GENERAL ASSEMBLY

INTRO TO DATA SCIENCE

LECTURE 8: PROBABILITY AND NAIVE BAYESIAN CLASSIFICATION

LAST TIME:

- LINEAR REGRESSION

QUESTIONS?

I. INTRO TO PROBABILITY

II. NAÏVE BAYESIAN CLASSIFICATION

EXERCISES:

III. IMPLEMENTING A SPAM FILTER

I. INTRO TO PROBABILITY

Q: What is a **probability***?*

Q: What is a **probability***?*

A: A number between 0 and 1 that characterizes the likelihood that some event will occur.

Q: What is a **probability***?*

A: A number between 0 and 1 that characterizes the likelihood that some event will occur.

The probability of event A *is denoted* $P(A)$*.*

Q: What is a **probability***?*

A: A number between 0 and 1 that characterizes the likelihood that some event will occur.

The probability of event A is denoted $P(A)$.

Examples

The probability of getting heads on a coin flip is .5

The probability of picking the 1 red ball in a bag of 8 balls is .125

Q: What is the set of all possible events called?

Q: What is the set of all possible events called?

A: This set is called the **sample space** Ω*. Event* A *is a member of the sample space, as is every other event.*

Q: What is the set of all possible events called?

A: This set is called the **sample space** Ω. *Event* A *is a member of the sample space, as is every other event.*

The probability of the sample space $P(\Omega)$ *is 1.*

Q: Consider two events A & B. How can we characterize the **intersection** *of these events?*

Q: Consider two events A & B. How can we characterize the **intersection** *of these events?*

A: With the **joint probability** *of A and* B*, written $P(AB)$.*

Examples

The probability of rolling a die as an ***odd****(A)* ***prime*** *(B) number is ...*

Q: Consider two events A & B. How can we characterize the **intersection** *of these events?*

A: With the **joint probability** *of A and* B, *written $P(AB)$.*

Examples

The probability of rolling a die as an ***odd****(A)* ***prime*** *(B) number is ...*

	1	2	3	4	5	6
A:	O	E	O	E	O	E
B:	N	P	P	N	P	N

Q: Consider two events A & B. How can we characterize the **intersection** *of these events?*

A: With the **joint probability** *of A and B, written $P(AB)$.*

Examples

The probability of rolling a die as an ***odd****(A)* ***prime*** *(B) number is 2/6, or .333*

	1	2	3	4	5	6
A:	O	E	O	E	O	E
B:	N	P	P	N	P	N

Q: Consider two events A & B. How can we characterize the **intersection** *of these events?*

A: With the **joint probability** *of A and B, written $P(AB)$.*

Question:

What's the probability of rolling an ***even prime number?***

	1	2	3	4	5	6
A:	O	E	O	E	O	E
B:	N	P	P	N	P	N

Q: Consider two events A & B. How can we characterize the **intersection** *of these events?*

A: With the **joint probability** *of A and B, written $P(AB)$.*

Question:

What's the probability of rolling an ***even prime number? 1/6 (.1667)***

	1	2	3	4	5	6
A:	O	E	O	E	O	E
B:	N	P	P	N	P	N

Q: Consider two events A & B. How can we characterize the **intersection** *of these events?*

A: With the **joint probability** *of A and* B*, written $P(AB)$.*

Q: Suppose event B has occurred. What quantity represents the probability of A **given** *this information about B?*

Q: Suppose event B has occurred. What quantity represents the probability of A **given** *this information about B?*

A: The intersection of A & B divided by region B.

Q: Suppose event B has occurred. What quantity represents the probability of A **given** *this information about B?*

A: The intersection of A & B divided by region B.

NOTE

This information about B transforms the sample space.

Take a moment to convince yourself of this!

Q: Suppose event B has occurred. What quantity represents the probability of A **given** *this information about B?*

A: The intersection of A & B divided by region B.

This is called the **conditional probability** *of A given B, written $P(A \mid B) = P(AB) \,/\, P(B)$.*

NOTE

This information about B transforms the sample space.

Take a moment to convince yourself of this!

Q: Suppose event B has occurred. What quantity represents the probability of A **given** *this information about B?*

A: The intersection of A & B divided by region B.

This is called the **conditional probability** *of A given B, written $P(A \mid B) = P(AB) \,/\, P(B)$.*

Notice, with this we can also write $P(AB) = P(A \mid B) * P(B)$.

