GENERAL ASSEMBLY

INTRO TO DATA SCIENCE

BAYES THEOREM, NAIVE BAYES CLASSIFIERS

AGENDA

PROBABILITY THEORY

BAYES THEOREM

NAIVE BAYES CLASSIFIERS

PROBABILITY THEORY

PROBABILITY vs STATISTICS

- Statistics:

- Probability:

PROBABILITY vs STATISTICS

- Statistics:
  - *given data, infer causes related to the data*
- Probability:

PROBABILITY vs STATISTICS

- Statistics:
  - *given data, infer causes related to the data*

- Probability:
  - *given a description of the causes, predict the data*

INTRO TO PROBABILITY

Q: What is a “probability” ?

INTRO TO PROBABILITY

Q: What is a “probability” ?

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

INTRO TO PROBABILITY

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)***

INTRO TO PROBABILITY

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

INTRO TO PROBABILITY

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

A: This set is called the sample space $\boldsymbol{\Omega}$. Event $\boldsymbol{A}$ is a member of the sample space, as is every other event.

INTRO TO PROBABILITY

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

INTRO TO PROBABILITY: Joint Probability

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

INTRO TO PROBABILITY: Joint Probability

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

A: With the ***joint probability*** of *A* and B,

$$\boldsymbol{P(AB)}$$

INTRO TO PROBABILITY: Conditional Probability

Suppose event ***B*** has occurred that affects ***A***.

Q: What quantity represents the probability of *A* given this information about *B*?

