CS 260: Foundations of Data Science

Prof. Thao Nguyen

Fall 2025

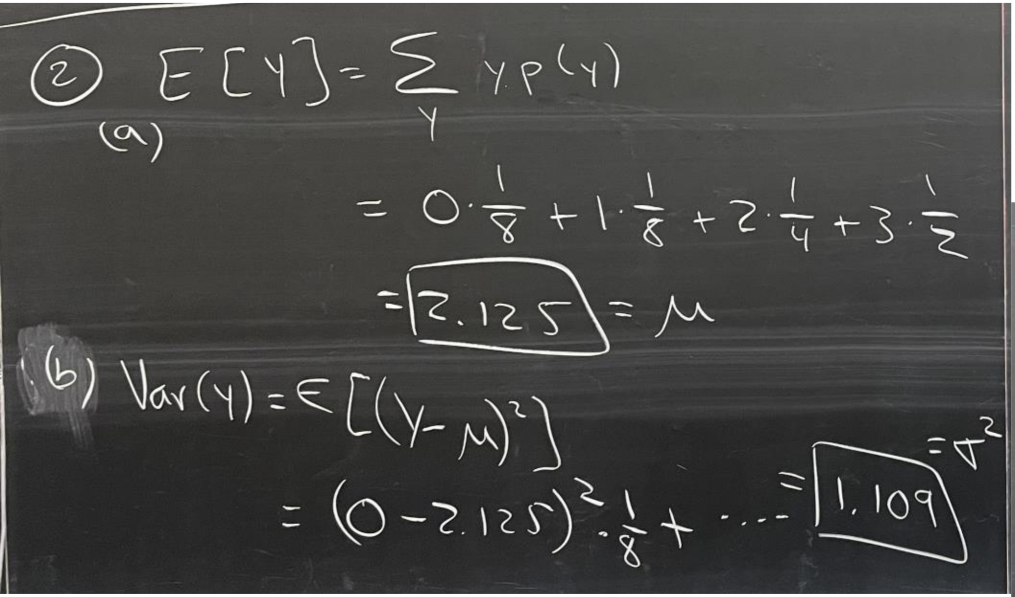
HAVERFORD
COLLEGE

Materials by Sara Mathieson

Outline for today

- Practice Midterm 2

Central Limit Theorem



② $E[Y] = \sum_{y} y\,p(y)$

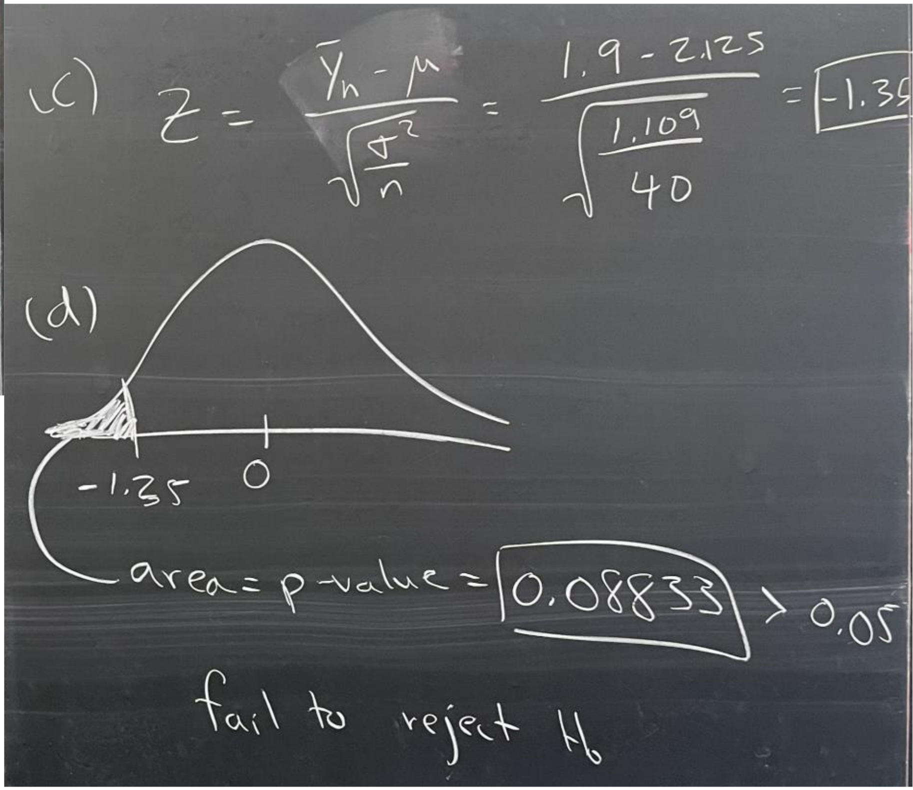
(a)

