Bayesian Classification

Dr. Daniel B. Rowe
Professor of Computational Statistics
Department of Mathematical and Statistical Sciences
Marquette University

Outline

The Classification Problem

Bayesian Classification

Simplified Bayes Classification

Naïve Bayes Classification

Discussion

Homework

The Classification Problem

The classification problem arises when have an observation $\underset{p\times 1}{x}$ that has come from/belongs to one of C known classes, but we don't know which.

So we want to probabilistically assign $\underset{p\times 1}{x}$ to each of the C classes.
Let y denote the class that $\underset{p\times 1}{x}$ came from/belongs to, $y=1,\ldots,C$.
Then $P(Y=y)=f(y)$ is the probability that $\underset{p\times 1}{x}$ came from/belongs to class $\underset{p\times 1}{y}$.
i.e. $y=1$ for female and $y=2$ for male.

There is a probability distribution associated with observations from each class. We write $f(x|y,\theta_y)$ for the distribution of $\underset{p\times 1}{x}$ given it came from/belongs to class y and its parameters θ_y.

The Classification Problem

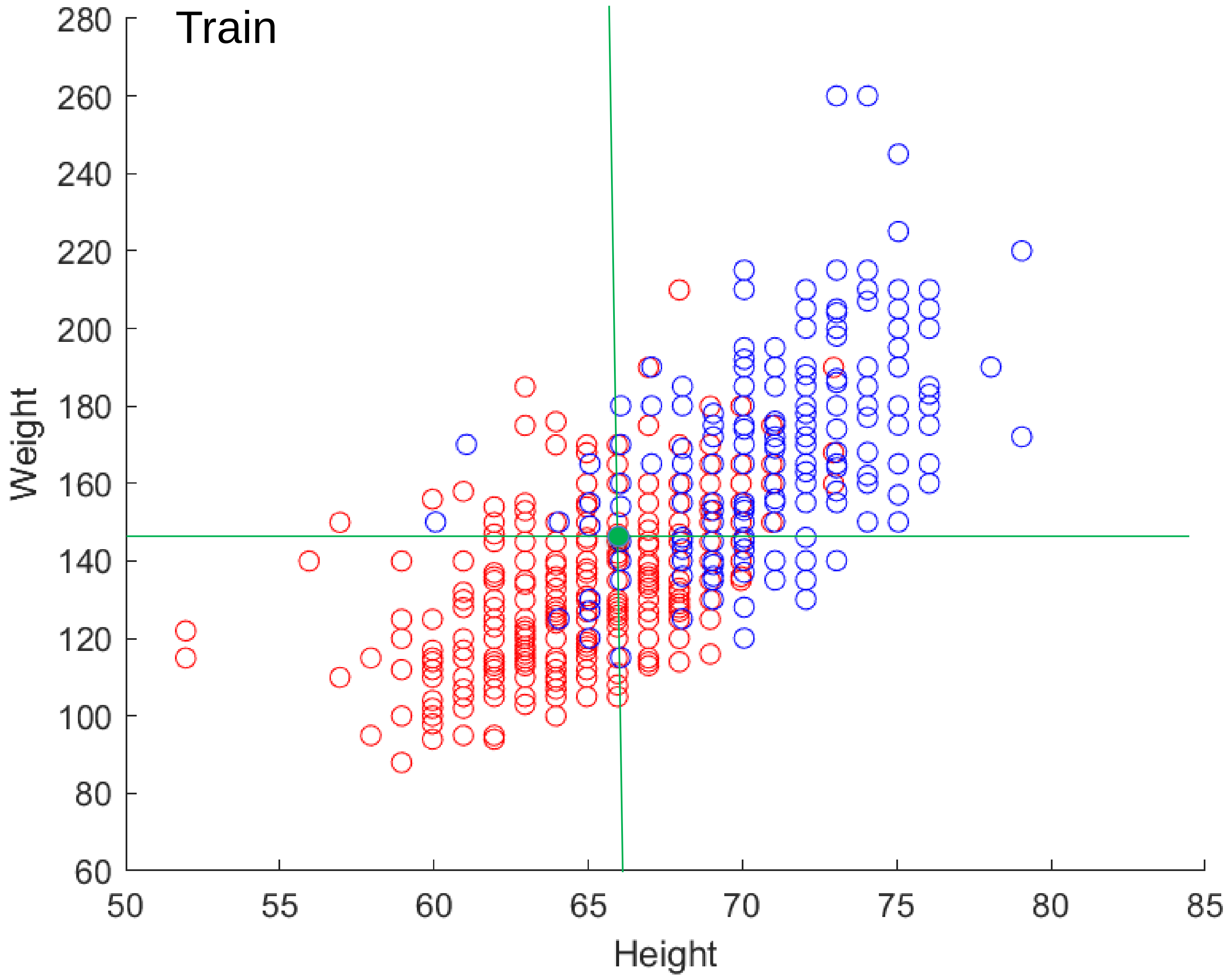
An observation $\underset{p \times 1}{x}$ came from/belongs to either the female or male class.

To the right are 683 student heights/weights with self reported gender.

$$y = 1, 2$$

I have a student that is 66 inches tall and weighs 145 pounds that did not report their gender. M or F?

$$\underset{2 \times 1}{x} = \begin{pmatrix} 66 \\ 145 \end{pmatrix}$$



Bayesian Classification

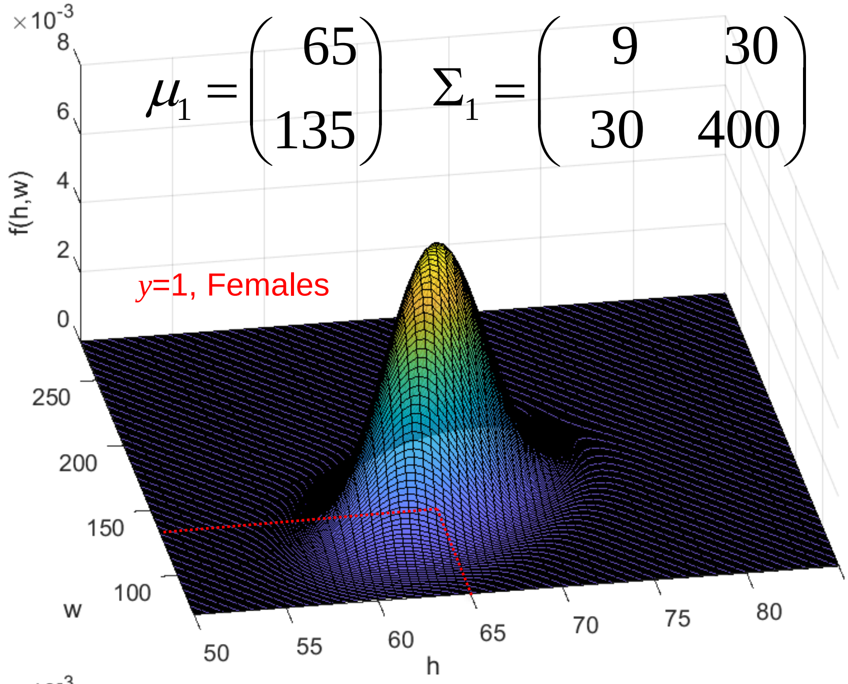
We often assume that the observation $\underset{p\times 1}{x}$ has a normal probability distribution function $f(x|y,\mu_y,\Sigma_y)$ given by

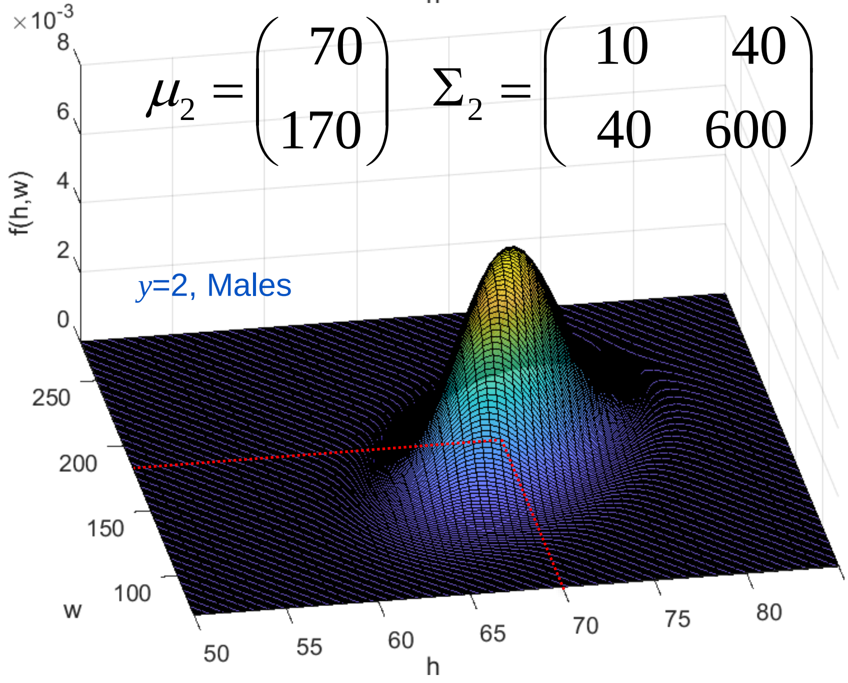
$$f(\underset{p\times 1}{x}\,|\,y,\underset{p\times 1}{\mu_y},\underset{p\times p}{\Sigma_y}) = (2\pi)^{-p/2}\,|\Sigma_y|^{-1/2}\,e^{-\frac{1}{2}(x-\mu_y)'\Sigma_y^{-1}(x-\mu_y)}$$

for $y=1,\ldots,C$. (But we don't have to.)

That is, if we knew that $\underset{p\times 1}{x}$ came from class y, then its PDF is as above with mean $\underset{p\times 1}{\mu_y}$ and covariance $\underset{p\times p}{\Sigma_y}$.

Example: $y=1,2$





Bayesian Classification

Generally we have prior knowledge about each of the C classes that we wish to probabilistically classify the observation $\underset{p\times 1}{x}$ into. This information is of the form of the class parameters, the mean $\underset{p\times 1}{\mu_y}$ and the covariance $\underset{p\times p}{\Sigma_y}$ for each class.

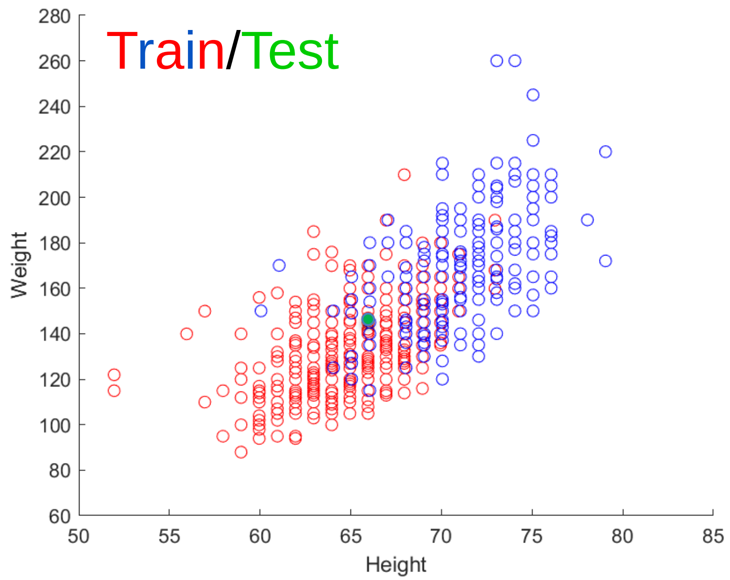
$$\underset{2\times 1}{x} = \begin{pmatrix} 66 \\ 145 \end{pmatrix}$$

