Classification: Introduction

David Dalpiaz

STAT 430, Fall 2017

Announcements

- BVT Review
- Regression Questions?
- General Questions?

- Supervised Learning
  - Regression
  - **Classification**
- Unsupervised Learning

Classification versus Regression

Regression

$$Y \in \mathbb{R}$$

Classification

$$Y \in \mathcal{G} = \{1, 2, 3, \ldots g\}$$

American Bulldog

Figure 1: A Good Dog

American Bulldog

Figure 2: A Good Dog

American Bulldog

Figure 3: A Good Dog

American Bulldog

Figure 4: A Good Dog

Golden Retriever

Figure 5: A Good Dog

Golden Retriever

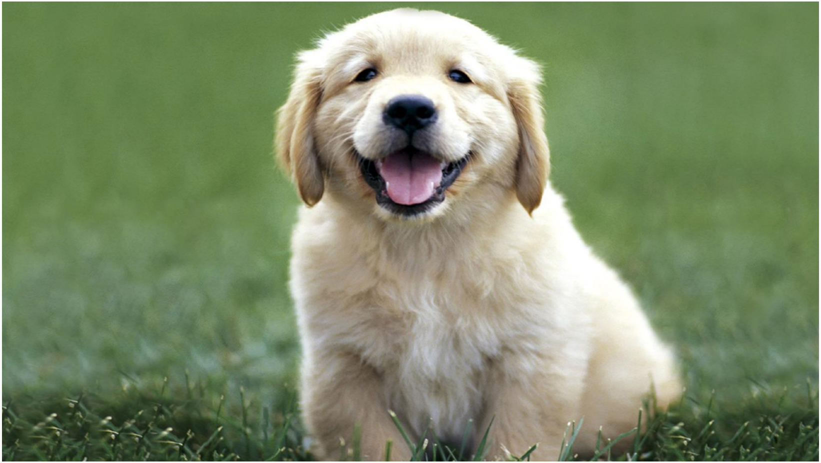


Figure 6: A Good Dog

Golden Retriever

Figure 7: A Good Dog

Golden Retriever

Figure 8: Some Good Dogs

Dog Probabilities

Let's make some assumptions about dogs. Let

- Y be the species, either B for American **B**ulldog, or R for Golden **R**etriever
- W be the weight of the dog, in kg
- H be the height of the dog, in cm

$$0.5 = \pi_B = P(Y = B) = P(Y = G) = \pi_R = 0.5$$

Dog Probabilities

American **B**ulldog

$$H \mid B \sim N(\mu = 56, \sigma^2 = 1.5^2)$$

$$W \mid B \sim N(\mu = 34, \sigma^2 = 2.25^2)$$

Golden **R**etriever

$$H \mid R \sim N(\mu = 53, \sigma^2 = 1.5^2)$$

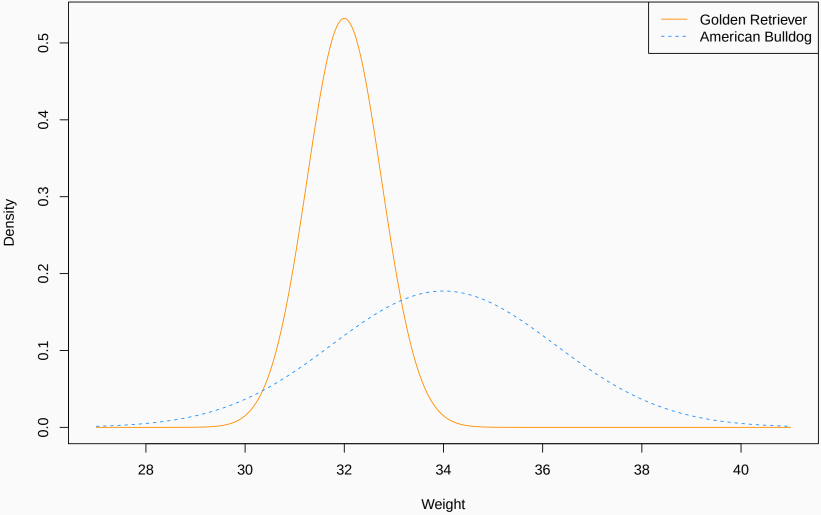
$$W \mid R \sim N(\mu = 30, \sigma^2 = 0.75^2)$$

Let's also assume that W and H are conditionally independent given B or R. (However, this is unrealistic.)

