

# DSC 140A

*Probabilistic Modeling & Machine Learning*

Lecture 11 | Part 1

**Probabilistic Modeling**

# Probabilistic Modeling

- ▶ Where does data come from? *How data are generated:  
→ measurements + their distributions*
- ▶ We might imagine that “Nature” generates it using some random (i.e., **probabilistic**) process.
- ▶ Maybe modeling this probabilistic process will suggest new ways of making predictions?

# Example: Flowers

- ▶ Suppose there are two species of flower.
- ▶ One species tends to have more petals.
- ▶ **Goal:** Given a new flower with  $X = x$  petals predict its species,  $Y$ .



# Example: Flowers

- ▶ **Idea:** The number of petals,  $X$ , and the species,  $Y$ , are **random variables**.
- ▶ **Assumption:** When Nature generates a new flower, it picks  $X$  and  $Y$  from some **probability distribution**.
- ▶ Let's imagine (for now) that we know this distribution.

# The Joint Distribution

- ▶ The **joint distribution**  $\mathbb{P}(X = x, Y = y)$  gives us full information.<sup>1</sup>

	Y = 0	Y = 1
X = 0	0%	0%
X = 1	5%	0%
X = 2	10%	5%
X = 3	15%	15%
X = 4	5%	20%
X = 5	0%	15%
X = 6	0%	10%

<sup>1</sup>Note: it's conventional to encode Y as 0 or 1 instead of 1 and -1.

# Observation

- ▶ The entries of the joint distribution table sum to 100%.
- ▶ Mathematically:  $\sum_{x \in \{0,1,\dots,6\}} \sum_{y \in \{0,1\}} \mathbb{P}(X = x, Y = y) = 1.$

	Y = 0	Y = 1
X = 0	0%	0%
X = 1	5%	0%
X = 2	10%	5%
X = 3	15%	15%
X = 4	5%	20%
X = 5	0%	15%
X = 6	0%	10%

# Marginal Distributions

- ▶ What is the probability that a new flower has  $X = 4$  petals (regardless of species)?

	$Y = 0$	$Y = 1$
$X = 0$	0%	0%
$X = 1$	5%	0%
$X = 2$	10%	5%
$X = 3$	15%	15%
$X = 4$	5%	20%
$X = 5$	0%	15%
$X = 6$	0%	10%

$$5\% + 20\% = 25\%$$

# Marginal Distributions

- ▶ The **marginal distribution** for  $X$  is found by summing over values of  $Y$ .
- ▶ That is:  $\mathbb{P}(X = x) = \sum_{y \in \{0,1\}} \mathbb{P}(X = x, Y = y)$

	$Y = 0$	$Y = 1$	
$X = 0$	0%	0%	0
$X = 1$	5%	0%	5
$X = 2$	10%	5%	15
$X = 3$	15%	15%	30
$X = 4$	5%	20%	25
$X = 5$	0%	15%	15
$X = 6$	0%	10%	10

joint

$X = 0$	0%
$X = 1$	5%
$X = 2$	15%
$X = 3$	30%
$X = 4$	25%
$X = 5$	15%
$X = 6$	10%

marginal in  $X$

# Marginal Distributions

- ▶ What is the probability that a new flower is species 1 (regardless of number of petals)?

	Y = 0	Y = 1
X = 0	0%	0%
X = 1	5%	0%
X = 2	10%	5%
X = 3	15%	15%
X = 4	5%	20%
X = 5	0%	15%
X = 6	0%	10%

# Marginal Distributions

- ▶ The **marginal distribution** for  $Y$  is found by summing over values of  $X$ .
- ▶ That is:  $\mathbb{P}(Y = y) = \sum_{x \in \{0, \dots, 6\}} \mathbb{P}(X = x, Y = y)$

	$Y = 0$	$Y = 1$
$X = 0$	0%	0%
$X = 1$	5%	0%
$X = 2$	10%	5%
$X = 3$	15%	15%
$X = 4$	5%	20%
$X = 5$	0%	15%
$X = 6$	0%	10%

joint

$Y = 0$	35%
$Y = 1$	65%

marginal in  $Y$

# Observation

- ▶ The probabilities in the marginal distributions also sum to 1.

## Exercise

Suppose flower A has 4 petals. What do you think its species is?

