

Lecture 11 | Part 1

Probabilistic Modeling

Probabilistic Modeling

- Where does data come from?
- 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.¹

	Y = 0	<i>Y</i> = 1
X = 0	0%	0%
<i>X</i> = 1	5%	0%
<i>X</i> = 2	10%	5%
<i>X</i> = 3	15%	15%
<i>X</i> = 4	5%	20%
X = 5	0%	15%
<i>X</i> = 6	0%	10%

¹Note: 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
<i>X</i> = 0	0%	0%
<i>X</i> = 1	5%	0%
<i>X</i> = 2	10%	5%
<i>X</i> = 3	15%	15%
<i>X</i> = 4	5%	20%
<i>X</i> = 5	0%	15%
<i>X</i> = 6	0%	10%

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

	Y = 0	Y = 1
X = 0	0%	0%
<i>X</i> = 1	5%	0%
<i>X</i> = 2	10%	5%
<i>X</i> = 3	15%	15%
<i>X</i> = 4	5%	20%
X = 5	0%	15%
X = 6	0%	10%

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
<i>X</i> = 0	0%	0%
<i>X</i> = 1	5%	0%
<i>X</i> = 2	10%	5%
X = 3	15%	15%
X = 4	5%	20%
X = 5	0%	15%
<i>X</i> = 6	0%	10%

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

marginal in X

joint

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

	<i>Y</i> = 0	Y = 1
<i>X</i> = 0	0%	0%
<i>X</i> = 1	5%	0%
<i>X</i> = 2	10%	5%
<i>X</i> = 3	15%	15%
<i>X</i> = 4	5%	20%
<i>X</i> = 5	0%	15%
<i>X</i> = 6	0%	10%

The marginal distribution for Y is found by summing over values of X.

• That is:
$$\mathbb{P}(Y = y) = \sum_{x \in \{0,...,6\}} \mathbb{P}(X = x, Y = y)$$

	Y = 0	Y = 1
X = 0	0%	0%
<i>X</i> = 1	5%	0%
X = 2	10%	5%
X = 3	15%	15%
X = 4	5%	20%
X = 5	0%	15%
<i>X</i> = 6	0%	10%

ioint

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

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

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

Conditional Probabilities

• This is captured by the **conditional probability** $\mathbb{P}(Y = y | X = x) = \mathbb{P}(X = x, Y = y)/\mathbb{P}(X = x).$

			$\mathbb{P}(Y = y \mid X)$
	Y = 0	Y = 1	Y = 0 10
X = 0	0%	0%	Y = 1
(= 1	5%	0%	
K = 2	10%	5%	$\mathbb{P}(Y = y \mid X)$
(=3	15%	15%	$Y = 0 \perp 6\theta$
(= 4	5%	20%	Y = 1 33
(= 5	0%	15%	<u> </u>
(= 6	0%	10%	
			$\mathbb{P}(Y = y \mid X)$
	ioint		Y = 0 2
	junit		V – 1 9

Conditional Probabilities

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

	<i>Y</i> = 0	Y = 1
X = 0	0%	0%
X = 1 X = 2	10%	0% 5%
X = 3 X = 4	15% 5%	15% 20%
X = 4 X = 5	0%	15%
<i>X</i> = 6	0%	10%

$\mathbb{P}(X = x \mid Y = 0)$	
$\begin{array}{c} X = 0 \\ X = 1 \end{array}$	0% 14.2%
X = 2 X = 3	28.5%
X = 3 X = 4 X = 5	14.2%
x = 5 X = 6	0% 0%

joint

Observation

Conditional probabilities sum to 1 as well.

For any fixed *x*:

$$\sum_{y} \mathbb{P}(Y = y \mid X = x) = 1$$

For any fixed *y*:

$$\sum_{x} \mathbb{P}(X = x \mid Y = y) = 1$$

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: $\mathbb{P}(Y = y | X = x)$
- **Conditional on** *Y*: $\mathbb{P}(X = x | Y = y)$
- If we know the **joint** distribution, we can compute any of the others.