A: This is called the conditional probability of *A* given *B*,

$$P(A|B) = P(AB) \,/\, P(B)$$

INTRO TO PROBABILITY: Conditional Independence

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

INTRO TO PROBABILITY: Conditional Independence

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

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

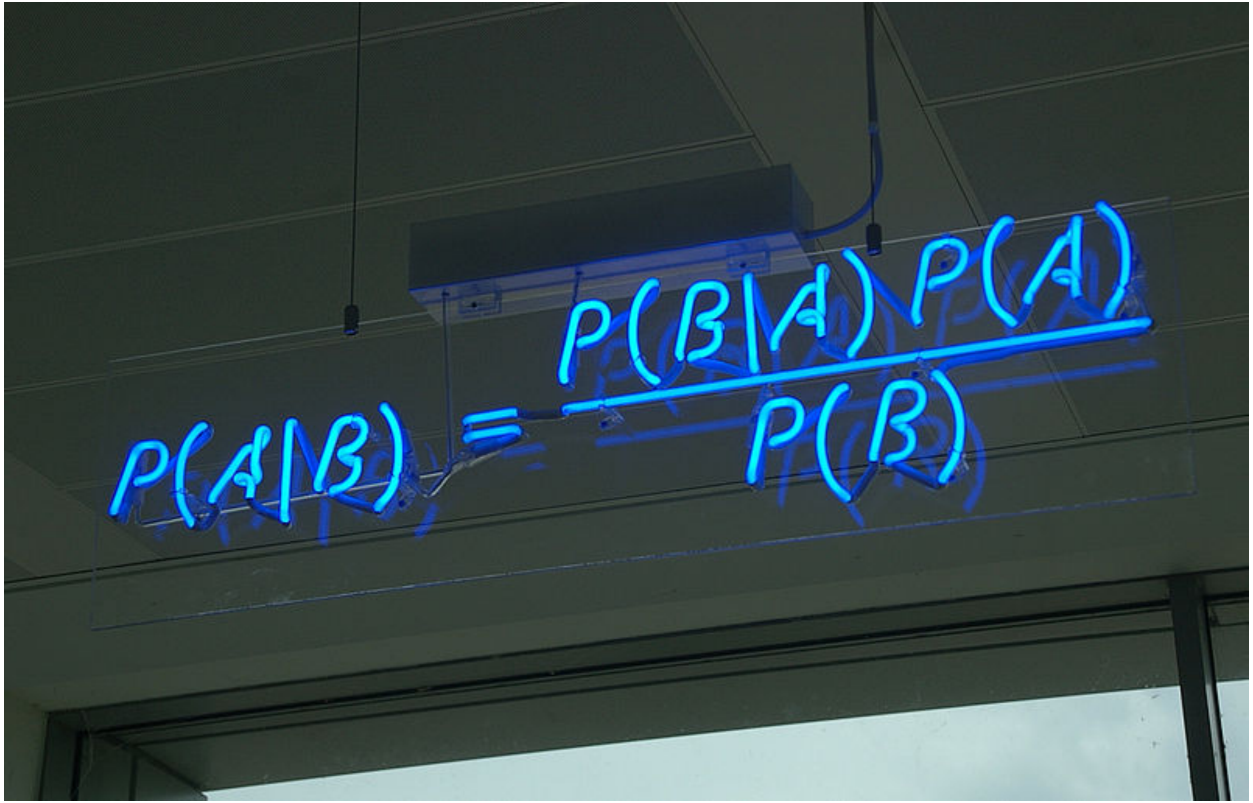
$$P(A|B) = P(A)$$

INTRO TO BAYES THEOREM

THOMAS BAYES

English statistician, philosopher and Presbyterian minister, known for having formulated a specific case of the theorem that bears his name

Bayes Theorem



Frequentist vs. Bayesian

- ***Frequentist***:

  Probability measures a proportion of outcomes

- ***Bayesian***:

  Probability measures a degree of belief

CHECK THIS OUT

Probably the only proof in the course:

CHECK THIS OUT

Probably the only proof in the course:

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

CHECK THIS OUT

Probably the only proof in the course:

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

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

CHECK THIS OUT

Probably the only proof in the course:

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

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

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

CHECK THIS OUT

Probably the only proof in the course:

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

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

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

$$\rightarrow \boldsymbol{P(A|B) * P(B) = P(B|A) * P(A)}$$

CHECK THIS OUT

Probably the only proof in the course:

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

CHECK THIS OUT

Probably the only proof in the course:

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

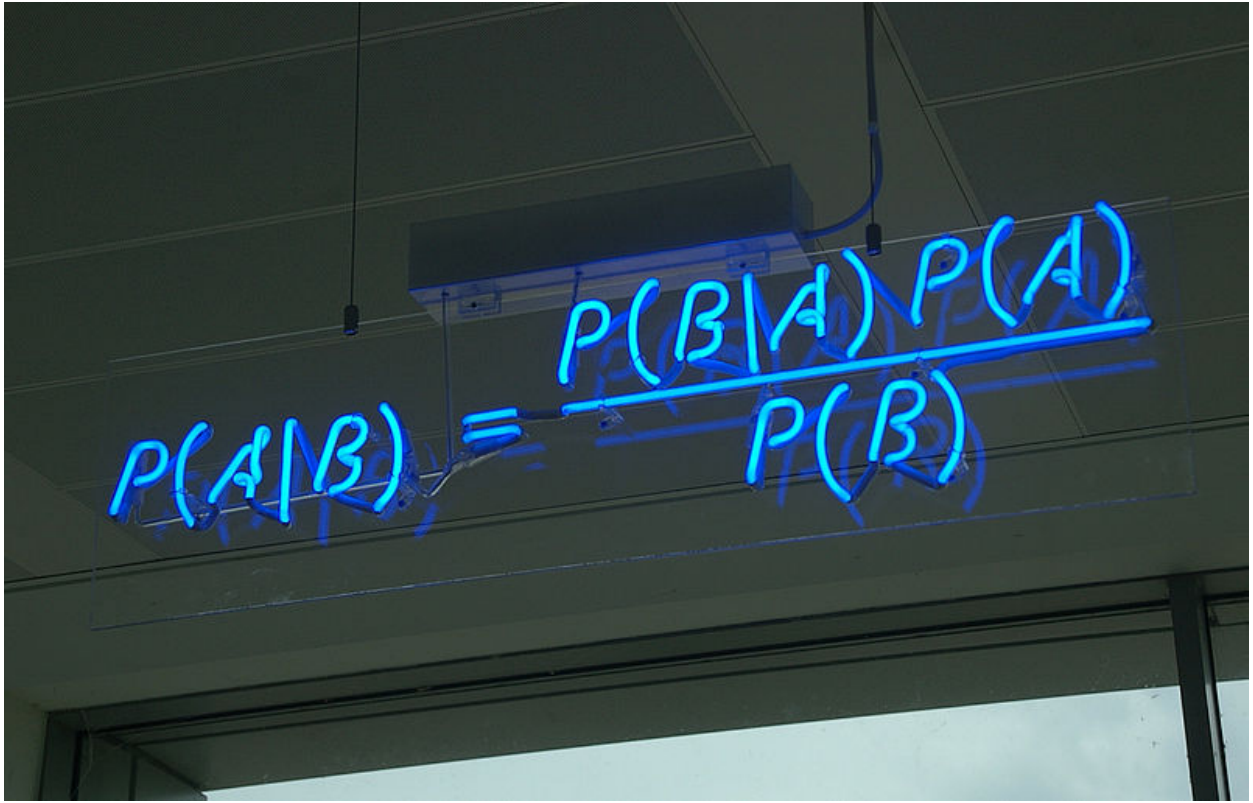
→ $P(A|B) = P(B|A) * P(A) / P(B)$ by rearranging