$$= 0\cdot\frac{1}{8} + 1\cdot\frac{1}{8} + 2\cdot\frac{1}{4} + 3\cdot\frac{1}{2}$$

$$= \boxed{2.125} = \mu$$

(b) $Var(y) = E[(y-\mu)^2]$

$$= (0-2.125)^2\cdot\frac{1}{8} + \dots = \boxed{1.109} = \sigma^2$$



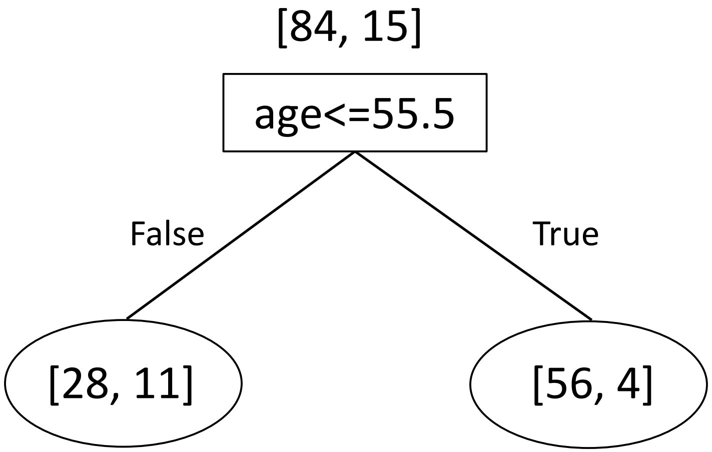
(c) $Z = \dfrac{\bar{Y}_n - \mu}{\sqrt{\dfrac{\sigma^2}{n}}} = \dfrac{1.9 - 2.125}{\sqrt{\dfrac{1.109}{40}}} = \boxed{-1.35}$

(d)

area = p-value = $\boxed{0.08833}$ > 0.05

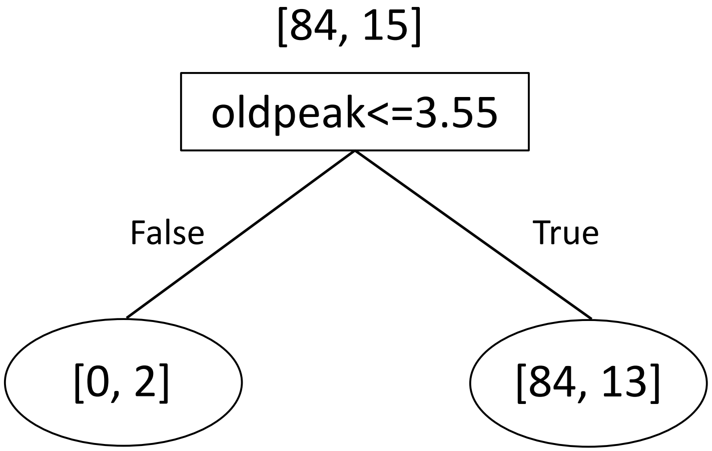
fail to reject H_0

Classification error



Classification error:

$$\frac{11+4}{99}=\frac{15}{99}$$



Classification error: $13/99$

Entropy

- $H(y) = -\sum_{c \in vals(y)} p(y = c) \log_2\big(p(y = c)\big)$

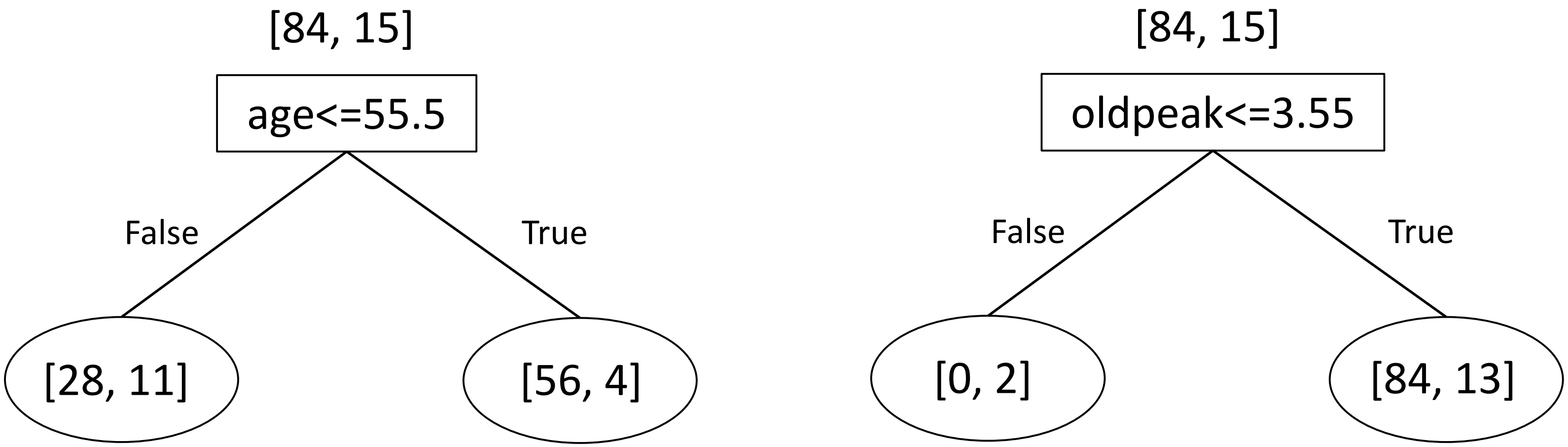
  $= -\left(\frac{84}{99}\log_2\frac{84}{99} + \frac{15}{99}\log_2\frac{15}{99}\right) = 0.61$

- $H(y|\text{oldpeak}) = \frac{2}{99} H(y|\text{oldpeak} = F) +$

  $\frac{97}{99} H(y|\text{oldpeak} = T)$

- $H(y|\text{oldpeak} = T) = -\left(\frac{84}{97}\log_2\frac{84}{97} + \frac{13}{97}\log_2\frac{13}{97}\right)$

Entropy



[84, 15]

age<=55.5

False

True

[28, 11]

[56, 4]

[84, 15]

oldpeak<=3.55

False

True

[0, 2]

[84, 13]

H(Y) = 0.6136190195993708

H(Y|age<=55.5) = 0.5522480910534322

H(Y|oldpeak<=3.55) = 0.5568804630596093

=> Age feature produces more information gain!

