CS 260: Foundations of Data Science

Prof. Thao Nguyen

Fall 2024

Materials by Sara Mathieson

Admin

- **Lab 7** grades & feedback posted on Moodle
- **Lab 8** due tonight
- **Optional lab** tomorrow (Tuesday)
- **Midterm 2 handed out today in class**
  - Do not open the exam until you're ready to start
  - Time limit: **3 hours**
  - Resources: hand-written study sheet, calculator
  - Due Monday (11/25) at the beginning of class

Outline for today

- Practice Midterm 2

Handout 21

1. Bootstrap

Unordered

- $n = 2 \Rightarrow \{n_1, n_1\},\ \{n_1, n_2\},\ \{n_2, n_2\}$ 3 sets
- $n = 3$
  - $\{n_1, n_1, n_1\} \Rightarrow$ 3 sets
  - $\{n_1, n_2, n_3\} \Rightarrow$ 1 set
  - $\{n_1, n_1, n_2\} \Rightarrow$ 6 sets

  } 10 sets

Ordered: n^n sets

(2) (a) $E[Y] = \sum_y y\,p(y)$

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

$= \boxed{2.125}$

(b) $Var(Y) = \sum_y (y-\mu)^2 p(y)$

$= (0-2.125)^2 \cdot \frac{1}{8} + \cdots$

$= \boxed{1.109}$

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

$\boxed{z = -1.35}$

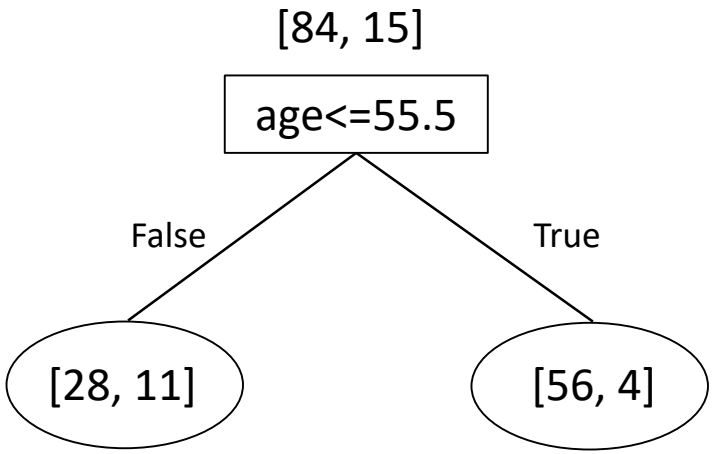
area $\Rightarrow$ p-value

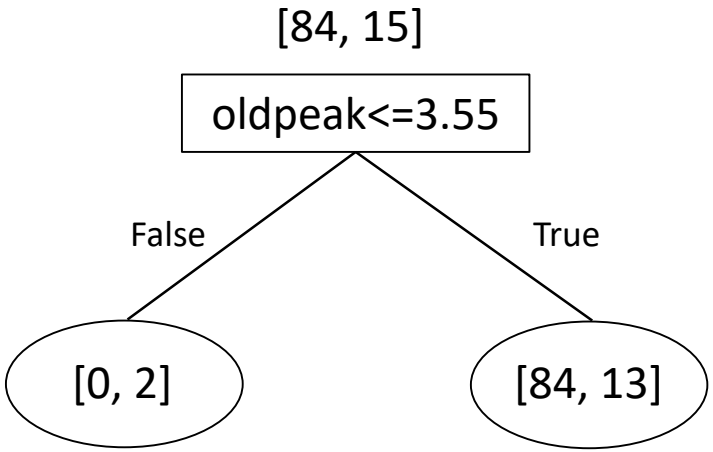
p-value = 0.08...