	Y = 0	Y = 1
X = 0	0%	0%
X = 1	5%	0%
X = 2	10%	5%
X = 3	15%	15%
X = 4	5%	20%
X = 5	0%	15%
X = 6	0%	10%

# Intuition

- ▶ It seems **more likely** that a petal with 4 flowers is from species 1.

	Y = 0	Y = 1
X = 0	0%	0%
X = 1	5%	0%
X = 2	10%	5%
X = 3	15%	15%
X = 4	5%	20%
X = 5	0%	15%
X = 6	0%	10%

# Conditional Probabilities

- This is captured by the **conditional probability**

$$\mathbb{P}(Y = y | X = x) = \mathbb{P}(X = x, Y = y) / \mathbb{P}(X = x).$$

*conditioned variable*

	Y = 0	Y = 1
X = 0	0%	0%
X = 1	5%	0%
X = 2	10%	5%
X = 3	15%	15%
X = 4	5%	20%
X = 5	0%	15%
X = 6	0%	10%

joint

$$\mathbb{P}(Y = y | X = 1)$$

Y = 0	100%
Y = 1	0%

$$\mathbb{P}(Y = y | X = 2)$$

Y = 0	66.5%
Y = 1	33.3%

$$\mathbb{P}(Y = y | X = 4)$$

Y = 0	20%
Y = 1	80%

# Conditional Probabilities

- The **conditional probability**

$$\mathbb{P}(X = x | Y = y) = \mathbb{P}(X = x, Y = y) / \mathbb{P}(Y = y).$$

	Y = 0	Y = 1
X = 0	0%	0%
X = 1	5%	0%
X = 2	10%	5%
X = 3	15%	15%
X = 4	5%	20%
X = 5	0%	15%
X = 6	0%	10%

joint

$\mathbb{P}(X = x   Y = 0)$	
X = 0	0%
X = 1	14.2%
X = 2	28.5%
X = 3	42.8%
X = 4	14.2%
X = 5	0%
X = 6	0%

# Observation

- ▶ Conditional probabilities sum to 1 as well.
- ▶ For any fixed  $x$ :

$$\sum_y \mathbb{P}(Y = y | X = x) = 1$$

- ▶ For any fixed  $y$ :

$$\sum_x \mathbb{P}(X = x | Y = y) = 1$$

# Five Distributions

- ▶ We've seen five distributions:
  - ▶ **Joint:**  $\mathbb{P}(X = x, Y = y)$
  - ▶ **Marginal in X:**  $\mathbb{P}(X = x)$
  - ▶ **Marginal in Y:**  $\mathbb{P}(Y = y)$
  - ▶ **Conditional on X:**  $\mathbb{P}(Y = y \mid X = x)$
  - ▶ **Conditional on Y:**  $\mathbb{P}(X = x \mid Y = y)$
  
- ▶ If we know the **joint** distribution, we can compute any of the others.

# Bayes' Theorem li'kol'ho'd

- **Bayes' Theorem** relates conditional probabilities and provides another way of computing them:

$$\left\{ \begin{array}{l} \text{posterior} \\ \mathbb{P}(Y = y | X = x) = \frac{\mathbb{P}(X = x | Y = y) \mathbb{P}(Y = y)}{\mathbb{P}(X = x)} \\ \mathbb{P}(X = x | Y = y) = \frac{\mathbb{P}(Y = y | X = x) \mathbb{P}(X = x)}{\mathbb{P}(Y = y)} \end{array} \right. \text{prior}$$

# Bayes' Theorem

Recall  $P(X=x|Y=y) = \frac{P(X=x, Y=y)}{P(Y=y)}$

► Derivation:

$$\begin{aligned} P(X=x, Y=y) &= P(X=x|Y=y) P(Y=y) \\ &= P(Y=y|X=x) P(X=x) \end{aligned}$$

$$\Rightarrow P(Y=y|X=x) P(X=x) = P(X=x|Y=y) P(Y=y)$$

$$\Rightarrow P(Y=y|X=x) = \frac{P(X=x|Y=y) P(Y=y)}{P(X=x)} \quad \# \text{ Bayes' Theorem}$$

# DSC 140A

*Probabilistic Modeling & Machine Learning*

Lecture 11 | Part 2

## Bayes Decision Theory

*Probabilistic classification  
(soft classification)*

# Probabilistic Classification

- ▶ We have seen several **paradigms** for classification:
  - ▶ Nearest neighbors
  - ▶ Linear predictors
  
- ▶ Now we see a **probabilistic** paradigm.