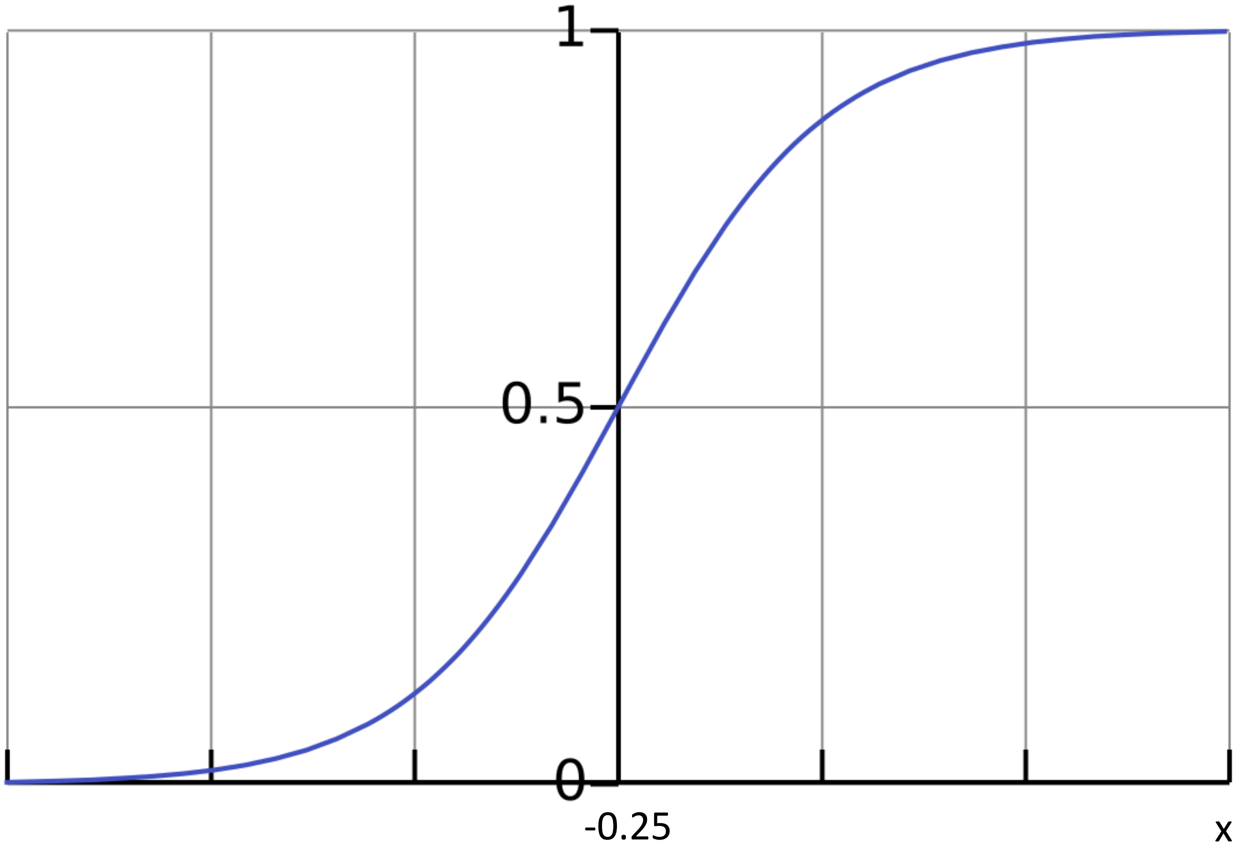
Logistic regression

Say I train a binary logistic regression model (i.e. outcomes $\in \{0, 1\}$) and end up with $\hat{w} = [\hat{w}_0, \hat{w}_1]^T = [1, 4]^T$. What is the decision boundary? Sketch a graph of this logistic model and label the decision boundary. How would you classify a new point $x_{\text{test}} = -0.3$? $< -0.25 \Rightarrow$ predict 0

$$w_0 + w_1 x \geq 0 \Rightarrow \hat{y} = 1$$

$$1 + 4x \geq 0$$

$$x \geq -1/4$$



SGD

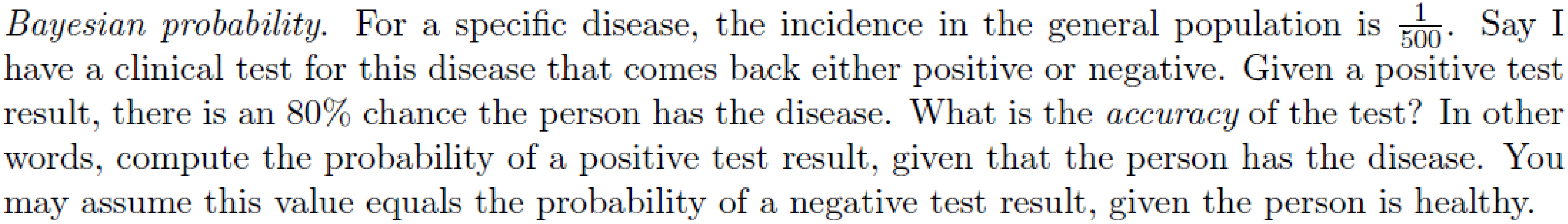
We are performing SGD to train a logistic regression model. We start with $\vec{w} = [w_0, w_1]^T = [0, 0]^T$ and $\alpha = 0.01$. What are the new weights after analyzing data point $(x, y) = (-3, 1)$?