NOTE

This information about B transforms the sample space.

Take a moment to convince yourself of this!

Conditional probability: $P(A \mid B) = P(AB) \,/\, P(B)$

Back to the roll problem: If someone rolls a prime number, what is the probability that they rolled an odd?

Conditional probability: $P(A \mid B) = P(AB) \,/\, P(B)$

Back to the roll problem: If someone rolls a prime number, what is the probability that they rolled an odd?

$P(AB) = P(\text{odd and prime}) = .333 \quad (1/3, \text{ or } 2/6)$

$P(B) = P(\text{prime}) = .5 \quad (1/2, \text{ or } 3/6)$

Conditional probability: $P(A \mid B) = P(AB) \,/\, P(B)$

Question: *If someone announces they rolled an even number, what is the probability that it was prime?*

$P(AB) = P(even\ and\ prime) = .166 \quad (1/6)$
$P(B) = P(even) = .5 \quad (1/2)$

$P(A \mid B) = P(prime\ given\ even) = .166 \,/\, .5 = .333\ (1/3)$

Review time. *Determine conditional probability for each!*

We have ten brown balls, 15 brown cubes, 18 green balls, and 25 green cubes in a bag. Michael takes an item out of the bag and announces…

1) It’s green. What’s the probability it’s a cube?
2) It’s brown. What’s the probability it’s a cube?
3) It’s a ball. What’s the probability it’s green?
4) It’s a cube. What’s the probability it’s a ball?

Q: What does it mean for two events to be **independent***?*

Q: What does it mean for two events to be **independent***?*

A: Information about one does not affect the probability of the other.

Q: What does it mean for two events to be **independent***?*

A: Information about one does not affect the probability of the other.

This can be written as $P(A \mid B) = P(A)$.

Q: What does it mean for two events to be **independent**?

A: Information about one does not affect the probability of the other.

This can be written as $P(A \mid B) = P(A)$.

Using the definition of the conditional probability, we can also write:

$$P(A \mid B) = P(AB) \,/\, P(B) = P(A) \rightarrow \qquad P(AB) = P(A) * P(B)$$

$$P(AB) = P(A \mid B) * P(B)$$

from last slide

$P(AB) = P(A \mid B) * P(B)$ *from last slide*

$P(BA) = P(B \mid A) * P(A)$ *by substitution*

$P(AB) = P(A \mid B) * P(B)$ *from last slide*

$P(BA) = P(B \mid A) * P(A)$ *by substitution*

But $P(AB) = P(BA)$ *since event* $AB =$ *event* BA

$P(AB) = P(A \mid B) * P(B)$ *from last slide*

$P(BA) = P(B \mid A) * P(A)$ *by substitution*

But $P(AB) = P(BA)$ *since event* AB = *event* BA

→ $P(A \mid B) * P(B) = P(B \mid A) * P(A)$ *by combining the above*

$P(AB) = P(A \mid B) * P(B)$ *from last slide*

$P(BA) = P(B \mid A) * P(A)$ *by substitution*

But $P(AB) = P(BA)$ *since event* $AB =$ *event* BA

$\rightarrow \quad P(A \mid B) * P(B) = P(B \mid A) * P(A)$ *by combining the above*

$\rightarrow \quad P(A \mid B) = P(B \mid A) * P(A) \,/\, P(B)$ *by rearranging last step*

This result is called **Bayes' theorem***. Here it is again:*

$$P(A \mid B) = P(B \mid A) * P(A) \,/\, P(B)$$

This result is called **Bayes' theorem**. *Here it is again:*

$$P(A \mid B) = P(B \mid A) * P(A) \,/\, P(B)$$

Some facts:

- This is a simple algebraic relationship using elementary definitions.

This result is called **Bayes' theorem**. *Here it is again:*

$$P(A \mid B) = P(B \mid A) * P(A) / P(B)$$

Some facts:

- This is a simple algebraic relationship using elementary definitions.

- It's interesting because it's kind of a "wormhole" between two different "interpretations" of probability.

This result is called **Bayes' theorem**. *Here it is again:*

$$P(A \mid B) = P(B \mid A) * P(A) / P(B)$$

Some facts:

- This is a simple algebraic relationship using elementary definitions.

- It's interesting because it's kind of a "wormhole" between two different "interpretations" of probability.