Priors:

$$P(Y=y) = f(y)$$

$$f(\underset{p\times 1}{\mu_y} \mid \underset{p\times 1}{\mu_{0y}}, \underset{p\times p}{\Sigma_y}, n_y) = (2\pi)^{-p/2} \mid \Sigma_y / n_y \mid^{-1/2} e^{-\frac{n_y}{2}(\mu_y-\mu_{0y})'\Sigma_y^{-1}(\mu_y-\mu_{0y})}$$

$$f(\underset{p\times p}{\Sigma_y} \mid \underset{p\times p}{H_y}, \nu_y) = k_y \mid H_k \mid^{\nu_y/2} \mid \Sigma_y \mid^{-(\nu_y+p+1)/2} e^{-\frac{1}{2}tr\Sigma_y^{-1}H_y}$$



Bayesian Classification

If we multiply the priors and the likelihood together we obtain

$$f(\underbrace{x}_{A},\underbrace{y}_{B},\underbrace{\mu_y}_{C},\underbrace{\Sigma_y}_{D}) = f(\underbrace{x}_{A}|\underbrace{y}_{B},\underbrace{\mu_y}_{C},\underbrace{\Sigma_y}_{D})\, f(\underbrace{y}_{B})\, f(\underbrace{\mu_y}_{C}|\mu_{0y},\underbrace{\Sigma_y}_{D},n_y)\, f(\underbrace{\Sigma_y}_{D}|H_y,\nu_y)$$

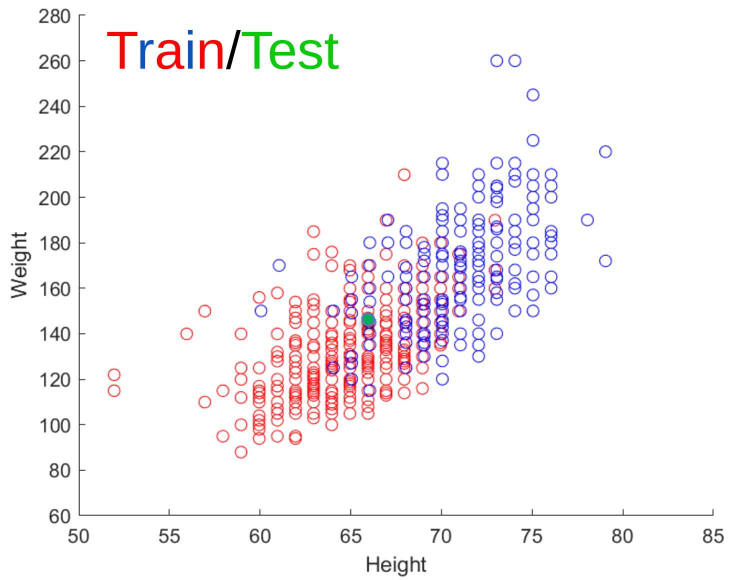
$$f(\underbrace{x}_{A},\underbrace{y}_{B},\underbrace{\mu_y}_{C},\underbrace{\Sigma_y}_{D}) = (2\pi)^{-p/2}|\Sigma_y|^{-1/2}\, e^{-\frac{1}{2}(x-\mu_y)'\Sigma_y^{-1}(x-\mu_y)}$$

$$\times f(y)$$

$$\times (2\pi)^{-p/2}|\Sigma_y/n_y|^{-1/2}\, e^{-\frac{n_y}{2}(\mu_y-\mu_{0y})'\Sigma_y^{-1}(\mu_y-\mu_{0y})}$$

$$\times k_y|H_y|^{\nu_y/2}|\Sigma_y|^{-(\nu_y+p+1)/2}\, e^{-\frac{1}{2}tr\Sigma_y^{-1}H_y}$$

$$\underset{2\times 1}{x} = \begin{pmatrix} 66 \\ 145 \end{pmatrix}$$



The posterior PDF of observations and parameters.

Bayesian Classification

We can integrate/sum over the parameter values

$$f(x,y,\mu_y,\Sigma_y) = f(x|y,\mu_y,\Sigma_y)\,f(y)\,f(\mu_y|\mu_{0y},\Sigma_y,n_y)\,f(\Sigma_y|H_y,\nu_y)$$

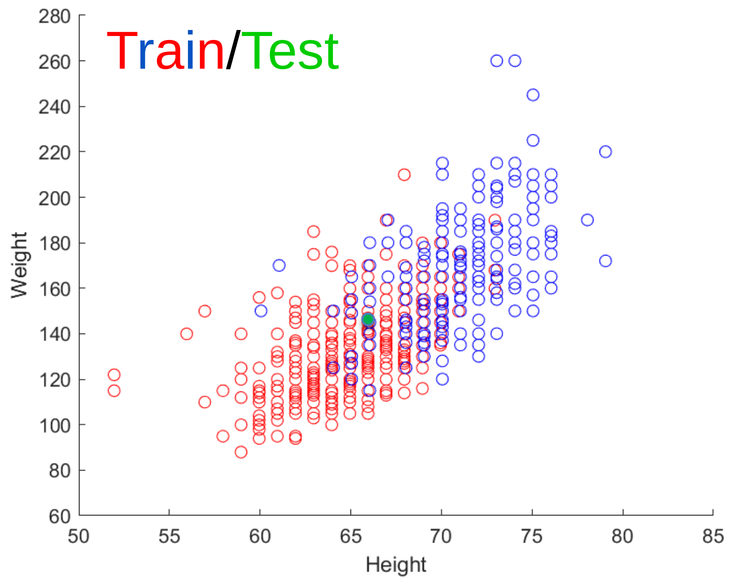
$$f(x) = \sum_{y=1}^{C}\int f(x,y,\mu_y,\Sigma_y)\,d\Sigma_y\,d\mu_y$$

$$\underset{2\times 1}{x} = \begin{pmatrix} 66 \\ 145 \end{pmatrix}$$

and divide

$$f(y,\mu_y,\Sigma_y|x) = \frac{f(x,y,\mu_y,\Sigma_y)}{f(x)}$$

to obtain the posterior distribution of the parameters.



Bayesian Classification

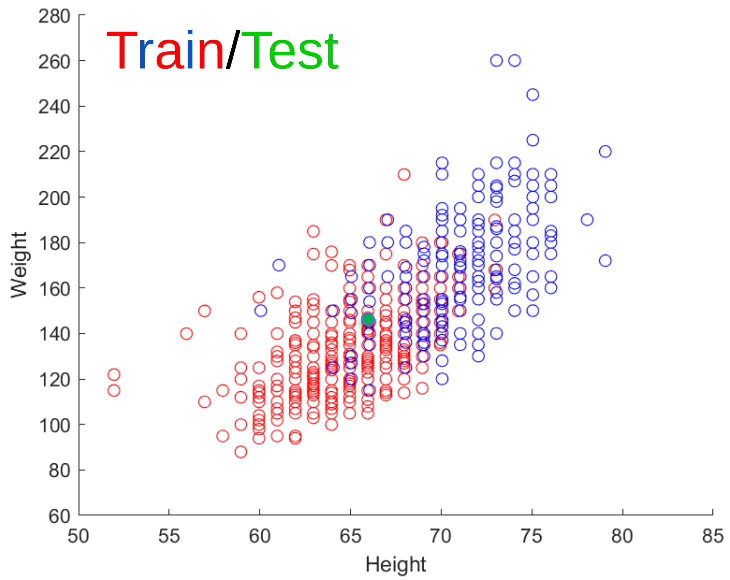
This posterior PDF of the parameters

$$f(y,\mu_y,\Sigma_y\,|\,x)=\frac{f(x,y,\mu_y,\Sigma_y)}{f(x)}$$

can then be integrated wrt Σ_y and then μ_y for each class y

$$\underset{2\times 1}{x}=\begin{pmatrix}66\\145\end{pmatrix}$$

$$f(y\,|\,x)=\int f(y,\mu_y,\Sigma_y\,|\,x)\,d\Sigma_y\,d\mu_y$$



so we can probabilistically classify the observation $\underset{p\times 1}{x}$.

$$f(y\,|\,x)=P(Y=y\,|\,x)\,,\quad y=1,\ldots,C\;.$$

Bayesian Classification

With our likelihood and conjugate priors, the joint PDF becomes

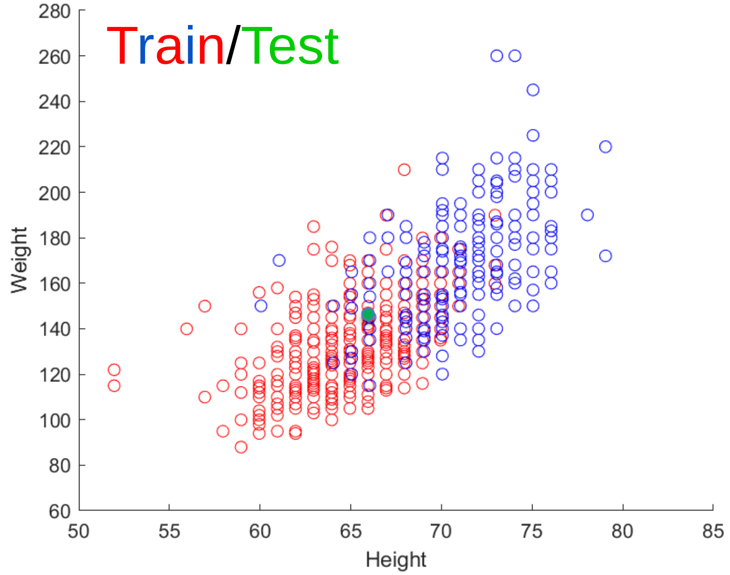
$$f(x,y,\mu_y,\Sigma_y) = f(y)(2\pi)^{-2p/2} n_y^{1/2} k_y |H_y|^{\nu_y/2} |\Sigma_y|^{-(\nu_y+p+3)/2} e^{-\frac{1}{2}tr\Sigma_y^{-1}[(\mu_y-\hat{\mu}_y)(\mu_y-\hat{\mu}_y)'+W]}$$

where $W = n_{0y}\mu_{0y}\mu_{0y}' + H_y + x'x - (n_{0y}\mu_{0y} + x)(n_{0y}\mu_{0y} + x)' / (n_{0y} + n)$

$$\underset{2\times 1}{x} = \begin{pmatrix} 66 \\ 145 \end{pmatrix}$$

which we first integrate wrt Σ_y by forming an inverse Wishart PDF which yields a multivariate student-t PDF factor for μ_y, then we integrate wrt μ_y .

The normalizing constants that remain are $f(x,y)$. We then sum over y to obtain $f(x)$.



Simplified Bayes Classification

However, in practice the full Bayesian statistical process previously described is not implemented by "Data Scientists."

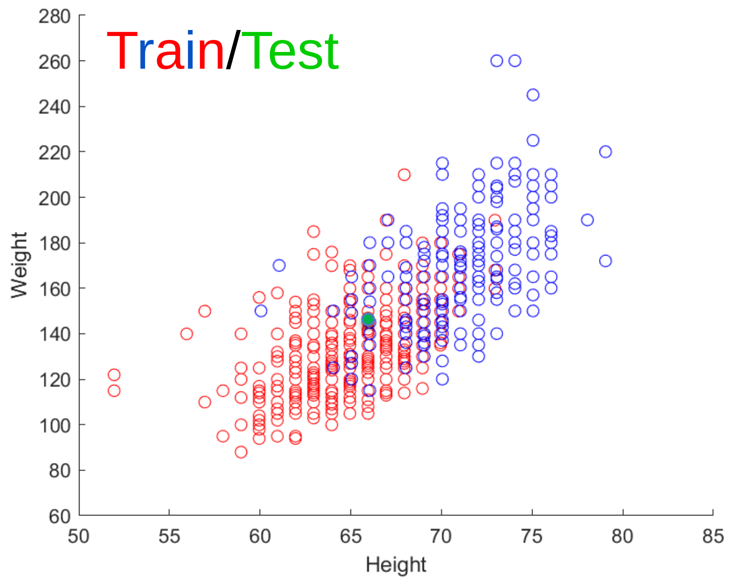
FYI: Statistics is the science of data.

