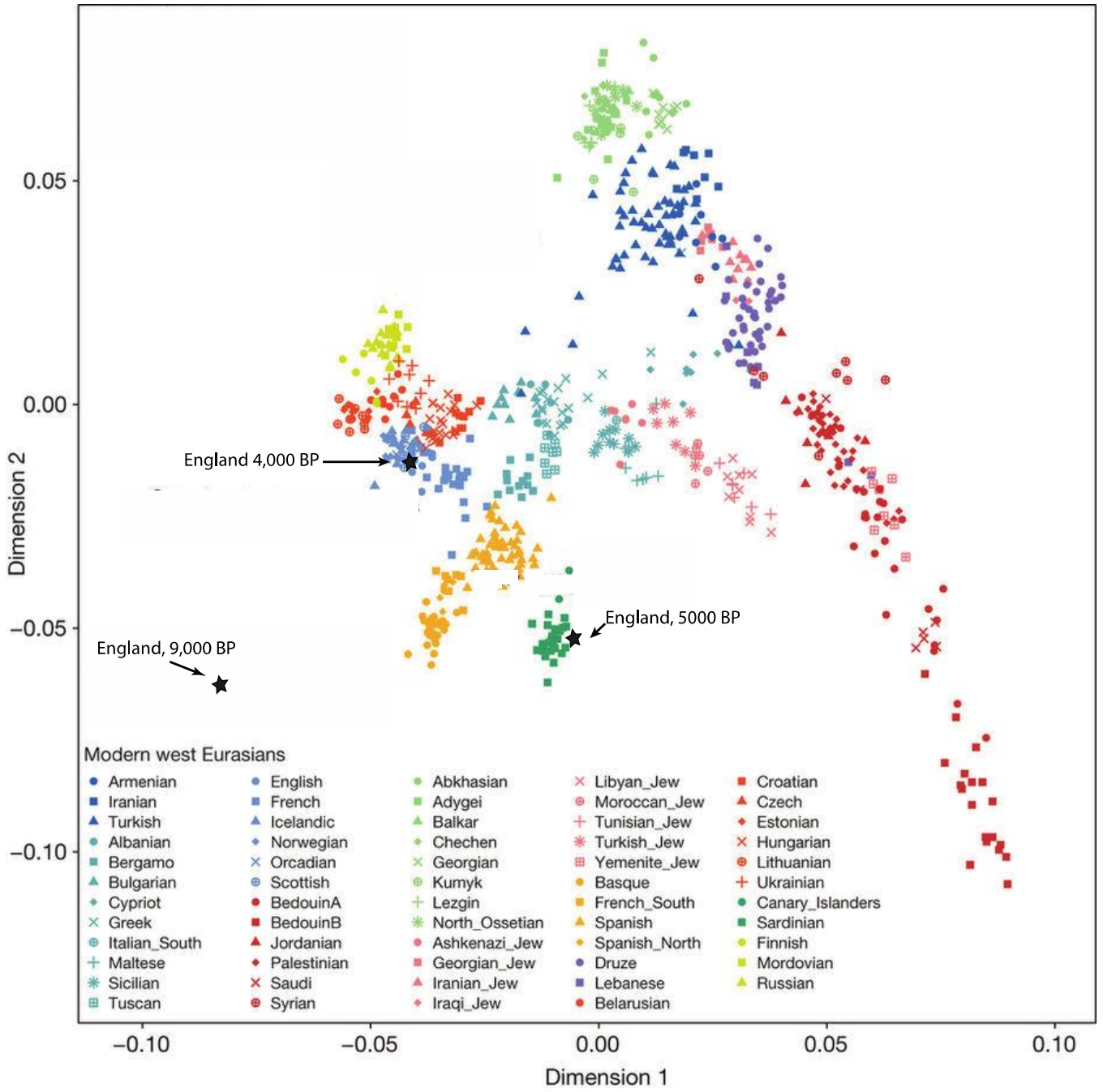
- Midterm 2 Review
  - PCA
  - Naïve Bayes
  - Logistic regression and cross entropy

Outline for today

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

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



BRYN MAW COLLEGE

Outline for today

- Midterm 2 Review
  - PCA
  - **Naïve Bayes**
  - Algorithmic Bias and Disparate Impact
  - Logistic regression and cross entropy

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: $P(A,B) = P(A)P(B)$

not always true!