# Probabilistic Classification

- ▶ In ML, we think of  $X$  as a feature and  $Y$  as a label.
  - ▶ **Example:** Given new flower with  $X = x$  petals, predict the species,  $Y$ .
- ▶ **Idea:** predict the **most likely** label  $y$  given  $X = x$ .
  - ▶ I.e., predict the  $y$  that maximizes  $\mathbb{P}(Y = y | X = x)$ .

# Binary Probabilistic Classification

- ▶ Predict 1 if  $\mathbb{P}(Y = 1 | X = x) > \mathbb{P}(Y = 0 | X = x)$ ; otherwise predict 0.
- ▶ That is, pick whichever label is more likely given the features.

# Bayes Classification Rule

- ▶ This is the **Bayes (binary) classification rule**:
  - ▶ Predict class 1 if  $\mathbb{P}(Y = 1 | X = x) > \mathbb{P}(Y = 0 | X = x)$ ;
  - ▶ Otherwise, predict class 0.

# Bayes Decision Theory

- ▶ Two equivalent forms of the Bayes classification rule:

- ▶ **“Original” form:**

- ▶ Predict class 1 if  $\mathbb{P}(Y = 1 | X = x) > \mathbb{P}(Y = 0 | X = x)$ ;
- ▶ Otherwise, predict class 0.

- ▶ **“Alternative” form:**

- ▶ Using Bayes' rule,

$$\mathbb{P}(Y = y | X = x) = \mathbb{P}(X = x | Y = y)\mathbb{P}(Y = y) / \mathbb{P}(X = x) \dots$$

- ▶ Predict class 1 if

$$\mathbb{P}(X = x | Y = 1)\mathbb{P}(Y = 1) > \mathbb{P}(X = x | Y = 0)\mathbb{P}(Y = 0)$$

- ▶ Otherwise, predict class 0.

## Main Idea

If we know the conditional probability of the label  $Y$  given feature  $X$ , the Bayes classification rule is a natural way to make predictions.

# Optimality

- ▶ We'll see that the Bayes classification rule is **optimal** in a certain sense.

# DSC 140A

*Probabilistic Modeling & Machine Learning*

Lecture 11 | Part 3

**Continuous Distributions**

## Example: Penguins

- ▶ Suppose there are two species of penguin.
- ▶ One species tends to have longer flippers.
- ▶ **Goal:** given a new penguin with flipper length  $X = x$ , predict its species,  $Y$ .

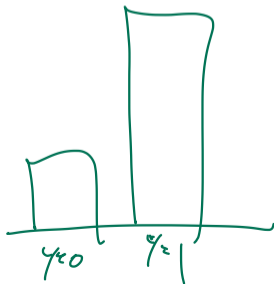
# Five Distributions

- ▶ In this situation, what do the five distributions look like?
  - ▶ Joint distribution of  $X$  and  $Y$
  - ▶ Marginal distribution in  $X$
  - ▶ Marginal distribution in  $Y$
  - ▶ Conditional on  $X$
  - ▶ Conditional on  $Y$

# Marginal in $Y$

- ▶ What is the probability that Nature generates a penguin from species  $Y$ ?
  - ▶ Marginal distribution:  $\mathbb{P}(Y = y)$ .
- ▶ This is a **discrete** distribution, as before.
- ▶ Example:

$Y = 0$	30%
$Y = 1$	70%

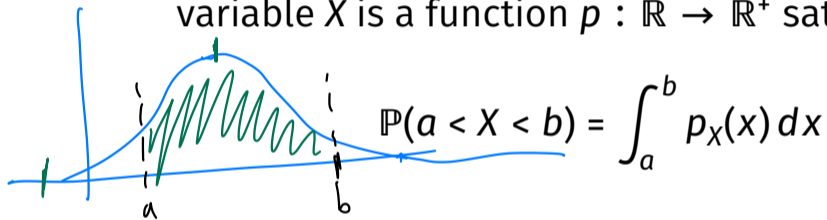


# Marginal in $X$

- ▶ What is the probability that Nature generates a flipper length of  $x$ , without regard to species?
- ▶ Flipper length is a **continuous** random variable.
- ▶ Distribution is described by a **probability density function (pdf)**,  $p : \mathbb{R} \rightarrow \mathbb{R}^+$ .