Generally, we have previous "training" data of both observation x and the class it is from y that we assess our parameters from.

$$\underset{2\times 1}{x} = \begin{pmatrix} 66 \\ 145 \end{pmatrix}$$

The mean μ_y and covariance Σ_y for each population are Estimated via MLE, $(\hat{\mu}_y, \hat{\Sigma}_y)$, then we reinsert to obtain

$$f(\underset{p\times 1}{x} \mid y, \underset{p\times 1}{\hat{\mu}_y}, \underset{p\times p}{\hat{\Sigma}_y}) = (2\pi)^{-p/2} \mid \hat{\Sigma}_y \mid^{-1/2} e^{-\frac{1}{2}(x-\hat{\mu}_y)'\hat{\Sigma}_y^{-1}(x-\hat{\mu}_y)} \qquad y=1,2 \quad .$$



Simplified Bayes Classification

Using the conditional probability distributions

$$f(x \mid y, \hat{\mu}_y, \hat{\Sigma}_y) = (2\pi)^{-p/2} \, |\hat{\Sigma}_y|^{-1/2} \, e^{-\frac{1}{2}(x-\hat{\mu}_y)'\hat{\Sigma}_y^{-1}(x-\hat{\mu}_y)} \qquad y=1,2$$

and Bayes' Rule

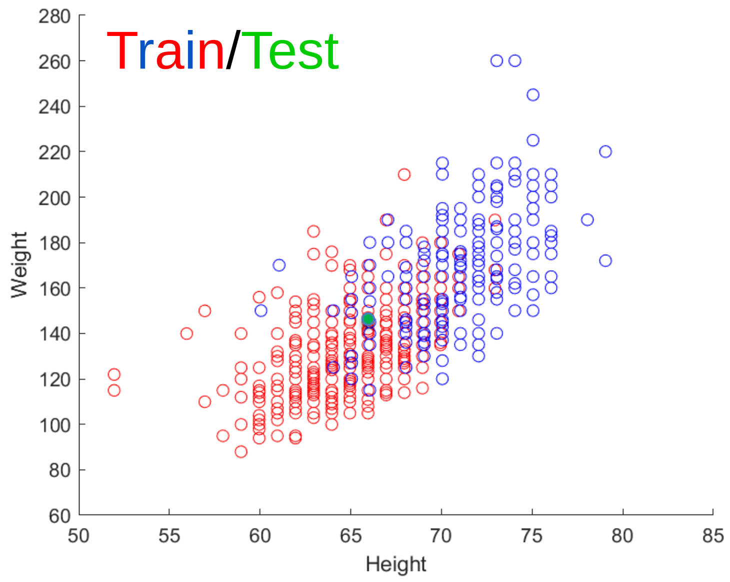
$$f(y \mid x, \hat{\mu}_y, \hat{\Sigma}_y) \propto f(x \mid y, \hat{\mu}_y, \hat{\Sigma}_y) \, f(y) \qquad y=1,2$$

$$\underset{2\times 1}{x} = \begin{pmatrix} 66 \\ 145 \end{pmatrix}$$

we now compute a MAP estimator

$$\underset{y}{\text{ArgMax}} \; f(y \mid x, \hat{\mu}_y, \hat{\Sigma}_y)$$

for classification.



Simplified Bayes Classification

Example:

Have students heights/weights with known class, $(\underset{p\times 1}{x_i},\underset{1\times 1}{y_i})$, $i=1,\ldots,n_0=683$.

Used this "training" data to estimate μ_y and Σ_y for each class.

Have another 633 that we want to probabilistically classify.

(Actually know true reported classes.)

Estimated class means and covariances using MLE to be

$$\hat{\mu}_1 = \begin{pmatrix} 65.1128 \\ 133.3850 \end{pmatrix} \quad \hat{\Sigma}_1 = \begin{pmatrix} 9.1824 & 26.2469 \\ 26.2469 & 335.3592 \end{pmatrix}$$

$$P(Y=1)=.6618$$

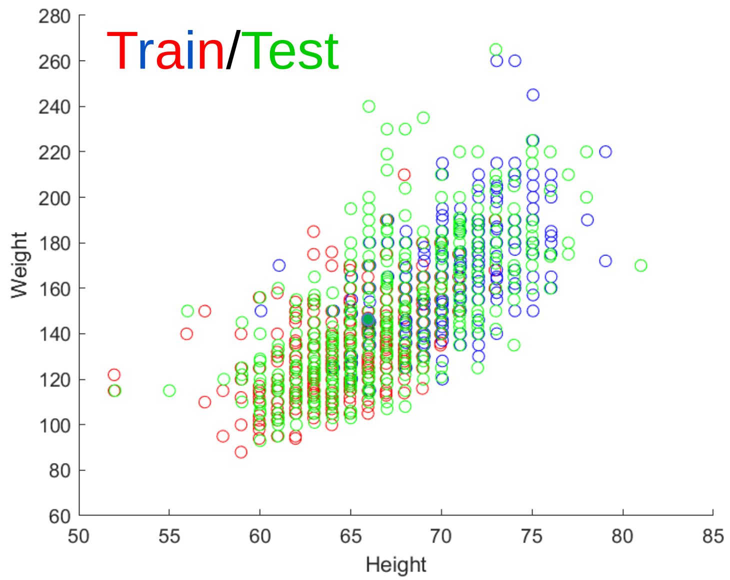
66.18% of past students female

$$\hat{\mu}_2 = \begin{pmatrix} 71.1082 \\ 170.3247 \end{pmatrix} \quad \hat{\Sigma}_2 = \begin{pmatrix} 9.4882 & 38.7343 \\ 38.7343 & 591.8724 \end{pmatrix}$$

$$P(Y=2)=.3382$$

33.82% of past students male

633 to classify



Simplified Bayes Classification

Example: Classify $x = \begin{pmatrix} 66 \\ 145 \end{pmatrix}$

$$f(y \mid x, \hat{\mu}_y, \hat{\Sigma}_y) \propto f(x \mid y, \hat{\mu}_y, \hat{\Sigma}_y) f(y)$$

$$f(y \mid x, \hat{\mu}_1, \hat{\Sigma}_1) \propto 0.0109 = \boxed{0.8947}$$

$$f(y \mid x, \hat{\mu}_2, \hat{\Sigma}_2) \propto 0.0014 = \boxed{0.1053}$$

$$\hat{\mu}_1 = \begin{pmatrix} 65.1128 \\ 133.3850 \end{pmatrix} \quad \hat{\Sigma}_1 = \begin{pmatrix} 9.1824 & 26.2469 \\ 26.2469 & 335.3592 \end{pmatrix}$$

$$\hat{\mu}_2 = \begin{pmatrix} 71.1082 \\ 170.3247 \end{pmatrix} \quad \hat{\Sigma}_2 = \begin{pmatrix} 9.4882 & 38.7343 \\ 38.7343 & 591.8724 \end{pmatrix}$$

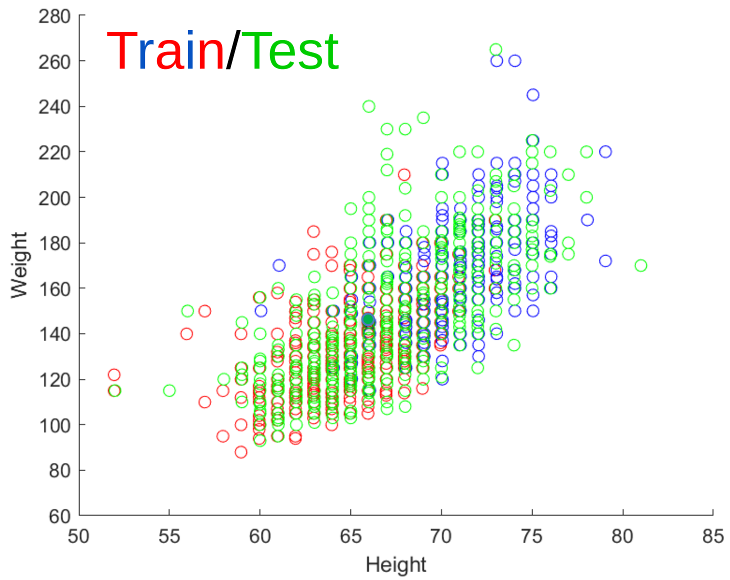
$$P(Y=1) = .6618$$

66.18% of past students female

$$P(Y=2) = .3382$$

33.82% of past students male

633 to classify



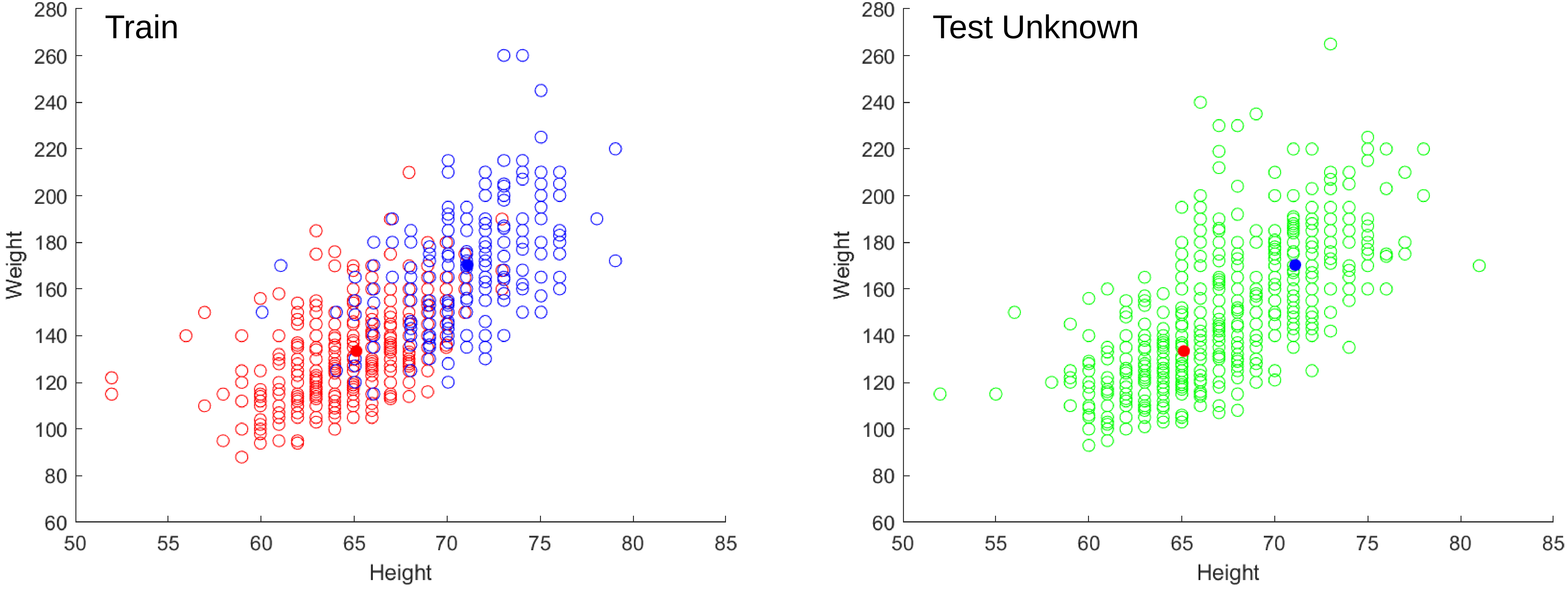
Simplified Bayes Classification

$$\hat{\mu}_1 = \begin{pmatrix} 65.1128 \\ 133.3850 \end{pmatrix} \quad \hat{\mu}_2 = \begin{pmatrix} 71.1082 \\ 170.3247 \end{pmatrix}$$

Example: Classify

$$f(y \mid x, \hat{\mu}_y, \hat{\Sigma}_y) \propto f(x \mid y, \hat{\mu}_y, \hat{\Sigma}_y) f(y)$$