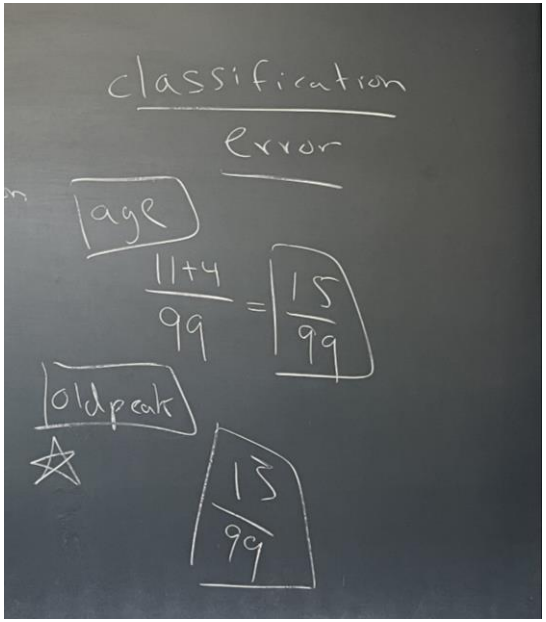
$\alpha = 0.05$

$\Rightarrow$ fail to reject H_0

Classification error







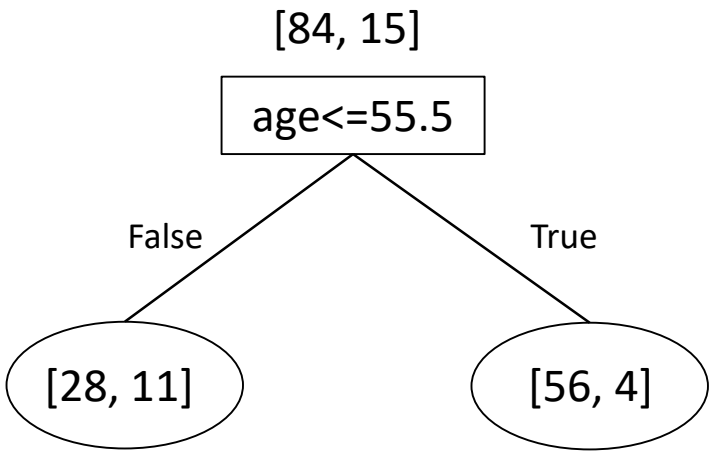
$$H(Y) = -\sum_{c \in vals(Y)} P(Y=c) \log_2 P(Y=c) \quad Y \in \{-1, +1\}$$

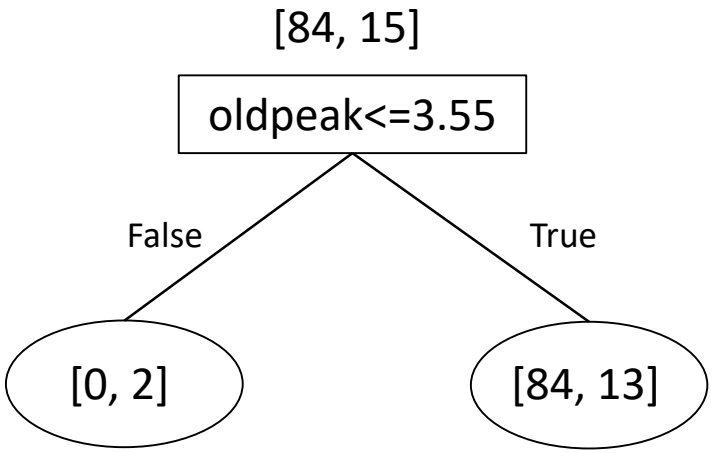
$$H(Y) = -\left(\frac{84}{99} \log_2 \frac{84}{99} + \frac{15}{99} \log_2 \frac{15}{99}\right) = 0.61$$

$$H(Y \mid oldpeak) = \frac{2}{99} \cancel{H(Y \mid oldpeak = F)} \to 0 + \frac{97}{99} H(Y \mid oldpeak = T)$$

$$H(Y \mid oldpeak = T) = -\left(\frac{84}{97} \log \frac{84}{97} + \frac{13}{97} \log \frac{13}{97}\right)$$

