GENERAL ASSEMBLY

# INTRO TO DATA SCIENCE

## NAIVE BAYES CLASSIFICATION

# I. INTRO TO PROBABILITY
# II. NAÏVE BAYES CLASSIFICATION

# LAB:
# III. 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 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** $\Omega$. 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** $\Omega$. Event $A$ is a member of the sample space, as is every other event.

The total 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 $P(AB)$.

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 $P(AB)$.

NOTE

The symbol ∩ is often used for intersection. For example, "$P(A \cap B)$".

Q: Is $P(AB)$ equal to $P(A)P(B)$?

Q: Is $P(AB)$ equal to $P(A)P(B)$?

A: Maybe, maybe not. More later…

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 $P(A|B) = P(AB) / P(B)$.

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 $P(A|B) = P(AB) / P(B)$.

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

NOTE

This information about B *transforms* the sample space.

Take a moment to convince yourself of this!

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 $P(A|B) = P(A)$.

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 $P(A|B) = P(A)$.

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

$$P(A|B) = P(AB) / P(B) = P(A) \rightarrow P(AB) = P(A) * P(B)$$

Probably the only proof in the course:

Probably the only proof in the course:

$P(AB) = P(A|B) * P(B)$ from earlier

Probably the only proof in the course:

$P(AB) = P(A|B) * P(B)$ from earlier

$P(BA) = P(B|A) * P(A)$ by substitution

Probably the only proof in the course:

$P(AB) = P(A|B) * P(B)$ from earlier

$P(BA) = P(B|A) * P(A)$ by substitution

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

Probably the only proof in the course:

$P(AB) = P(A|B) * P(B)$ from earlier

$P(BA) = P(B|A) * P(A)$ by substitution

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

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

Probably the only proof in the course:

$P(AB) = P(A|B) * P(B)$ from earlier

$P(BA) = P(B|A) * P(A)$ by substitution

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

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

→ $P(A|B) = P(B|A) * P(A) / P(B)$ by rearranging last step

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

$$P(A|B) = P(B|A) * P(A) / P(B)$$

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

$$P(A|B) = P(B|A) * P(A) / P(B)$$

Some facts:

- This is a simple algebraic relationship using elementary definitions.

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

$$P(A|B) = P(B|A) * P(A) / P(B)$$

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:

$$P(A|B) = P(B|A) * P(A) / P(B)$$

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.

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.

# II. NAÏVE BAYES 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 $c$.

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

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 approximate 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.

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

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.

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

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.

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

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 intractably difficult in practice?

# Remember the likelihood function?

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

Remember the likelihood function?

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

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:

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

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:

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

This "naïve" assumption simplifies the likelihood function to make it tractable.

# III. SPAM FILTER

# EXERCISE – SPAM FILTER (DOCUMENT CLASSIFICATION)



### KEY OBJECTIVES

- preprocess data
- perform naïve Bayes classification

### TOOLS

- tm, R