"distance" an observation is from each class mean

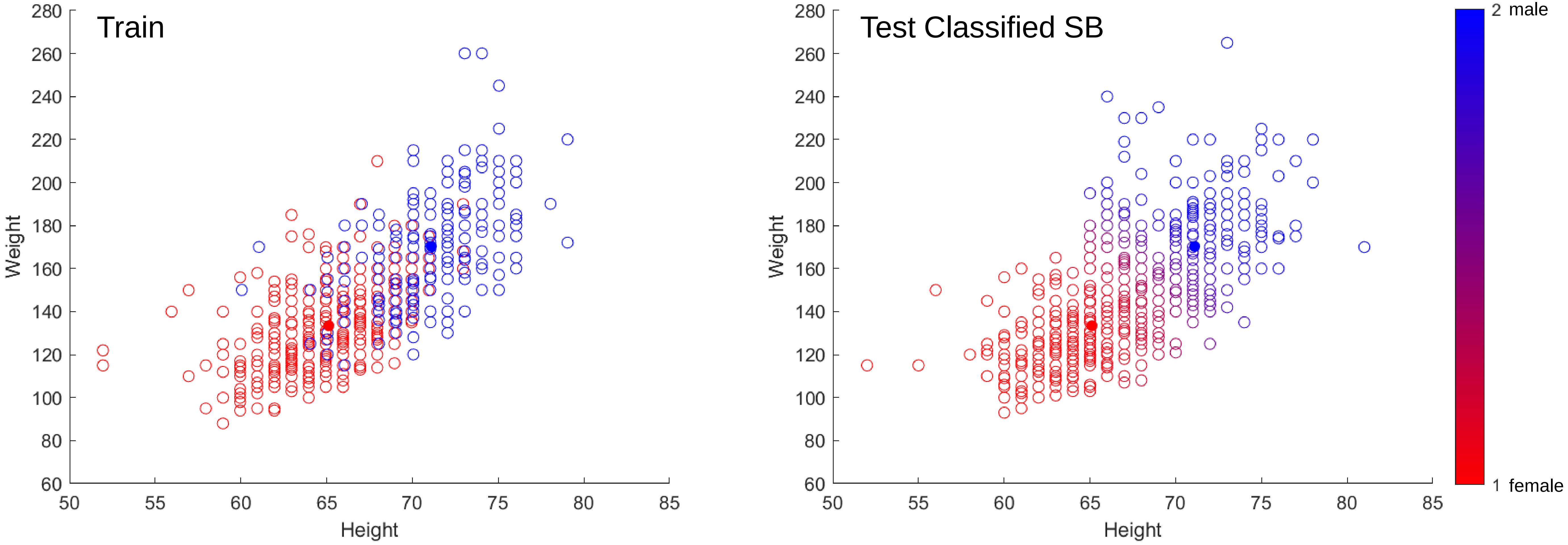


Simplified Bayes Classification

$$\hat{\mu}_1 = \begin{pmatrix} 65.1128 \\ 133.3850 \end{pmatrix} \quad \hat{\mu}_2 = \begin{pmatrix} 71.1082 \\ 170.3247 \end{pmatrix}$$

Example: Classify

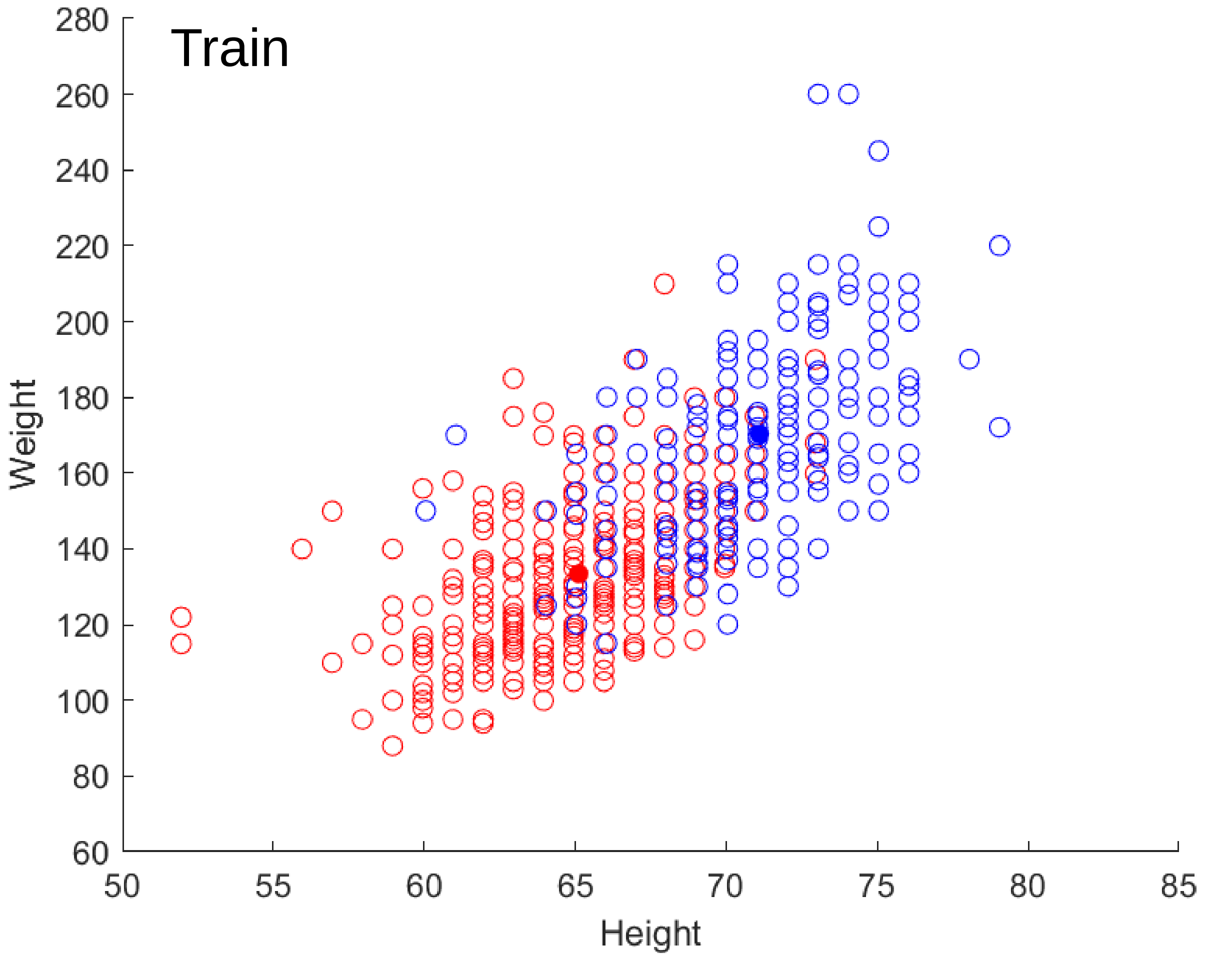
$$f(y \mid x, \hat{\mu}_y, \hat{\Sigma}_y) \propto f(x \mid y, \hat{\mu}_y, \hat{\Sigma}_y) f(y)$$

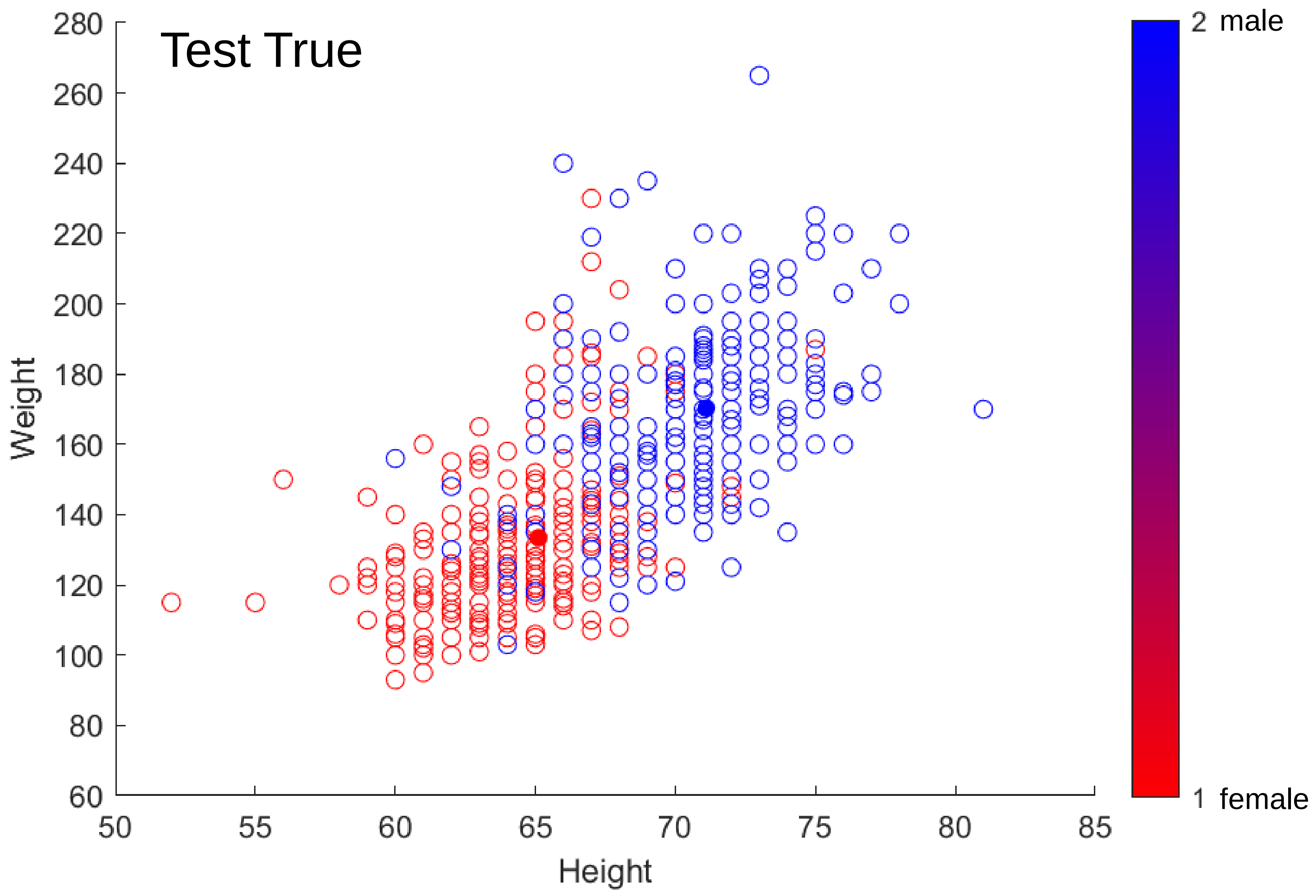


Simplified Bayes Classification

$$\hat{\mu}_1 = \begin{pmatrix} 65.1128 \\ 133.3850 \end{pmatrix} \quad \hat{\mu}_2 = \begin{pmatrix} 71.1082 \\ 170.3247 \end{pmatrix}$$

Example: Classify





Naïve Bayesian Classification

Naïve Bayes classification is a simpler form of the previous Bayesian classification. It assumes independence between the variables, elements of x.
i.e assumes independence between height and weight.

$$f(\underset{p\times 1}{x} \mid y, \underset{p\times 1}{\hat{\mu}_y}, \underset{p\times p}{\hat{\Sigma}_y}) = (2\pi)^{-p/2} \mid \hat{\Sigma}_y \mid^{-1/2} e^{-\frac{1}{2}(x-\hat{\mu}_y)'\hat{\Sigma}_y^{-1}(x-\hat{\mu}_y)}$$

becomes

$$f(x \mid y, \hat{\mu}_y, \hat{\sigma}_{y1}^2, \ldots, \hat{\sigma}_{yp}^2) = (2\pi)^{-p/2}\left(\prod_{j=1}^{p}(\sigma_j^2)^{-1/2}\right) e^{-\frac{1}{2}\sum_{j=1}^{p}(x_j-\hat{\mu}_{yj})^2/\sigma_j^2}$$

Naïve Bayesian Classification

Using the probability distributions

$$f(x|y,\hat{\mu}_y,\hat{\sigma}^2_{y1},...,\hat{\sigma}^2_{yp}) = (2\pi)^{-p/2}\left(\prod_{j=1}^{p}(\sigma_j^2)^{-1/2}\right)e^{-\frac{1}{2}\sum_{j=1}^{p}(x_j-\hat{\mu}_{yj})^2/\sigma_j^2}$$

$y=1,2$

and Bayes' Rule

$$f(y|x,\hat{\mu}_y,\hat{\sigma}^2_{y1},...,\hat{\sigma}^2_{yp}) \propto f(x|y,\hat{\mu}_y,\hat{\sigma}^2_{y1},...,\hat{\sigma}^2_{yp})f(y)$$

$y=1,2$

we now compute a MAP estimator

$$\underset{y}{\text{ArgMax}}\ f(y|x,\hat{\mu}_y,\hat{\sigma}^2_{y1},...,\hat{\sigma}^2_{yp})$$

for classification.