# Recall: Density Functions

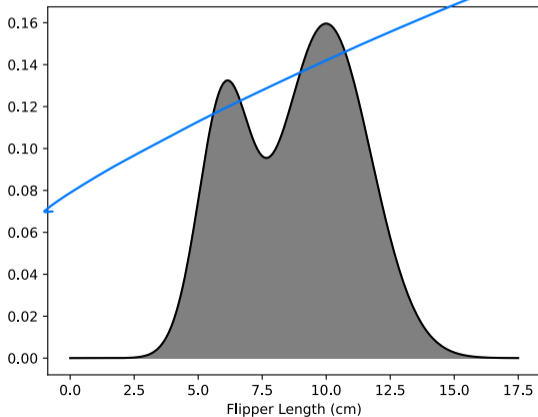
- ▶ A **probability density function (pdf)** for a random variable  $X$  is a function  $p : \mathbb{R} \rightarrow \mathbb{R}^+$  satisfying:



- ▶ That is, the pdf  $p$  describes how likely it is to get a value of  $X$  in any interval  $[a, b]$ .
- ▶ Note:  $\int_{-\infty}^{\infty} p_X(x) dx = 1$ , but  $p(x)$  can be larger than one.

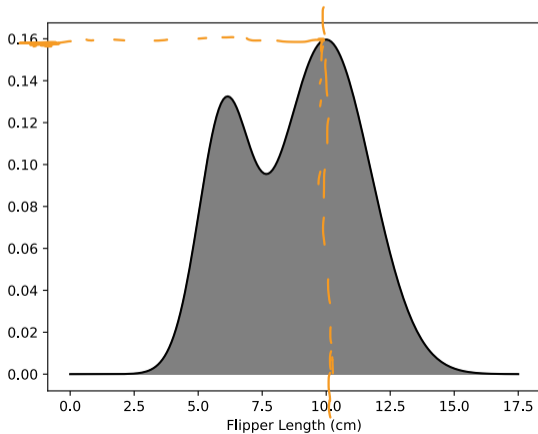
# Marginal in $X$

- ▶ The distribution of flipper lengths is described by a density function,  $p_X(x)$ .



## Exercise

What is the probability that Nature generates a penguin with flipper length equal to 10 cm?

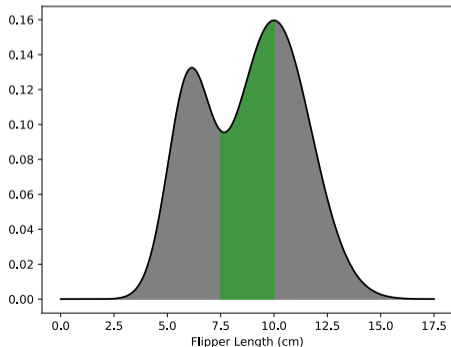


# Solution

- ▶ **Zero!**
- ▶  $p_X(x)$  is **not** the probability that  $X = x$ .
- ▶ Instead,  $\mathbb{P}(X = x) = \mathbb{P}(x < X < x) = \int_x^x p_X(x) dx = 0$
- ▶ The **probability** of a continuous random variable being *exactly* a certain value is zero.

# Example

- ▶ What is the probability that Nature generates a penguin whose flipper length is between 7.5 and 10 cm?