- It's a very powerful computational tool.

http://www.youtube.com/watch?v=Zxm4Xxvzohk

Things to consider:

Flipping a coin to see if it's heads or tails is our ***test*** *to see which coin was chosen.*

Things to consider:

*Flipping a coin to see if it's heads or tails is our **test** to see which coin was chosen.*

*Our probability can change **dependent** on previous results. Two heads in a row did not confirm anything, but only changed our perception of probability for each coin.*

Things to consider:

*Flipping a coin to see if it's heads or tails is our **test** to see which coin was chosen.*

*Our probability can change **dependent** on previous results. Two heads in a row did not confirm anything, but only changed our perception of probability for each coin.*

*In more common Bayes Probability, tests will more commonly produce **false positives** and **false negatives**, where error comes into play.*

Briefly, the two interpretations can be described as follows:

Briefly, the two interpretations can be described as follows:

The frequentist interpretation regards an event's probability as its limiting frequency across a very large number of trials.

Briefly, the two interpretations can be described as follows:

The frequentist interpretation regards an event's probability as its limiting frequency across a very large number of trials.

The Bayesian interpretation regards an event's probability as a "degree of belief," which can apply even to events that have not yet occurred.

If this sounds crazy to you, don't worry…we won't dwell on the theoretical details.

If this sounds crazy to you, don't worry…we won't dwell on the theoretical details.

If this sounds interesting, there are plenty of resources available to learn more about Bayesian inference.

If this sounds crazy to you, don't worry…we won't dwell on the theoretical details.

If this sounds interesting, there are plenty of resources available to learn more about Bayesian inference.

This a good direction to head if you like math and/or if you're interested in learning about cutting-edge data science techniques.

The Monty Hall Problem

https://www.youtube.com/watch?v=mhlc7peGlGg

II. NAÏVE BAYESIAN CLASSIFICATION

Suppose we have a dataset with features $x_1, \ldots, x_n$ and a class label C. What can we say about classification using Bayes' theorem?

Suppose we have a dataset with features $x_1, \ldots, x_n$ and a class label C. What can we say about classification using Bayes' theorem?

$$P(\text{class } C \mid \{x_i\}) = \frac{P(\{x_i\} \mid \text{class } C) \cdot P(\text{class } C)}{P(\{x_i\})}$$

Bayes' theorem can help us to determine the probability of a record belonging to a class, given the data we observe.

source: Data Analysis with Open Source Tools, by Philipp K. Janert. O'Reilly Media, 2011.

Each term in this relationship has a name, and each plays a distinct role in any Bayesian calculation (including ours).

$$P(\text{class } C \mid \{x_i\}) = \frac{P(\{x_i\} \mid \text{class } C) \cdot P(\text{class } C)}{P(\{x_i\})}$$

This term is the **likelihood function**. *It represents the joint probability of observing features* $\{x_i\}$ *given that that record belongs to class* C.

$$P(\text{class } C \mid \{x_i\}) = \frac{\boxed{P(\{x_i\} \mid \text{class } C)} \cdot P(\text{class } C)}{P(\{x_i\})}$$

This term is the **likelihood function**. *It represents the joint probability of observing features* $\{x_i\}$ *given that that record belongs to class* C.

$$P(\text{class } C \mid \{x_i\}) = \frac{P(\{x_i\} \mid \text{class } C) \cdot P(\text{class } C)}{P(\{x_i\})}$$

We can observe the value of the likelihood function from the training data.

This term is the **prior probability** *of* C*. It represents the probability of a record belonging to class* C *before the data is taken into account.*

$$P(\text{class } C \mid \{x_i\}) = \frac{P(\{x_i\} \mid \text{class } C) \cdot \boxed{P(\text{class } C)}}{P(\{x_i\})}$$

This term is the **prior probability** *of C. It represents the probability of a record belonging to class C before the data is taken into account.*

$$P(\text{class } C \mid \{x_i\}) = \frac{P(\{x_i\} \mid \text{class } C) \cdot P(\text{class } C)}{P(\{x_i\})}$$

The value of the prior is also observed from the data.

This term is the **normalization constant.** *It doesn't depend on* C, *and is generally ignored until the end of the computation.*

$$P(\text{class } C \mid \{x_i\}) = \frac{P(\{x_i\} \mid \text{class } C) \cdot P(\text{class } C)}{\boxed{P(\{x_i\})}}$$

This term is the **normalization constant.** *It doesn't depend on C, and is generally ignored until the end of the computation.*

$$P(\text{class } C \mid \{x_i\}) = \frac{P(\{x_i\} \mid \text{class } C) \cdot P(\text{class } C)}{P(\{x_i\})}$$

The normalization constant doesn't tell us much.