Naïve Bayesian Classification

Example:

Have students heights/weights with known class, (x_i,y_i), $i=1,\ldots,n_0=683$.

Used this "training" data to estimate μ_y and Σ_y for each class.

Have another 633 that we want to probabilistically classify.

(Actually know true reported classes.)

Estimated class means and variances using MLE to be

$$\hat{\mu}_1=\begin{pmatrix}65.1128\\133.3850\end{pmatrix}\qquad \hat{\sigma}_{11}^2=9.1824\qquad \hat{\sigma}_{12}^2=335.3592$$

$$\hat{\mu}_2=\begin{pmatrix}71.1082\\170.3247\end{pmatrix}\qquad \hat{\sigma}_{21}^2=9.4882\qquad \hat{\sigma}_{22}^2=591.8724$$

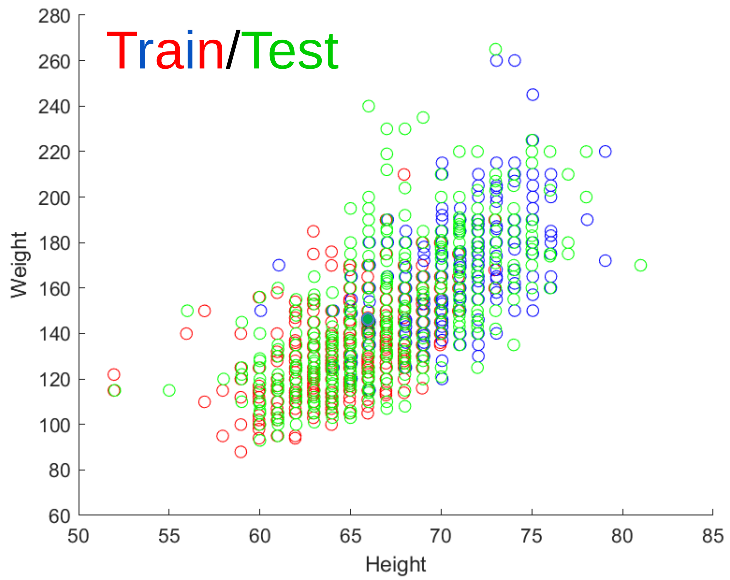
$$P(Y=1)=.6618$$

66.18% of past students female

$$P(Y=2)=.3382$$

33.82% of past students male

633 to classify



Naïve Bayesian Classification

$$x=\begin{pmatrix}66\\145\end{pmatrix}$$

Example:

$$f(y\,|\,x,\hat{\mu}_y,\hat{\sigma}_{y1}^2,\hat{\sigma}_{y2}^2)\propto f(x\,|\,y,\hat{\mu}_y,\hat{\sigma}_{y1}^2,\hat{\sigma}_{y2}^2)\,f(y)$$

$$f(y\,|\,x,\hat{\mu}_1,\hat{\sigma}_{11}^2,\hat{\sigma}_{12}^2)\propto 0.0130 = \boxed{0.9085}$$

$$f(y\,|\,x,\hat{\mu}_2,\hat{\sigma}_{21}^2,\hat{\sigma}_{22}^2)\propto 0.0013 = \boxed{0.0915}$$

$$\hat{\mu}_1=\begin{pmatrix}65.1128\\133.3850\end{pmatrix}\qquad \hat{\sigma}_{11}^2=9.1824\qquad \hat{\sigma}_{12}^2=335.3592$$

$$\hat{\mu}_2=\begin{pmatrix}71.1082\\170.3247\end{pmatrix}\qquad \hat{\sigma}_{21}^2=9.4882\qquad \hat{\sigma}_{22}^2=591.8724$$

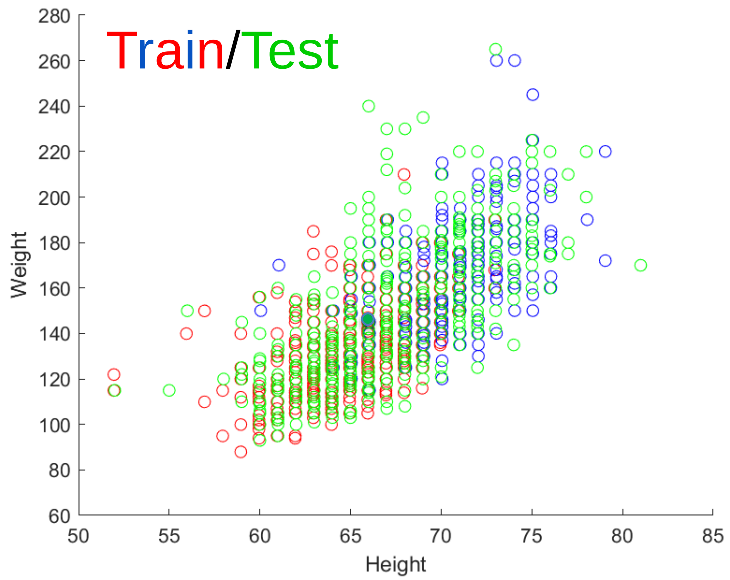
$$P(Y=1)=.6618$$

66.18% of past students female

$$P(Y=2)=.3382$$

33.82% of past students male

633 to classify

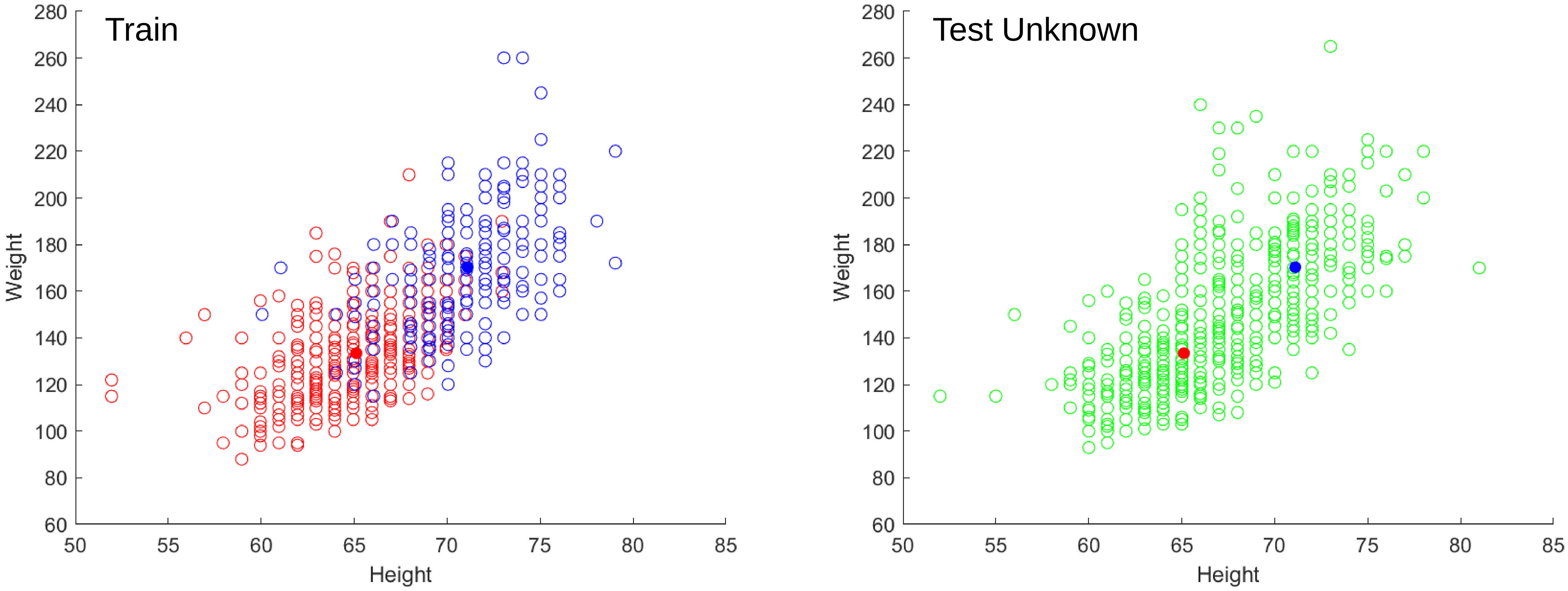


Simplified Bayes Classification

$$\hat{\mu}_1 = \begin{pmatrix} 65.1128 \\ 133.3850 \end{pmatrix} \quad \hat{\mu}_2 = \begin{pmatrix} 71.1082 \\ 170.3247 \end{pmatrix}$$

Example: Classify

$$f(y \mid x, \hat{\mu}_y, \hat{\sigma}^2_{y1}, \hat{\sigma}^2_{y2}) \propto f(x \mid y, \hat{\mu}_y, \hat{\sigma}^2_{y1}, \hat{\sigma}^2_{y2}) f(y)$$