Entropy





H(Y) = 0.613619019599370 8

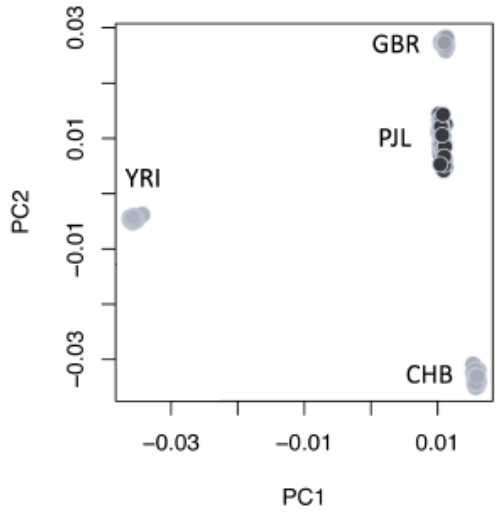
H(Y|age<=55.5) = 0.552248091053432 2

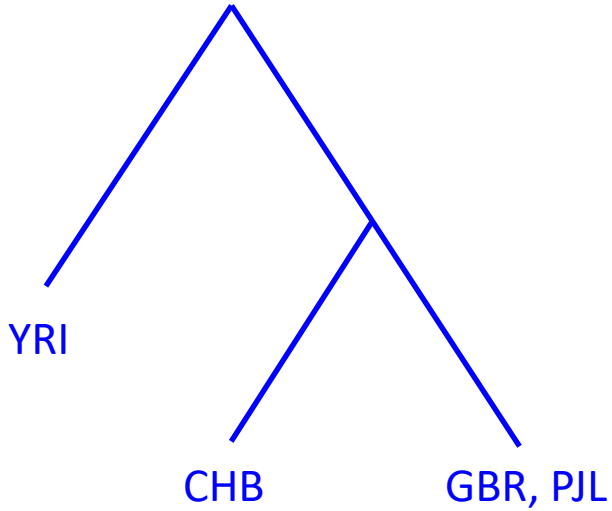
H(Y|oldpeak<=3.55) = 0.556880463059609 3

=> Age feature produces more information gain!

PCA

Consider the following four populations from the 1000 Genomes Project: CHB ("Chinese from Beijing"), GBR ("Great Britain"), PJL ("Punjabi from Lahore"), YRI ("Yoruba from Nigeria"). The first two principal components for this data are shown below:

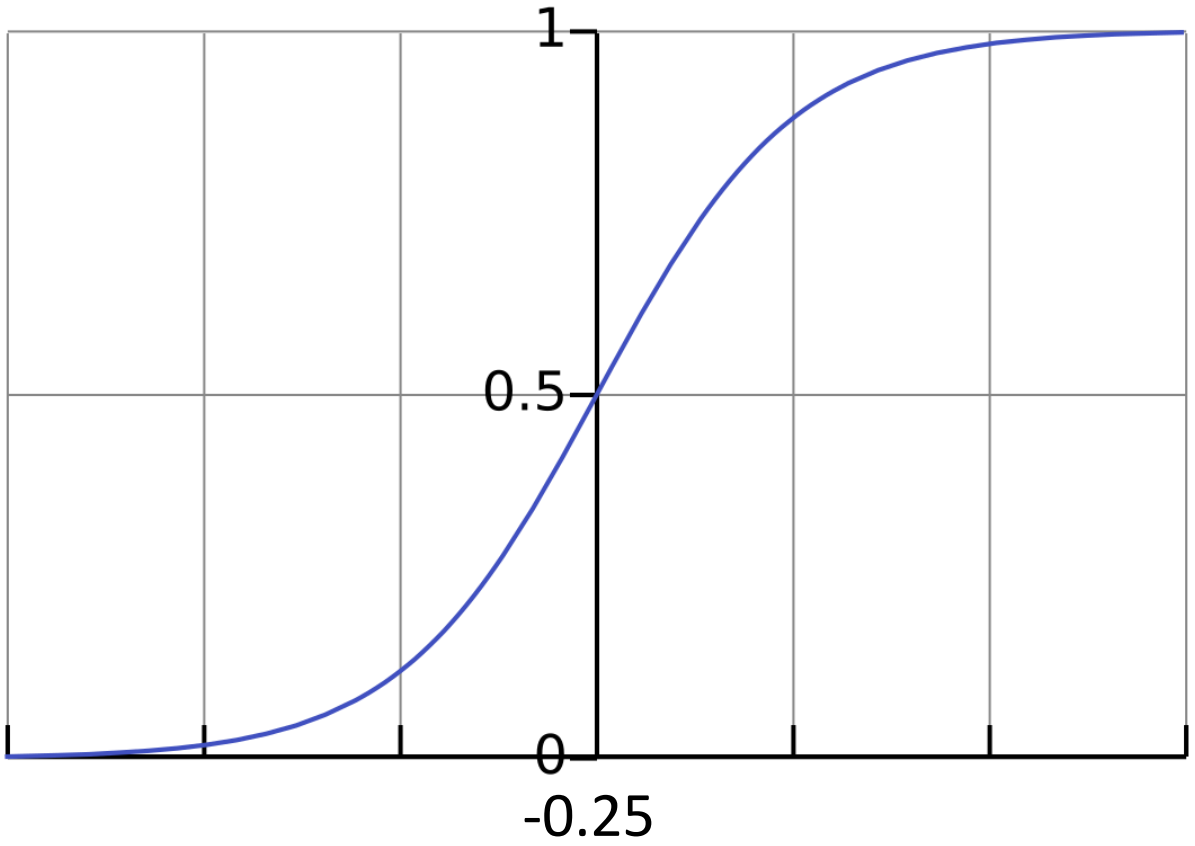




GBR & PJL are most closely related

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



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

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 $y = 1$