Dog Parameters

```r
# ht, wt
# cm, kg

b_mu = c(56, 34)
b_sigma = matrix(c(1.5 ^ 2, 0, 0, 2.25 ^ 2), 2, 2)

r_mu = c(53, 32)
r_sigma = matrix(c(1.5 ^ 2, 0, 0, 0.75 ^ 2), 2, 2)
```

Weight Distribution



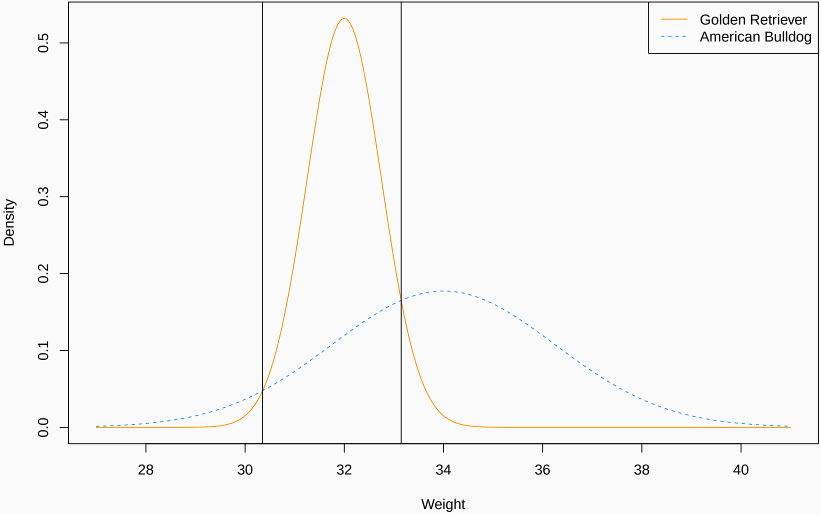
Bayes Classifier

$$C^B(x) = \underset{k}{\mathrm{argmax}} \;\; P(\ Y = k \mid X = x\ )$$

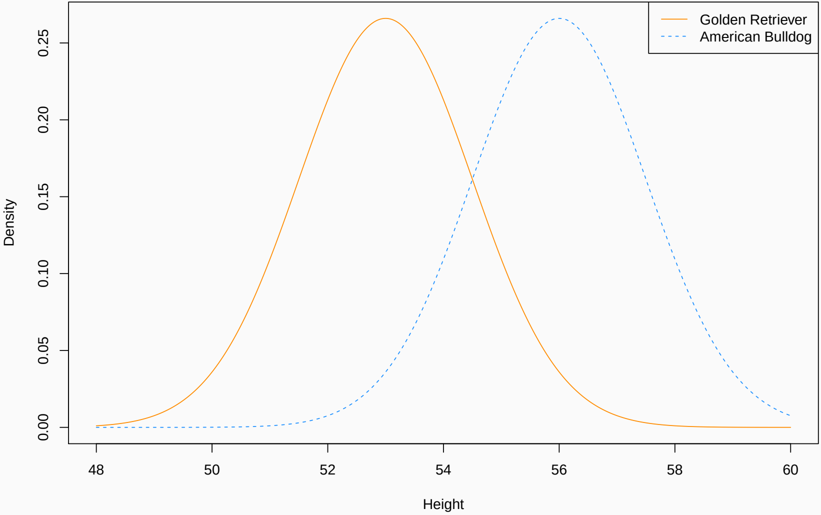
Decision Boundary

$$x :\;\; P(\ Y = B \mid X = x\ ) = P(\ Y = R \mid X = x\ )$$

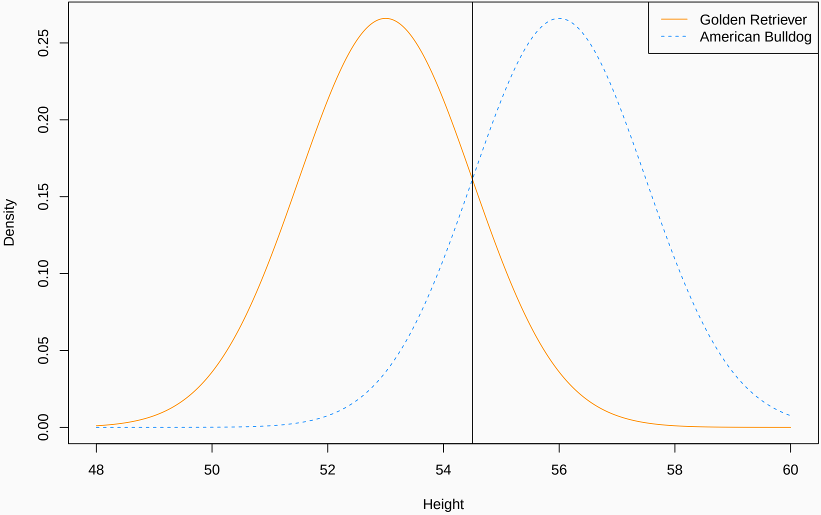
Weight Bayes, Decision Boundary



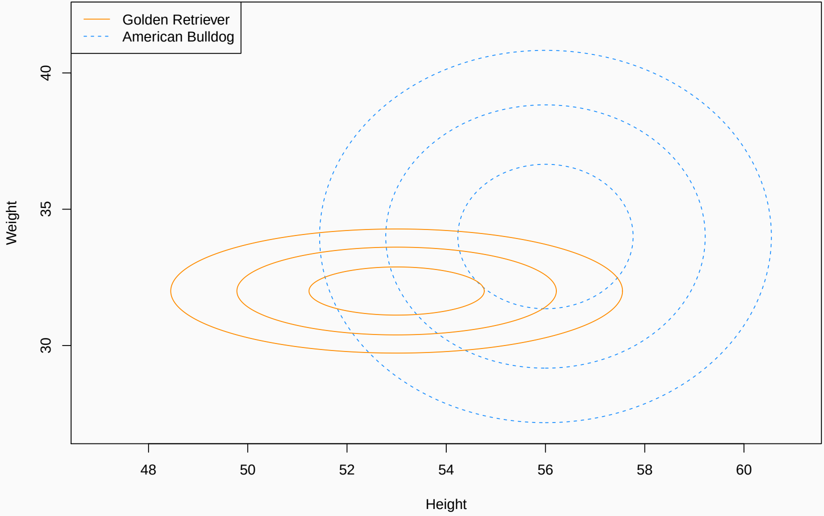
Height Distribution



Height Distribution, Decision Boundary