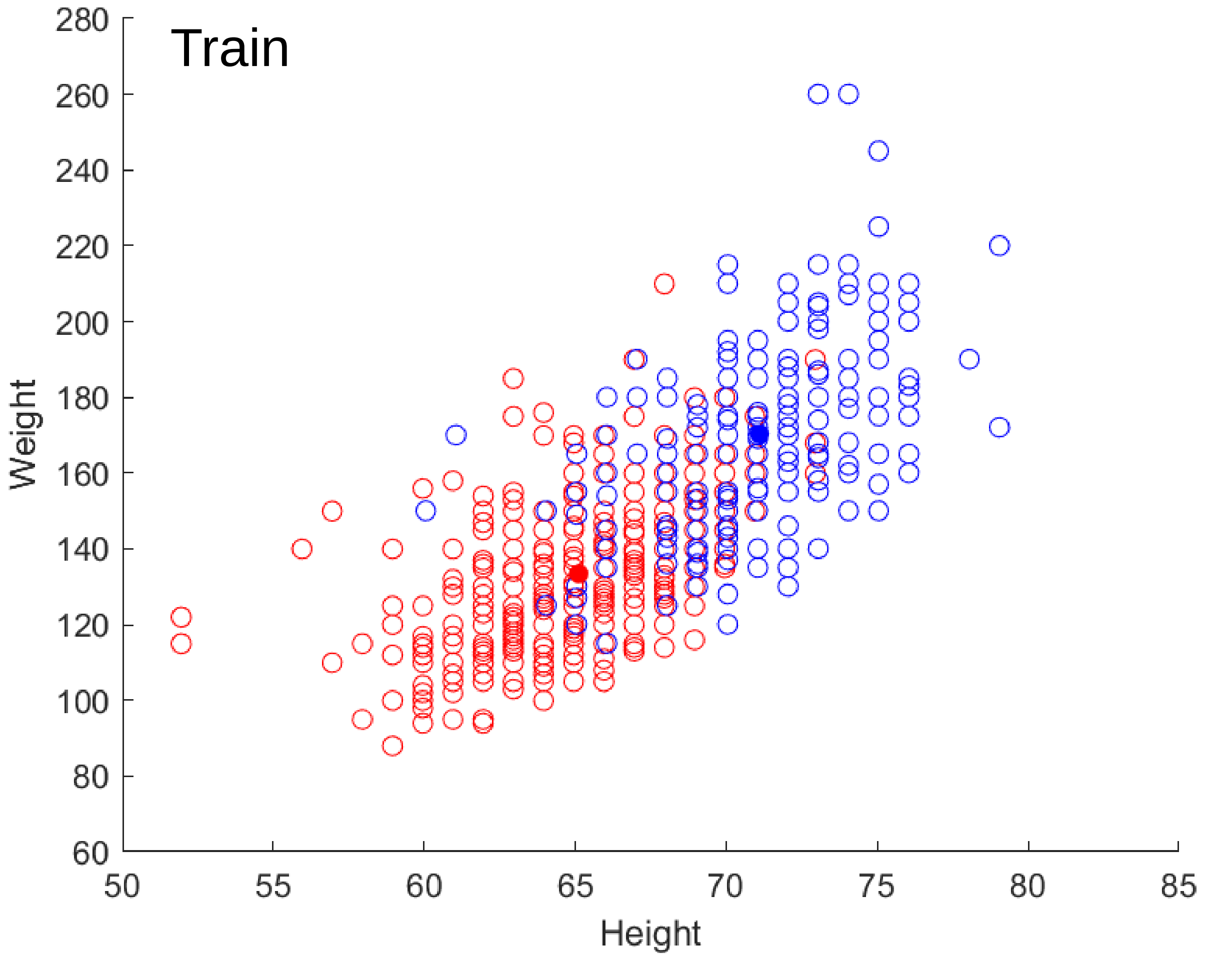


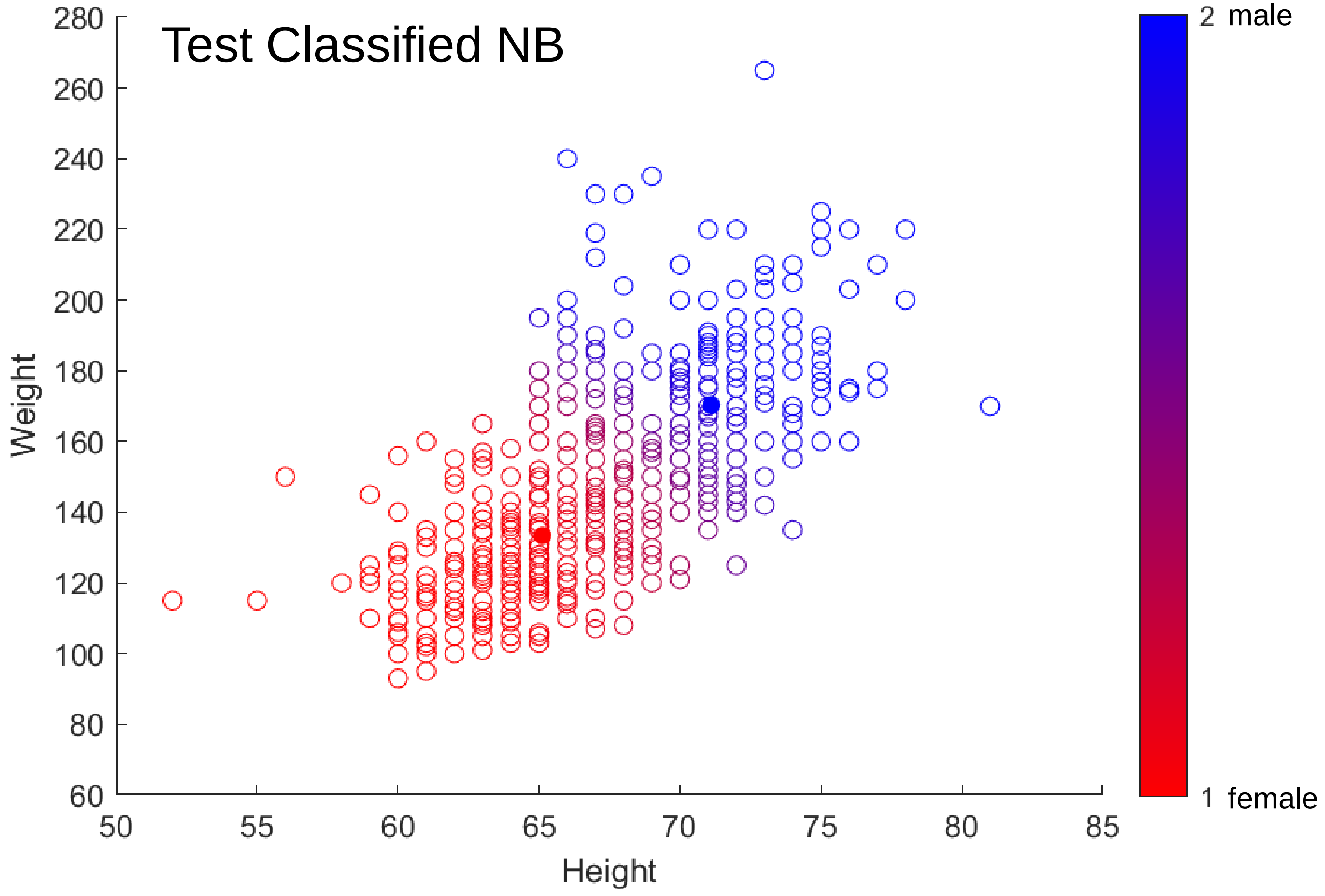
Simplified Bayes Classification

$$\hat{\mu}_1 = \begin{pmatrix} 65.1128 \\ 133.3850 \end{pmatrix} \quad \hat{\mu}_2 = \begin{pmatrix} 71.1082 \\ 170.3247 \end{pmatrix}$$

Example: Classify

$$f(y \mid x, \hat{\mu}_y, \hat{\sigma}^2_{y1}, \hat{\sigma}^2_{y2}) \propto f(x \mid y, \hat{\mu}_y, \hat{\sigma}^2_{y1}, \hat{\sigma}^2_{y2}) f(y)$$

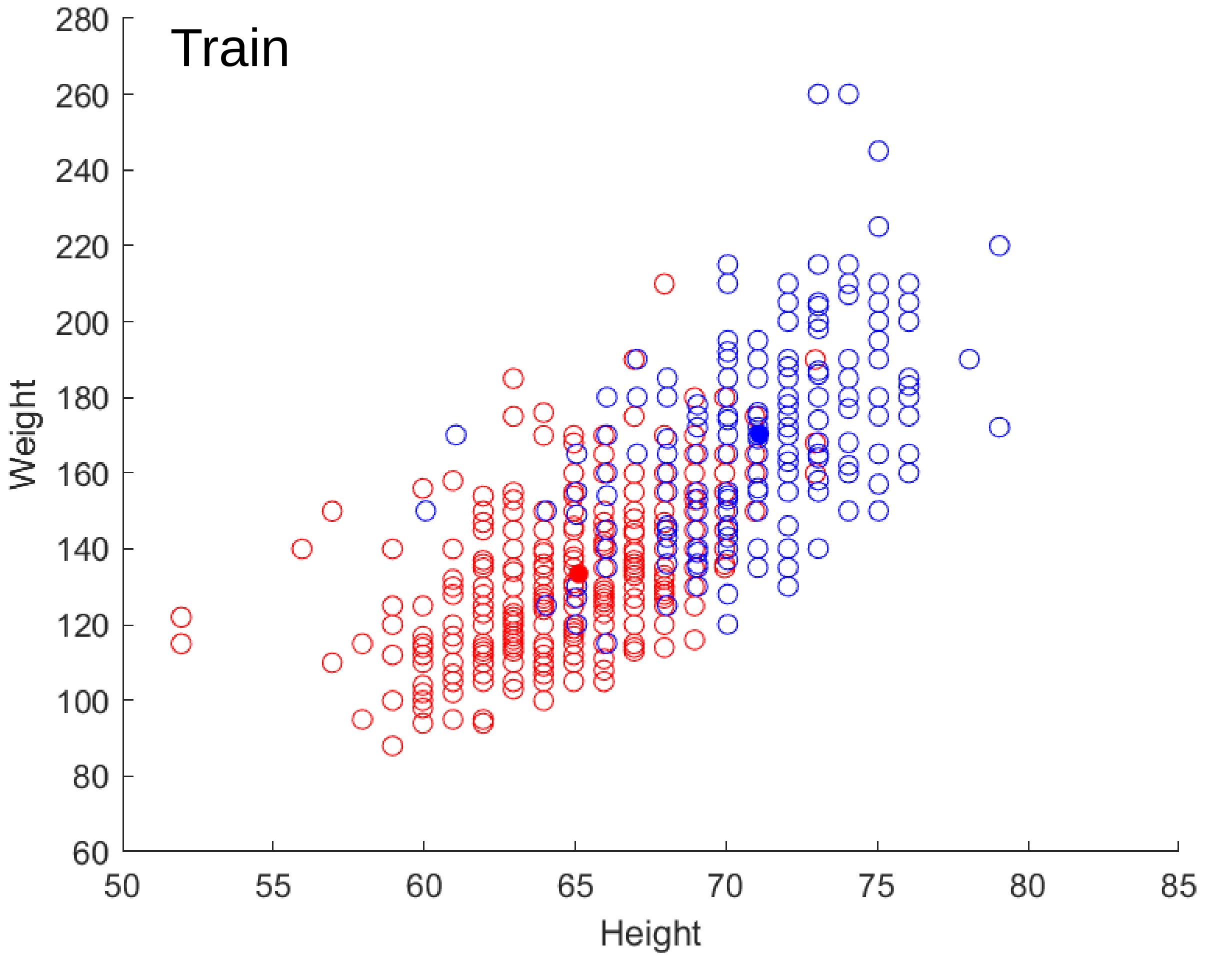


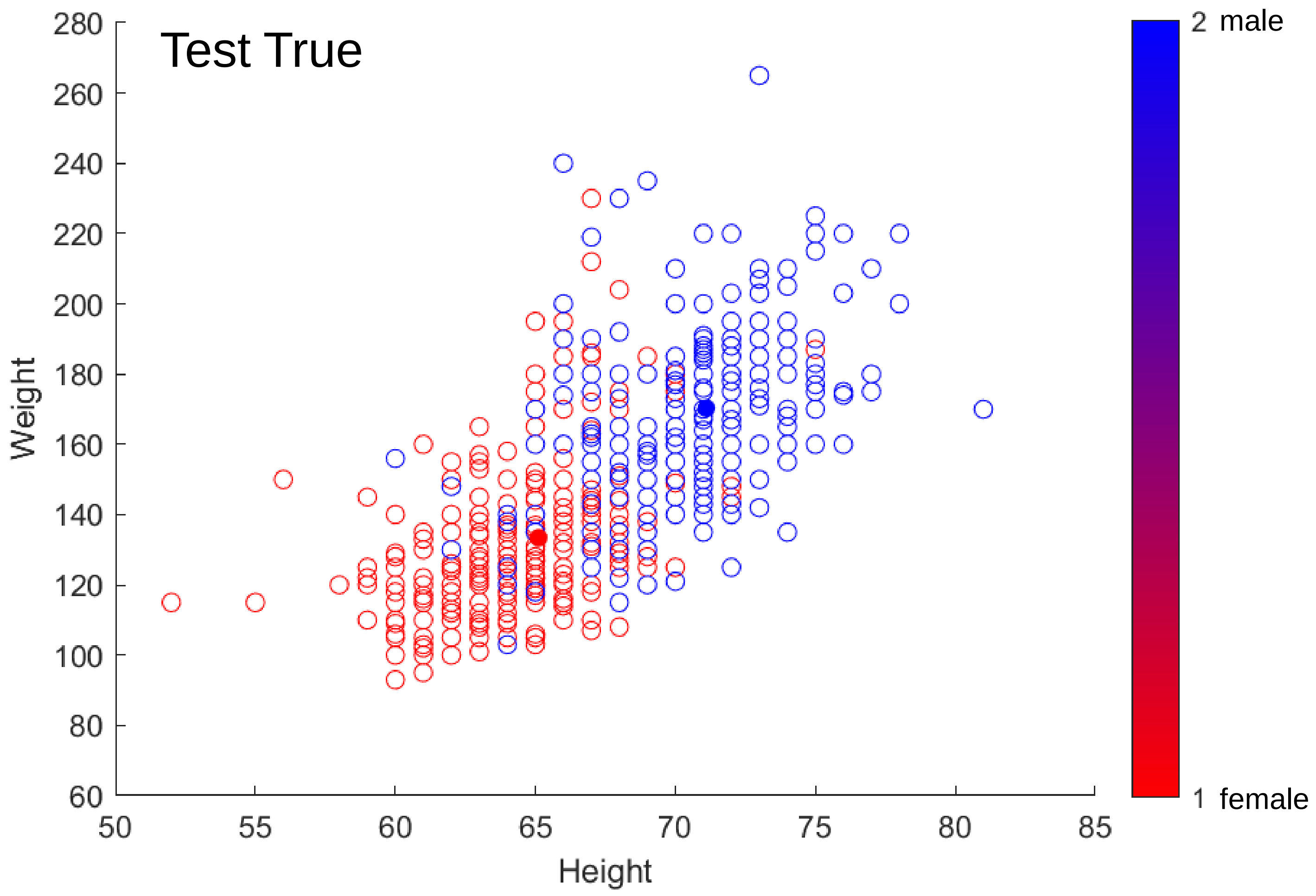


Simplified Bayes Classification

$$\hat{\mu}_1 = \begin{pmatrix} 65.1128 \\ 133.3850 \end{pmatrix} \quad \hat{\mu}_2 = \begin{pmatrix} 71.1082 \\ 170.3247 \end{pmatrix}$$