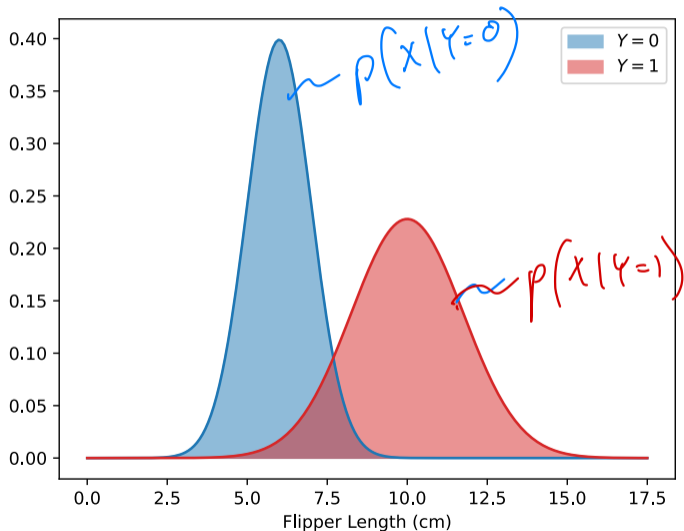


$$\mathbb{P}(7.5 < X < 10) = \int_{7.5}^{10} p_X(x) dx$$

# Conditional on $Y$

- ▶ What is the probability of a certain flipper length, given that the species is  $y$ ?
- ▶ Also a continuous distribution, described by **conditional density**  $p(x | Y = y)$ . "likelihood"
- ▶ Two conditional density functions: one for  $Y = 0$  and one for  $Y = 1$ .
  - ▶ Each integrates to one.

# Conditional on $Y$



# Conditional on X

$$P(Y | X=x)$$

- ▶ What is the probability that the species is  $y$  given a flipper length of  $x$ ?
- ▶ The conditional distribution of  $Y$  given  $X$ .

## Exercise

Is this distribution **continuous** or **discrete**?

## Conditional on $X$

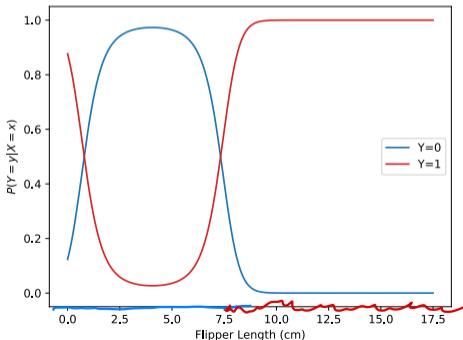
- ▶ Answer: **discrete**, because  $Y$  is discrete.
- ▶ One distribution  $\mathbb{P}(Y = y \mid X = x)$  for each possible value of  $X$  (infinitely many).

# Conditional on X

- ▶ Although for any fixed  $x$ ,  $\mathbb{P}(Y = y | X = x)$  is discrete, we can plot the functions

$$f_0(x) = \mathbb{P}(Y = 0 | X = x) \text{ and } f_1(x) = \mathbb{P}(Y = 1 | X = x)$$

likelyhood



# Bayes' Rule

- ▶ Bayes' Rule applies to densities, too:

$$\mathbb{P}(Y = y | X = x) = \frac{p(x | Y = y)\mathbb{P}(Y = y)}{p_X(x)}$$

# Bayes Decision Theory

- ▶ The Bayes classification rule applies to continuous distributions, too.

- ▶ **“Original” form:**

- ▶ Predict class 1 if  $\mathbb{P}(Y = 1 | X = x) > \mathbb{P}(Y = 0 | X = x)$ ;
- ▶ Otherwise, predict class 0.

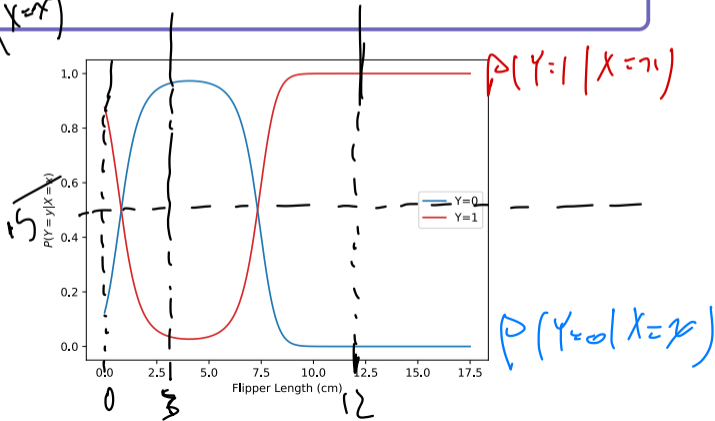
- ▶ **“Alternative” form:**

- ▶ Using Bayes' rule,  
 $\mathbb{P}(Y = y | X = x) = \mathbb{P}(X = x | Y = y)\mathbb{P}(Y = y) / \mathbb{P}(X = x) \dots$
- ▶ Predict class 1 if  
 $\mathbb{P}(X = x | Y = 1)\mathbb{P}(Y = 1) > \mathbb{P}(X = x | Y = 0)\mathbb{P}(Y = 0)$
- ▶ Otherwise, predict class 0.