Height and Weight Distribution



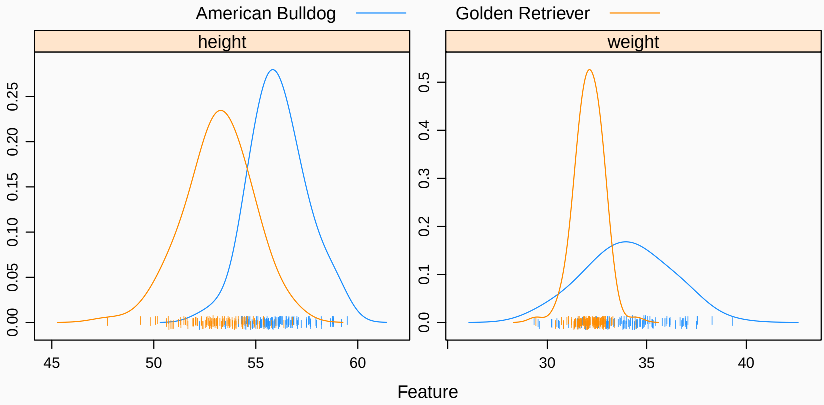
Let's Make Some Dogs

```
sim_dog_data = function(n_obs = 200,
                        b_mu, b_sigma, r_mu, r_sigma) {

  species = c(rep("American Bulldog", n_obs / 2),
              rep("Golden Retriever", n_obs / 2))
  ht_wt = rbind(mvtnorm::rmvnorm(n = n_obs / 2,
                                 mean = b_mu, sigma = b_sigma),
                mvtnorm::rmvnorm(n = n_obs / 2,
                                 mean = r_mu, sigma = r_sigma))

  data.frame(
    species,
    height = ht_wt[, 1],
    weight = ht_wt[, 2])
}
```