Example: Classify





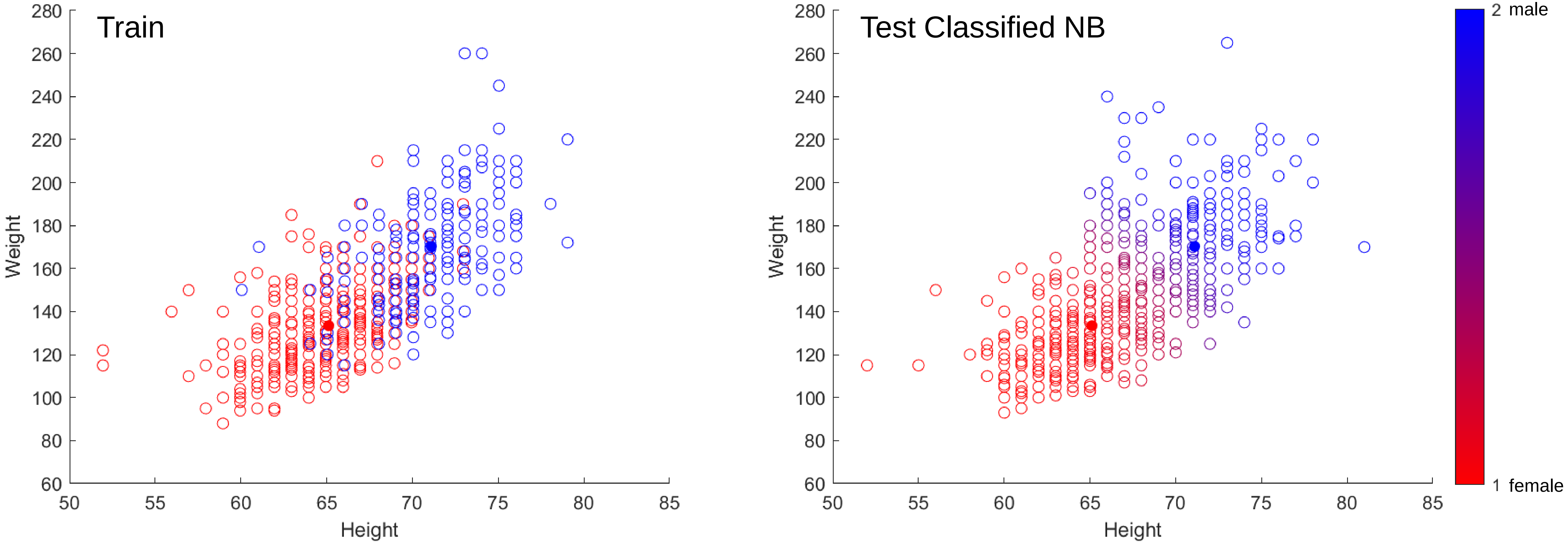
Toggle

Simplified Bayes Classification

$$\hat{\mu}_1 = \begin{pmatrix} 65.1128 \\ 133.3850 \end{pmatrix} \quad \hat{\mu}_2 = \begin{pmatrix} 71.1082 \\ 170.3247 \end{pmatrix}$$

Example: Classify

$$f(y \mid x, \hat{\mu}_y, \hat{\sigma}^2_{y1}, \hat{\sigma}^2_{y2}) \propto f(x \mid y, \hat{\mu}_y, \hat{\sigma}^2_{y1}, \hat{\sigma}^2_{y2}) f(y)$$

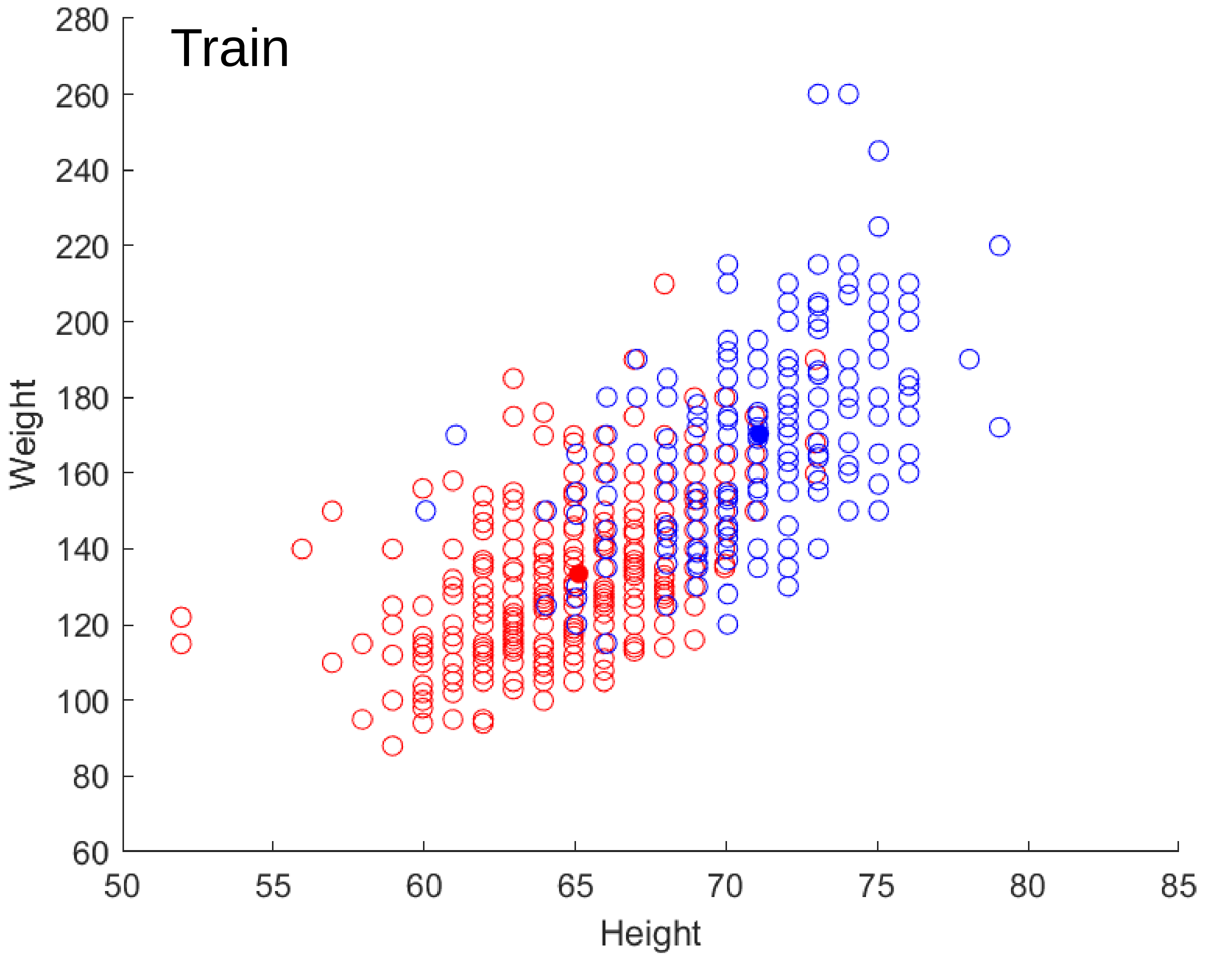


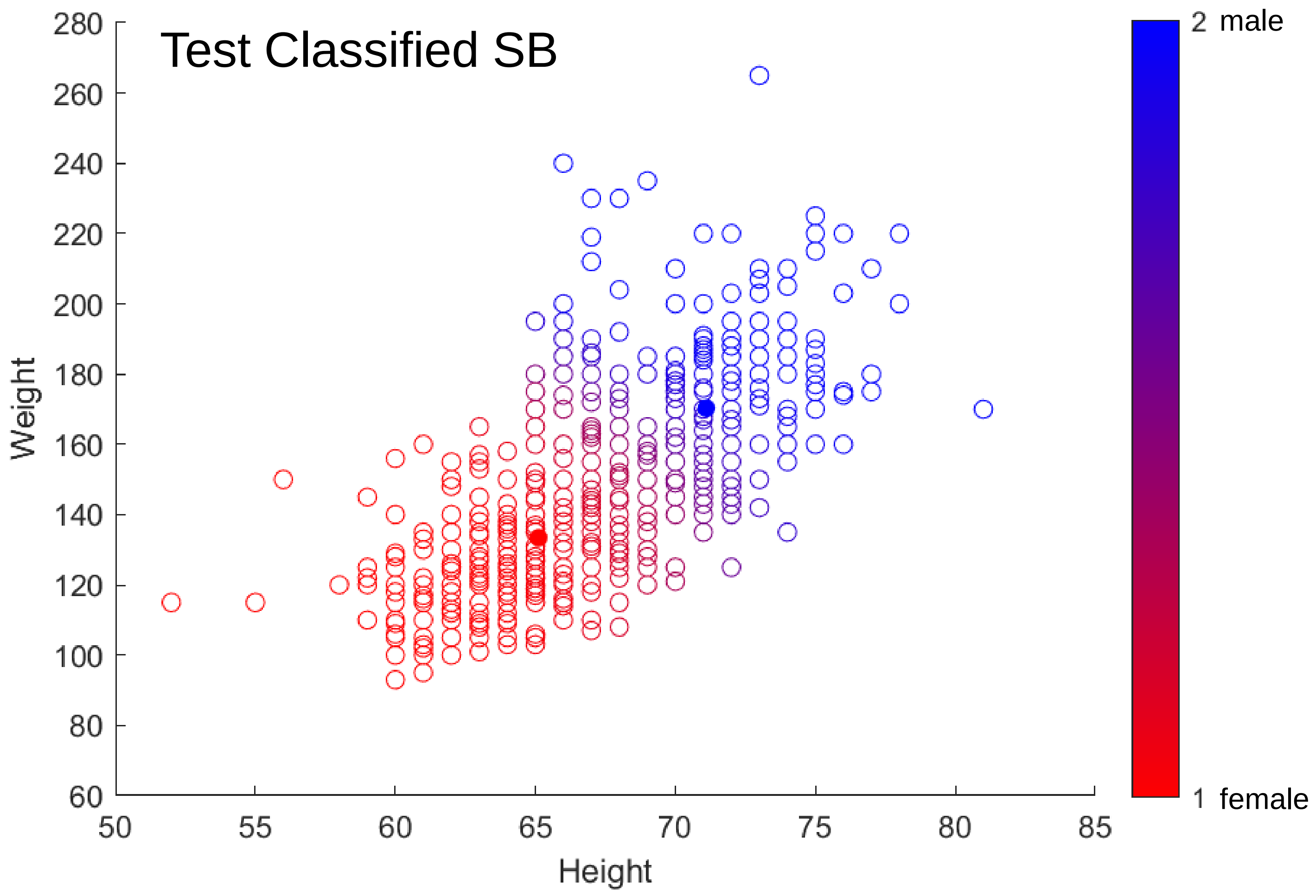
Simplified Bayes Classification

$$\hat{\mu}_1 = \begin{pmatrix} 65.1128 \\ 133.3850 \end{pmatrix} \quad \hat{\mu}_2 = \begin{pmatrix} 71.1082 \\ 170.3247 \end{pmatrix}$$