## Exercise

Penguins with flippers of length 0, 3, and 12 are observed. What are their predicted species according to the Bayes' classification rule?

$$P(Y=1|X=0) > P(Y=0|X=0)$$



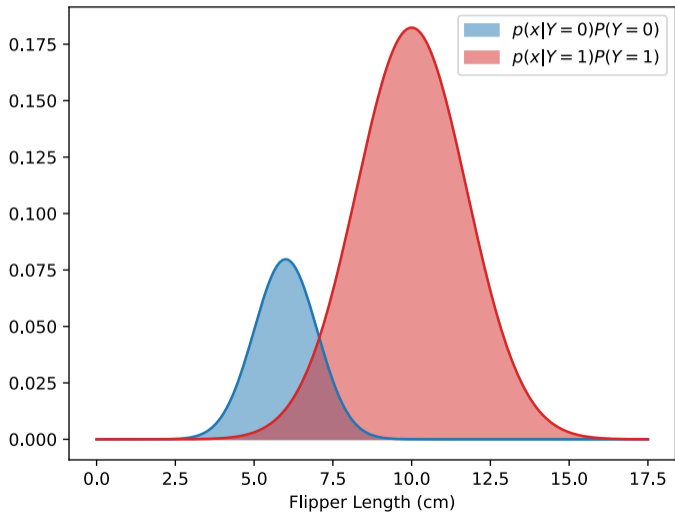
# Joint

- ▶ The **joint** distribution in this case is neither totally continuous nor totally discrete.
- ▶ From Bayes' rule:

$$p(x, 0) = p(x | Y = 0)\mathbb{P}(Y = 0)$$

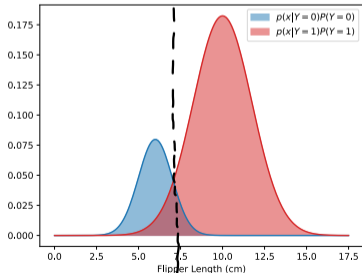
$$p(x, 1) = p(x | Y = 1)\mathbb{P}(Y = 1)$$

# Joint Distribution



## Exercise

Where does the Bayes decision rule make a prediction for class 1?



- ▶ Predict class 1 if  $p(x | Y = 1)P(Y = 1) > p(X = x | Y = 0)P(Y = 0)$
- ▶ Otherwise, predict class 0.

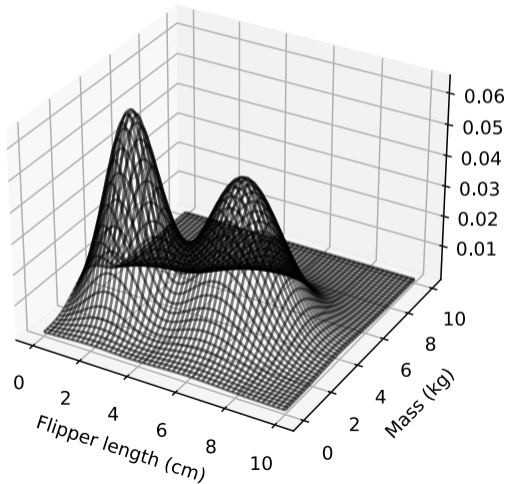
# Multivariate Distributions

- ▶ In binary classification,  $y \in \{0, 1\}$ .
- ▶ But we usually deal with feature vectors,  $\vec{x}$ .
- ▶ The previous applies with straightforward changes.