$$\vec{w} \leftarrow \vec{w} - \alpha\left(h_{\vec{w}}(\vec{x_i}) - y_i\right)\vec{x_i}$$

$$h_{\vec{w}}(\vec{x_i}) = \frac{1}{1 + e^{-\vec{w}\cdot\vec{x_i}}} = \frac{1}{1 + e^{0}} = 0.5$$

$$\begin{bmatrix} w_0 \\ w_1 \end{bmatrix} \leftarrow \begin{bmatrix} 0 \\ 0 \end{bmatrix} - 0.01(0.5 - 1)\begin{bmatrix} 1 \\ -3 \end{bmatrix} = \begin{bmatrix} 0.005 \\ -0.015 \end{bmatrix}$$

Bayesian probability



Bayesian probability. For a specific disease, the incidence in the general population is $\frac{1}{500}$. Say I have a clinical test for this disease that comes back either positive or negative. Given a positive test result, there is an 80% chance the person has the disease. What is the *accuracy* of the test? In other words, compute the probability of a positive test result, given that the person has the disease. You may assume this value equals the probability of a negative test result, given the person is healthy.

D=has disease H=healthy

$$x = p(pos|D) = \frac{p(D|pos)p(pos)}{p(D)} = \frac{0.8 * p(pos)}{1/500}$$

$$p(pos) = p(pos, D) + p(pos, H) = p(pos|D)p(D) + p(pos|H)p(H)$$

$$p(pos|H) = 1 - p(neg|H) = 1 - p(pos|D)$$

$$x = \frac{0.8 * p(pos)}{1/500} = \frac{0.8 * \left(\frac{x}{500} + \frac{499}{500}(1-x)\right)}{1/500} \approx 0.9995$$

Naïve Bayes

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

- $\theta_0 = \frac{N_0+1}{n+K} = \frac{4}{7}$; $\theta_1 = \frac{3}{7}$

Naïve Bayes

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

- $\theta_0 = \frac{N_0+1}{n+K} = \frac{4}{7};\ \theta_1 = \frac{3}{7}$ $\quad \theta_{k,j,v} = \frac{N_{k,j,v} + 1}{N_k + |f_j|}$
- $\vec{x} = [1, D]$
  - $p(\vec{x}|y = 0) = \theta_{0,1,1}\theta_{0,2,D}$

Naïve Bayes

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

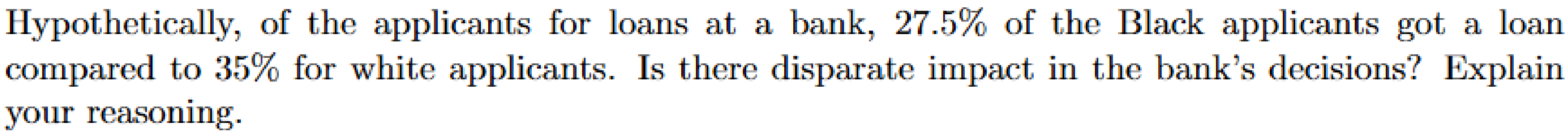
- $\theta_0 = \frac{N_0+1}{n+K} = \frac{4}{7};\ \theta_1 = \frac{3}{7}$ $\quad \theta_{k,j,v} = \frac{N_{k,j,v}+1}{N_k+|f_j|}$
- $\vec{x} = [1, D]$
  - ○ $p(\vec{x}|y=0) = \theta_{0,1,1}\theta_{0,2,D} = \frac{2}{6} * \frac{1}{8} = \frac{1}{24}$
  - ○ $p(\vec{x}|y=1) = \frac{2}{5} * \frac{1}{7} = \frac{2}{35}$

Naïve Bayes

- $p(y = 0|\vec{x}) \propto \frac{4}{7} * \frac{1}{24} \approx 0.0238$
- $p(y = 1|\vec{x}) \propto \frac{3}{7} * \frac{2}{35} \approx 0.0245$