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

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: $p(x_j=v \mid y=k)$

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

Outline for today

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

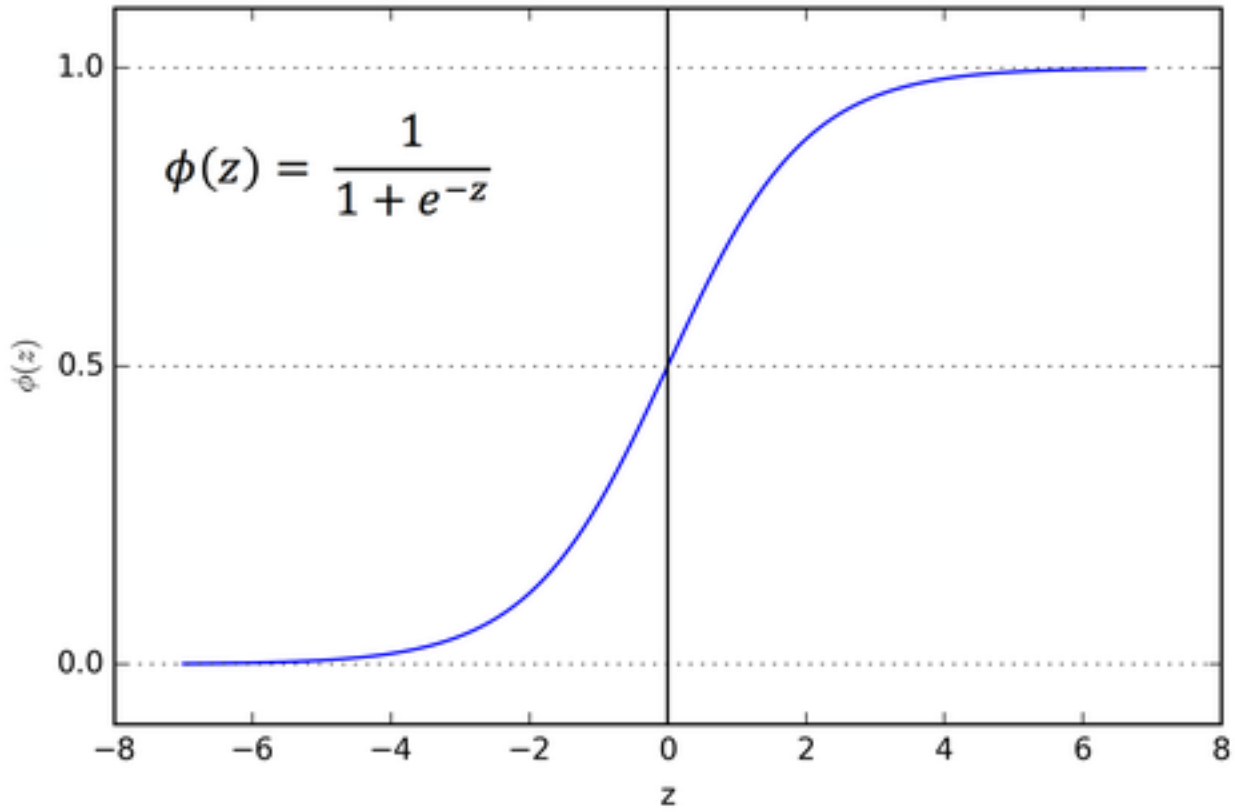
3 important pieces to SGD

- Hypothesis function (prediction)

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

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)

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

- Cost function (want to minimize)

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

- Gradient of cost wrt single data point x_i

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

Stochastic Gradient Descent for Logistic Regression (binary classification)

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

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

shuffle data points

for i = 1,...,n:

$$\vec{w} \leftarrow \vec{w} - \alpha \nabla J_{\vec{x_i}}(\vec{w})$$

store $J(\vec{w})$

derivative of $J(\vec{w})$ wrt x_i

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!