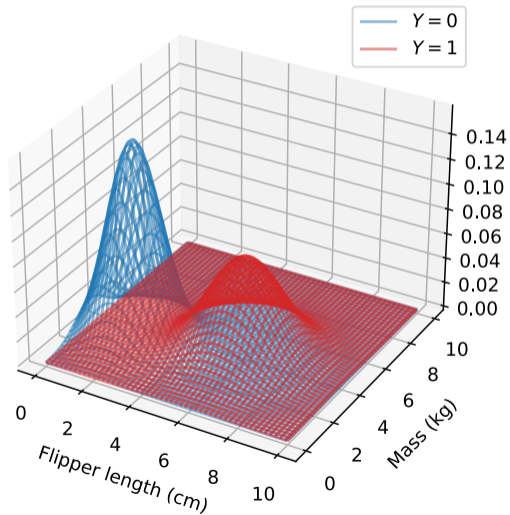
# Example: Penguins

- ▶ Again consider penguins of two species, but now consider both flipper length and body mass.
- ▶ Each penguin's measurements are a **random vector**:  $\vec{X}$ .
- ▶ Densities are now functions of a vector.
  - ▶ E.g., marginal:  $p_X(\vec{x}) : \mathbb{R}^2 \rightarrow \mathbb{R}^+$

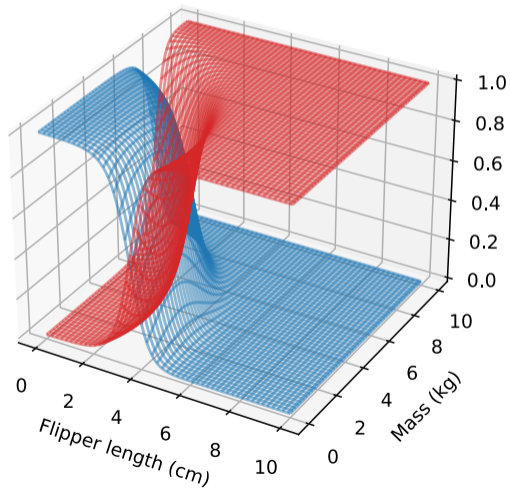
# Marginal in $\vec{X}$



# Conditional on $Y$



# Conditional on $X$



# DSC 140A

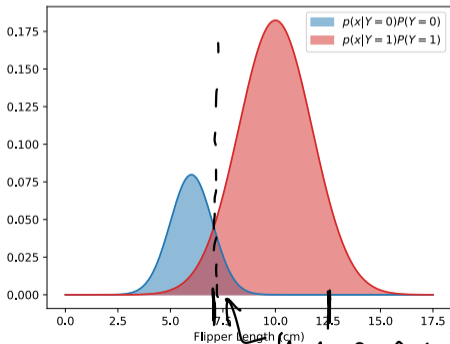
*Probabilistic Modeling & Machine Learning*

Lecture 11 | Part 4

**Bayes Error**

## Exercise

Can a classifier exist that predicts the species of an unseen penguin with 0% chance of error?



Non-separable

# Bayes Error

- ▶ The **Bayes classification rule** is a natural approach to making predictions.
- ▶ It is also the **best you can do**, in a sense.



# Errors

- ▶ What is the probability that the Bayes classifier makes a mistake?
- ▶ In binary classification, there are two types of errors:
  - ▶ Predicted 0, but the right answer is 1 (Case 1).
  - ▶ Predicted 1, but the right answer is 0 (Case 2).
- ▶ The probability of making a classification error is:

$$\mathbb{P}(\text{error}) = \mathbb{P}(\text{Case 1}) + \mathbb{P}(\text{Case 2})$$

# Probability of Error

- ▶ Case 1: it predicts 0, but the right answer is 1.
- ▶ The probability of this error is:

$$\begin{aligned}\mathbb{P}(\text{Case 1}) &= \mathbb{P}(\underbrace{Y \text{ is actually } 1}, \underbrace{\text{predict } 0}) \\ &= \underbrace{\mathbb{P}(Y \text{ is actually } 1)} \times \underbrace{\mathbb{P}(\text{predict } 0 \mid Y \text{ is actually } 1)}\end{aligned}$$

# Probability of Error

- ▶ Case 2: it predicts 1, but the right answer is 0.
- ▶ The probability of this error is:

$$\begin{aligned}\mathbb{P}(\text{Case 2}) &= \mathbb{P}(Y \text{ is actually } 0, \text{ predict } 1) \\ &= \mathbb{P}(Y \text{ is actually } 0) \times \mathbb{P}(\text{predict } 1 \mid Y \text{ is actually } 0)\end{aligned}$$