Bayes' Theorem

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

$$\mathbb{P}(Y = y \mid X = x) = \frac{\mathbb{P}(X = x \mid Y = y)\mathbb{P}(Y = y)}{\mathbb{P}(X = x)}$$

Bayes' Theorem

Derivation:



Lecture 11 | Part 2

Bayes Decision Theory

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 P(Y = y | X = x).

Binary Probabilistic Classification

Predict 1 if P(Y = 1 | X = x) > 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,

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

Predict class 1 if

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

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.



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: P(Y = y).
- This is a discrete distribution, as before.
- Example:

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 : ℝ → ℝ⁺.

Recall: Density Functions

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

$$\mathbb{P}(a < X < b) = \int_a^b p_X(x) \, dx$$

- 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!

h

• $p_X(x)$ 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?



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

- 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).
- Two conditional density functions: one for Y = 0 and one for Y = 1.
 - Each integrates to one.



- 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?

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

Although for any fixed x, P(Y = y | X = x) is discrete, we can plot the functions f₀(x) = P(Y = 0 | X = x) and f₁(x) = P(Y = 1 | X = x)



Bayes' Rule

Bayes' Rule applies to densities, too:

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

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, $\mathbb{P}(Y = y | X = x) = \mathbb{P}(X = x | Y = y)\mathbb{P}(Y = y)/\mathbb{P}(X = x)...$
 - Predict class 1 if
 - $\mathbb{P}(X = x \mid Y = 1)\mathbb{P}(Y = 1) > \mathbb{P}(X = x \mid Y = 0)\mathbb{P}(Y = 0)$
 - 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?



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: X.
- Densities are now functions of a vector. • Eq. marginal: $p(\vec{x}) : \mathbb{P}^2 \to \mathbb{R}^+$
 - ► E.g., marginal: $p_{\chi}(\vec{x})$: $\mathbb{R}^2 \rightarrow \mathbb{R}^+$

Marginal in \vec{X}









Lecture 11 | Part 4

Bayes Error



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



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:

P(Case 1) = P(Y is actually 1, predict 0)
 = P(Y is actually 1) × P(predict 0 | Y is actually 1)

Probability of Error

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

Bayes Error

- The probability that the Bayes classifier makes a mistake is:
 - $\mathbb{P}(\text{error}) = \mathbb{P}(\text{Case 1}) + \mathbb{P}(\text{Case 2})$ = $\mathbb{P}(\text{Y is actually 1}) \times \mathbb{P}(\text{predict 0} | \text{Y is actually 1})$ + $\mathbb{P}(\text{Y is actually 0}) \times \mathbb{P}(\text{predict 1} | \text{Y is actually 0})$
- 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 ≥ the Bayes error.²

In other words, the Bayes classifier is optimal.

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

Exercise

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



 $\begin{array}{l} -p_0(x \mid Y = 0) \\ \cdots & p_1(x \mid Y = 1) \end{array} \end{array} \qquad \mathbb{P}(Y \text{ is actually 1}) \times \mathbb{P}(\text{predict 0} \mid Y \text{ is actually 1}) \\ + \mathbb{P}(Y \text{ is actually 0}) \times \mathbb{P}(\text{predict 1} \mid Y \text{ is actually 0}) \end{array}$

Exercise

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



 $\begin{array}{l} & -p_0(x \mid Y = 0) \\ & \cdots p_1(x \mid Y = 1) \end{array} \end{array}$ $\begin{array}{l} \mathbb{P}(Y \text{ is actually 1}) \times \mathbb{P}(\text{predict 0} \mid Y \text{ is actually 1}) \\ & + \mathbb{P}(Y \text{ is actually 0}) \times \mathbb{P}(\text{predict 1} \mid Y \text{ is actually 0}) \end{array}$

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.