This term is the **posterior probability** *of* C*. It represents the probability of a record belonging to class* C *after the data is taken into account.*

$$\boxed{P(\text{class } C \mid \{x_i\})} = \frac{P(\{x_i\} \mid \text{class } C) \cdot P(\text{class } C)}{P(\{x_i\})}$$

This term is the **posterior probability** *of C. It represents the probability of a record belonging to class C after the data is taken into account.*

$$P(\text{class } C \mid \{x_i\}) = \frac{P(\{x_i\} \mid \text{class } C) \cdot P(\text{class } C)}{P(\{x_i\})}$$

The goal of any Bayesian computation is to find ("learn") the posterior distribution of a particular variable.

The idea of Bayesian inference, then, is to **update** *our beliefs about the distribution of* C *using the data ("evidence") at our disposal.*

$$P(\text{class } C \mid \{x_i\}) = \frac{P(\{x_i\} \mid \text{class } C) \cdot P(\text{class } C)}{P(\{x_i\})}$$

Then we can use the posterior for prediction.

Methods	*Predictions*
"classical" (frequentist)	*point estimates*
Bayesian	*distributions*

Q: What piece of the puzzle we've seen so far looks like it could be intractably difficult in practice?

Remember the likelihood function?

$$P(\{x_i\} \mid C) = P(\{x_1, x_2, \ldots, x_n\}) \mid C)$$

Remember the likelihood function?

$$P(\{x_i\} \mid C) = P(\{x_1, x_2, \ldots, x_n\}) \mid C)$$

Observing this exactly would require us to have enough data for every possible combination of features to make a reasonable estimate.

Q: What piece of the puzzle we've seen so far looks like it could intractably difficult in practice?

A: Estimating the full likelihood function.

Q: So what can we do about it?

Q: So what can we do about it?

A: Make a simplifying assumption. In particular, we assume that the features x_i are conditionally independent from each other:

Q: So what can we do about it?

A: Make a simplifying assumption. In particular, we assume that the features x_i are conditionally independent from each other:

$$P(\{x_i\} \mid C) = P(x_1, x_2, \ldots, x_n \mid C) \approx P(x_1 \mid C) * P(x_2 \mid C) * \ldots * P(x_n \mid C)$$

LAB: NAÏVE BAYESIAN CLASSIFICATION

Q: What is this classification best suited for?

Q: What is this classification best suited for?

A: More often than not, NaiveBayes makes a great **text classifier**.

(Classic) Example: Classifying email as either spam or ham

Q: What are our features?

A: The text available in emails

(Classic) Example: Classifying email as either spam or ham

Q: How do we turn the text into features?

(Classic) Example: Classifying email as either spam or ham

Q: How do we turn the text into features?

A: Word counts, frequency matrices, tf-idf

(Classic) Example: Classifying email as either spam or ham

Q: How do we turn the text into features?

A: Word counts, frequency matrices, tf-idf

We can make alterations to this dictionary/features: dropping stop words, for example.

(Classic) Example: Classifying email as either spam or ham

html,	table,	Nigerian,	prince,	lunch,	break,	U.S.	spam
1	1	1	1	0	0	1	1
0	1	0	0	1	1	1	0

(Classic) Example: Classifying email as either spam or ham

html,	table,	Nigerian,	prince,	lunch,	break,	U.S.	spam
1	1	1	1	0	0	1	1
0	1	0	0	1	1	1	0

Now we want to learn P(word|spam) ie, what's the probability this word shows up given that it's spam?

This is what makes it supervised learning!

(Classic) Example: Classifying email as either spam or ham

html	table	Nigerian	prince	lunch	break	U.S.	spam
1	1	1	1	0	0	1	1
0	1	0	0	1	1	1	0

Now we want to learn P(word|spam) ie, what's the probability this word shows up given that it's spam?

(Classic) Example: Classifying email as either spam or ham

html	table	Nigerian	prince	lunch	break	U.S.	spam
1	1	1	1	0	0	1	1
0	1	0	0	1	1	1	0

Now that we've trained our data, we want to compute probability for each class (spam = 1 and spam = 0)