# Bayes Error

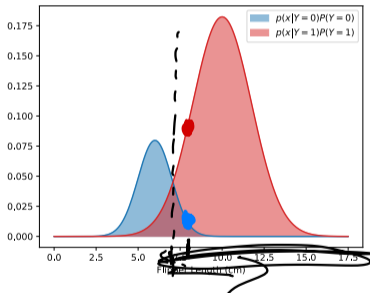
- ▶ The probability that the Bayes classifier makes a mistake is:

$$\begin{aligned}\mathbb{P}(\text{error}) &= \mathbb{P}(\text{Case 1}) + \mathbb{P}(\text{Case 2}) \\ &= \mathbb{P}(Y \text{ is actually 1}) \times \mathbb{P}(\text{predict 0} \mid Y \text{ is actually 1}) \\ &\quad + \mathbb{P}(Y \text{ is actually 0}) \times \mathbb{P}(\text{predict 1} \mid Y \text{ is actually 0})\end{aligned}$$

- ▶ This is called the **Bayes error**.
  - ▶ It depends on the distribution.

# Interpretation

- ▶ The Bayes error measures the **overlap** between the class-conditional distributions.



- ▶ It is the unavoidable error rate for any classifier on that distribution.

# Fact

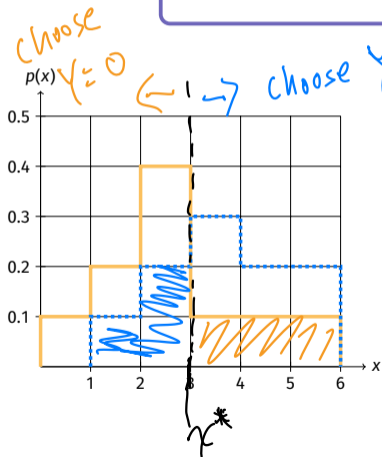
- ▶ The Bayes error is the **lowest possible** error rate for any classifier.
- ▶ Take *any* classifier. The probability that it makes an error is  $\geq$  the Bayes error.<sup>2</sup>
- ▶ In other words, the **Bayes classifier** is **optimal**.

---

<sup>2</sup>Note that this is a statement about a probability, and not about any finite data set!

## Exercise

What is the Bayes error if  $\mathbb{P}(Y = 0) = \mathbb{P}(Y = 1) = 0.5$ ?



$$\begin{aligned} & \frac{1}{2} \times \frac{3}{10} \\ & + \frac{1}{2} \times \frac{3}{10} \end{aligned}$$

$$P(\text{error}) = \frac{1}{2} \cdot \frac{3}{10} + \frac{1}{2} \cdot \frac{3}{10} = \frac{3}{10}$$

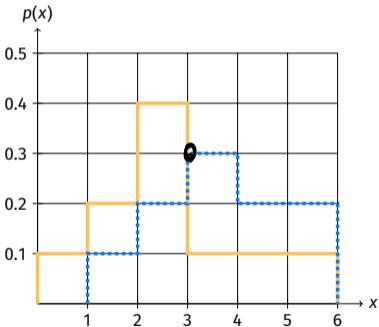
$P(Y=1|X=x) > P(Y=0|X=x)$  Boundary  $x^*$  s.t.  $P(Y=1|X=x^*) = P(Y=0|X=x^*)$

$P(Y=0|X=x^*)$

## Exercise

What is the Bayes error if  $P(Y=0)=0.3$  and  $P(Y=1)=0.7$ ?

$P(X=x^*|Y=1)P(Y=1) = P(X=x^*|Y=0)P(Y=0)$



—  $p_0(x|Y=0)$   
 - - -  $p_1(x|Y=1)$

$P(Y \text{ is actually } 1) \times P(\text{predict } 0 | Y \text{ is actually } 1)$   
 $+ P(Y \text{ is actually } 0) \times P(\text{predict } 1 | Y \text{ is actually } 0)$

# Bayes Error vs. Accuracy

- ▶ But wait... can't we usually make a classifier with 100% training accuracy?
- ▶ Yes! But training accuracy is **not the same** as the probability of making a mistake on new data.
- ▶ In fact, training accuracy is one way of estimating the true error rate.
  - ▶ Test accuracy is another (better) estimate.
- ▶ If train accuracy is much larger than the true error rate, we are **overfitting**.

# Problem

- ▶ The Bayes classifier is optimal.
- ▶ But it requires knowing the joint distribution; we **almost never know this**.
- ▶ Next time: **estimating** probability distributions from data.