$\Rightarrow$ predict $y = 1$

Disparate impact

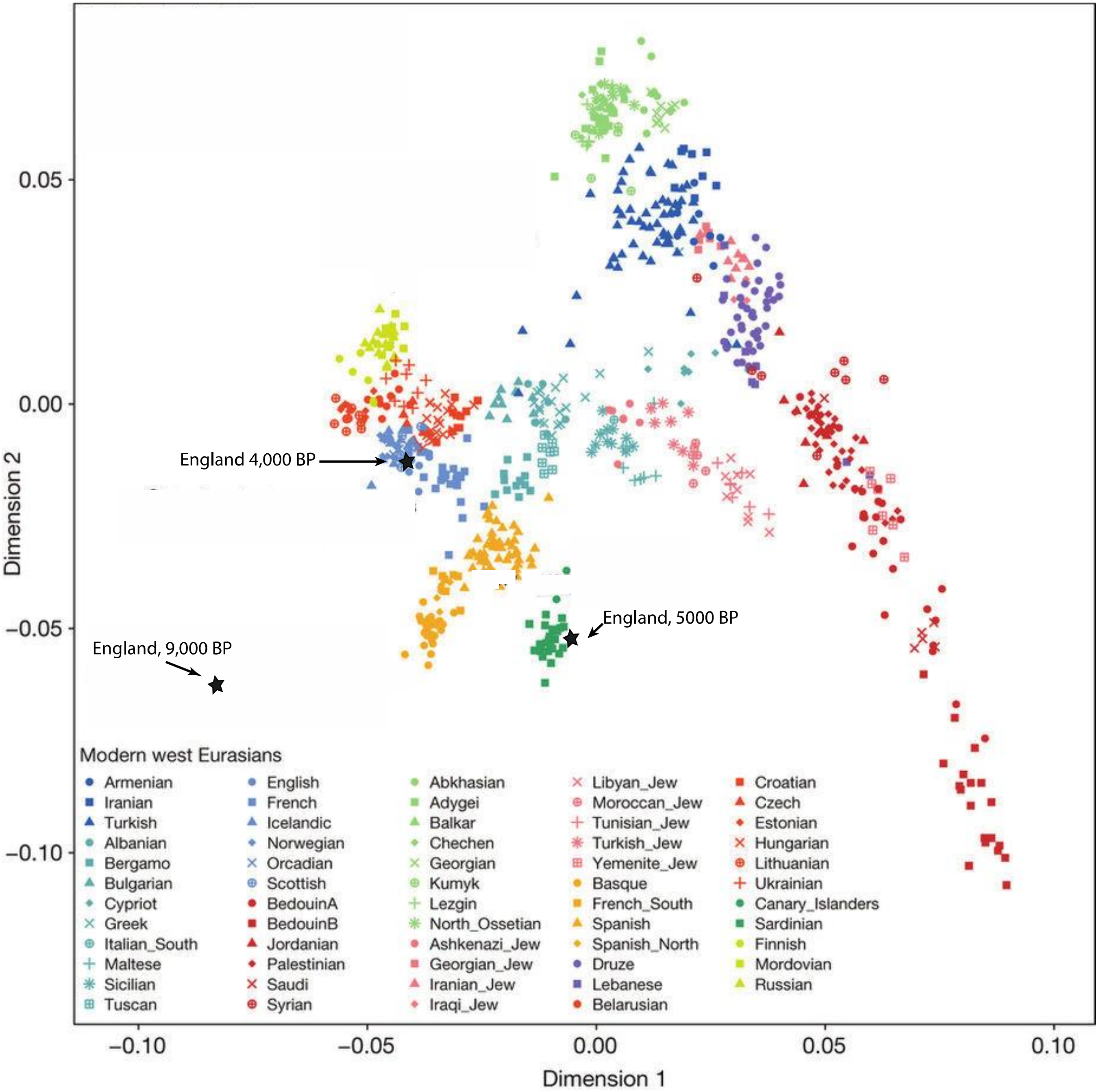


Hypothetically, of the applicants for loans at a bank, 27.5% of the Black applicants got a loan compared to 35% for white applicants. Is there disparate impact in the bank's decisions? Explain your reasoning.

If $P(C = 1|X = 0) < 0.8 * P(C = 1|X = 1)$

$\Rightarrow$ disparate impact

PCA



England 4,000 BP

England, 5000 BP

England, 9,000 BP