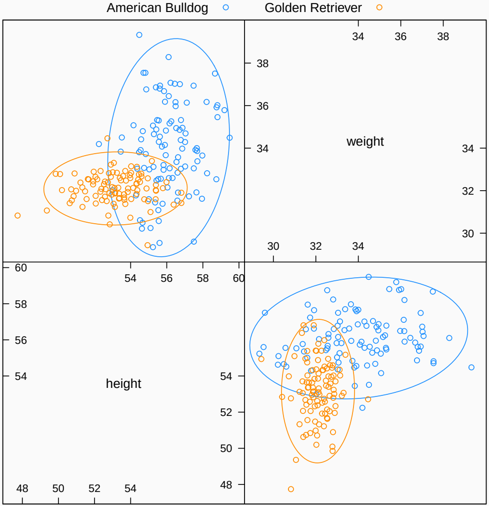
Let's Make Some Dogs

```
set.seed(66)
dog_trn = sim_dog_data(n_obs = 200,
                       b_mu, b_sigma, r_mu, r_sigma)
dog_tst = sim_dog_data(n_obs = 800,
                       b_mu, b_sigma, r_mu, r_sigma)
```

Simulated Dogs, Univariate Density Estimates



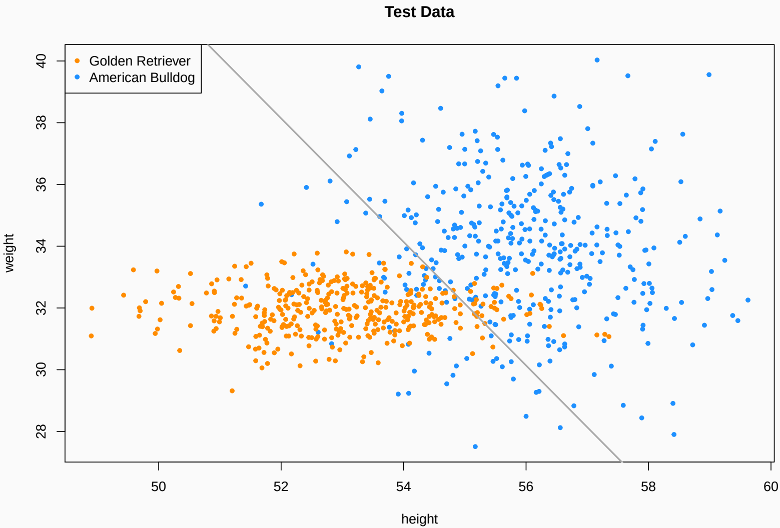
Simulated Dogs, Bivariate Density Estimates



Scatter Plot Matrix

Simulated Train Dogs, Decision?

Simulated Test Dogs, Decision?



Classification Error

$$I(y_i \neq \hat{C}(x)) = \begin{cases} 1 & y_i \neq \hat{C}(x) \\ 0 & y_i = \hat{C}(x) \end{cases}$$

$$\mathrm{Err}(\hat{C}, \mathrm{Data}) = \frac{1}{n} \sum_{i=1}^{n} I(y_i \neq \hat{C}(x_i))$$

Decision Boundary ?

$$w = 142.1745763 - 2.0006502 \cdot h$$

```
C = function(data) {
  with(data,
       ifelse(weight > (142.1746 - 2.00065 * height),
              "American Bulldog",
              "Golden Retriever"))
}
```

```
# train error
mean(C(dog_trn) != dog_trn$species)
```

```
## [1] 0.125
```

```
# test error
mean(C(dog_tst) != dog_tst$species)
```

```
## [1] 0.11
```