BAYES’ THEOREM

This result is called Bayes’ theorem. Here it is again:

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

Bayes Theorem



BAYES’ THEOREM

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

Some comments:

- This is a simple algebraic relationship using elementary definitions

BAYES' THEOREM

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

Some comments:

- This is a simple algebraic relationship using elementary definitions
- It's a very powerful computational tool

BAYES' THEOREM

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

Some comments:

- This is a simple algebraic relationship using elementary definitions
- It's a very powerful computational tool
- It's kind of a "wormhole" between two different "interpretations" of probability

INTERPRETATIONS OF PROBABILITY

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.

INTERPRETATIONS OF PROBABILITY

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

INTERPRETATIONS OF PROBABILITY

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.

INTRO TO DATA SCIENCE

Naive Bayes Classification

BAYESIAN INFERENCE

Suppose we have a dataset with features $\boldsymbol{x_1}, \ldots, \boldsymbol{x_n}$ and class label C.

What can we say about classification using Bayes’ theorem?

BAYESIAN INFERENCE

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.

SOME TERMINOLOGY

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

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

THE LIKELIHOOD FUNCTION

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

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

THE LIKELIHOOD FUNCTION

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

We can approximate the value of the ***likelihood*** from the training data.

THE PRIOR

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 PRIOR

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

The value of the prior is also observed from data.

THE NORMALIZATION CONSTANT

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 POSTERIOR

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 POSTERIOR

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

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

BAYESIAN INFERENCE

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.

A QUICK COMPARISON

Method	Prediction
Frequentist (classical)	point estimates
Bayesian	probability distributions

NAÏVE BAYES CLASSIFICATION

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

NAÏVE BAYES CLASSIFICATION

Remember the likelihood function?

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

NAÏVE BAYES CLASSIFICATION

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.

NAÏVE BAYES CLASSIFICATION

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.

NAÏVE BAYES CLASSIFICATION

Q: So what can we do about it?

NAÏVE BAYES CLASSIFICATION

Q: So what can we do about it?

A: Make a simplifying assumption.

In particular, we assume that the features $\boldsymbol{x_i}$ are ***conditionally independent*** from each other:

NAÏVE BAYES CLASSIFICATION

- In particular, we assume that the features x_i are ***conditionally independent*** from each other:

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

PRACTICE: NAIVE BAYES!