Example: Classify

$$f(y \mid x, \hat{\mu}_y, \hat{\Sigma}_y) \propto f(x \mid y, \hat{\mu}_y, \hat{\Sigma}_y) f(y)$$

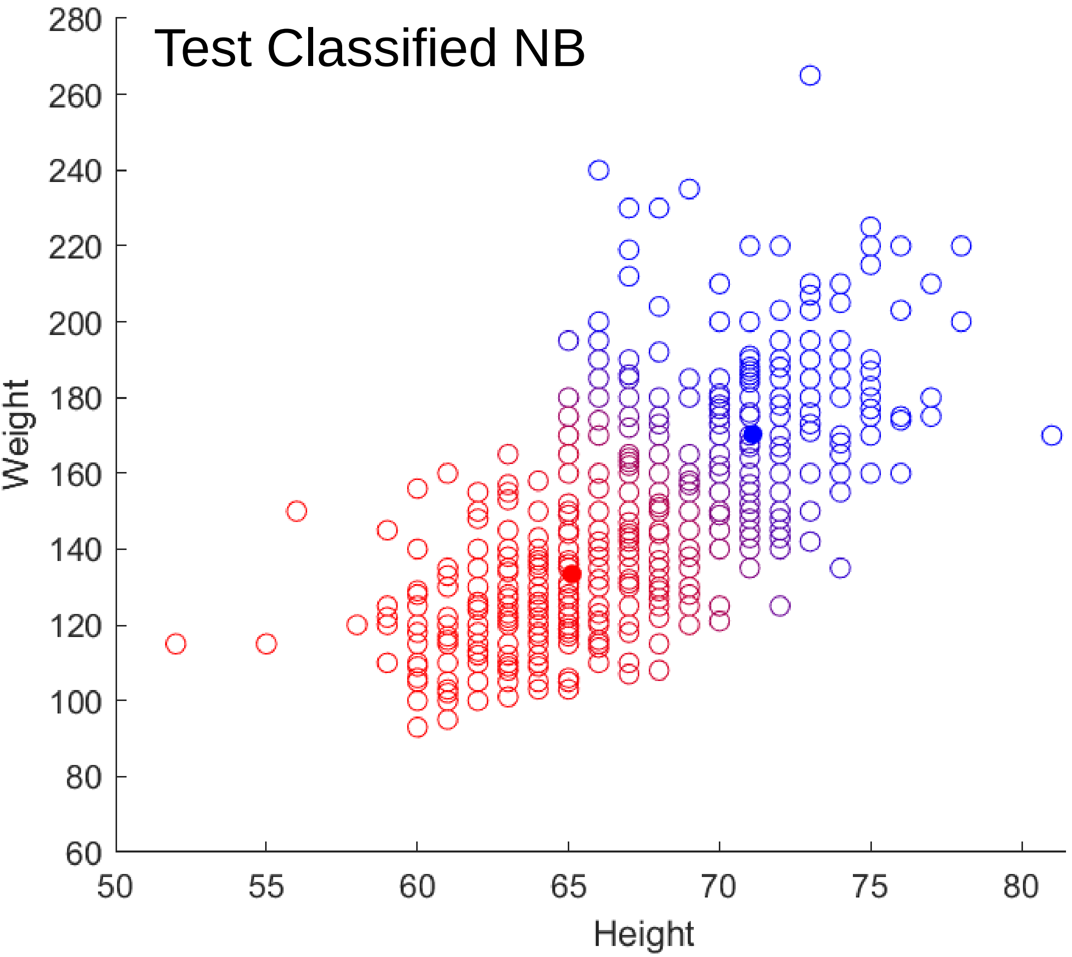


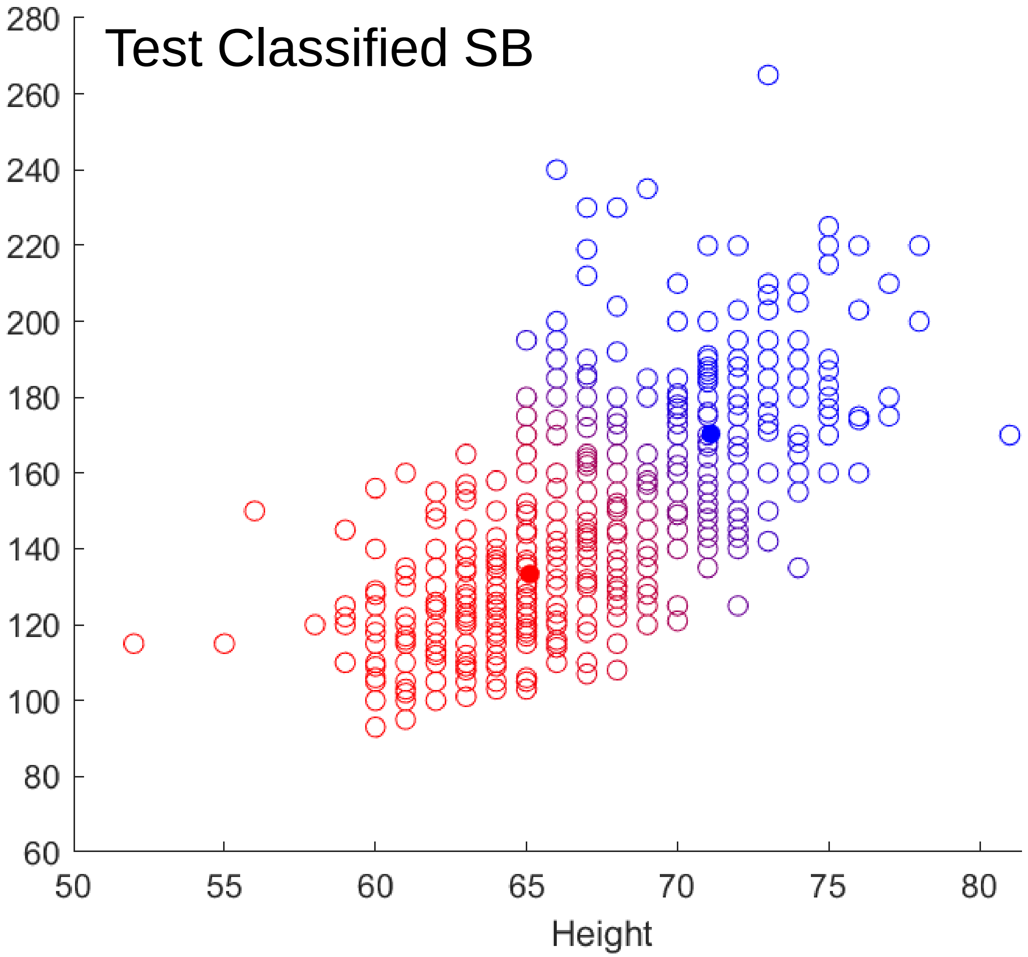


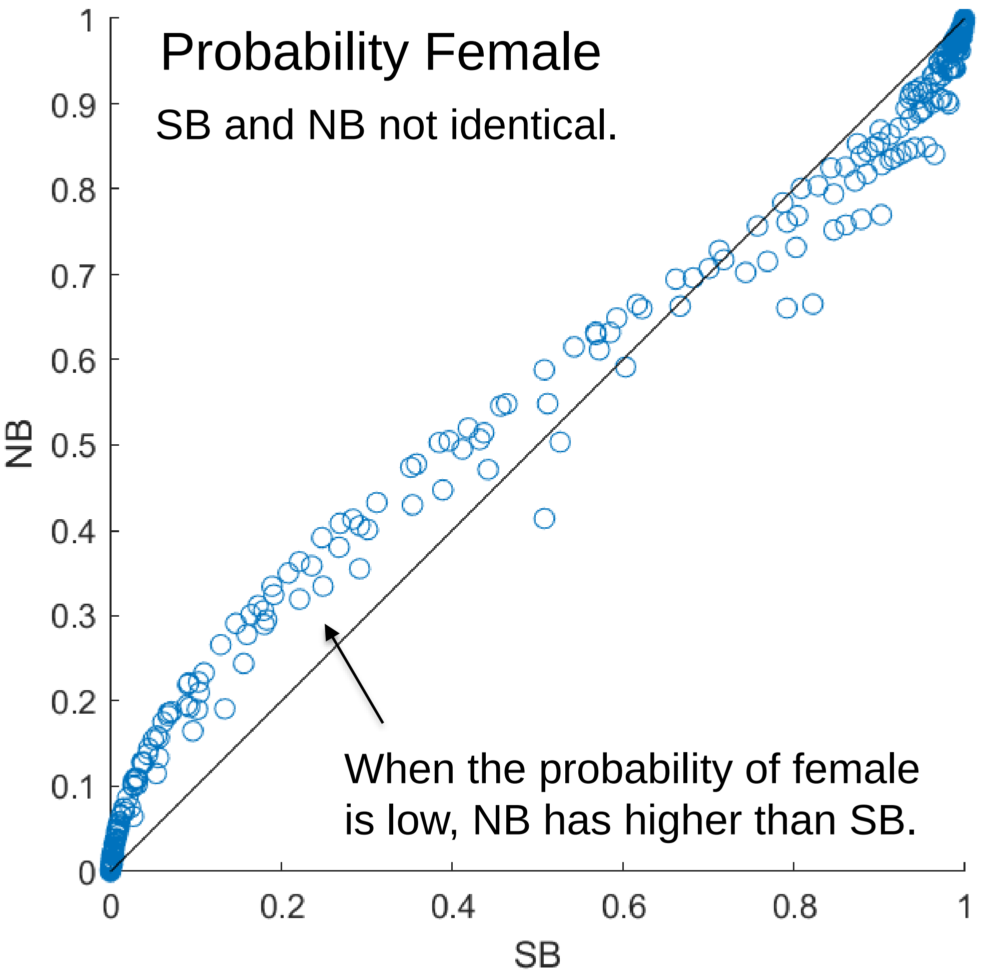
Simplified Bayes Classification

$$\hat{\mu}_1 = \begin{pmatrix} 65.1128 \\ 133.3850 \end{pmatrix} \quad \hat{\mu}_2 = \begin{pmatrix} 71.1082 \\ 170.3247 \end{pmatrix}$$

Example: Classify







Discussion

We explored the formal method for Bayesian classification that assessed a prior distribution on the mean vector and covariances matrix for each class along with unconditional class probabilities. The observed vector x could then be classified *a posteriori*.

We utilized the most common way, Naïve Bayes classification that Bayesian Statistics is used for item classification based upon their attributes (features).

Discussion

Questions?

Full Bayesian

$$f(y|x)=\int f(y,\mu_y,\Sigma_y|x)d\Sigma_y d\mu_y$$

Simplified Bayes

$$f(y|x,\hat{\mu}_y,\hat{\Sigma}_y)\propto f(x|y,\hat{\mu}_y,\hat{\Sigma}_y)f(y)$$

Naïve Bayes

$$f(y|x,\hat{\mu}_y,\hat{\sigma}_{y1}^2,\hat{\sigma}_{y2}^2)\propto f(x|y,\hat{\mu}_y,\hat{\sigma}_{y1}^2,\hat{\sigma}_{y2}^2)f(y)$$

Homework 13

1. Select true expected values, variances and covariances for heights and weights for each class of females and males. i.e. correlated observations for training.
   Generate some number m_1 and m_2 from each.
   Estimate sample means, variances, and covariances.
   Generate new correlated observations for testing
   n_1 and n_2 from each female and male.

   Separately classify the new observations to male/female using both Simplified Bayesian and Naïve Bayes classification. Comment!

Homework 13

2***.Repeat problem 1 but now assess conjugate prior distributions for the parameters. Go through the described Bayesian Process. Comment!

*** For enthusiastic students.