Bayesian Statistics and Classification

Data Science camp

June 7, 2018

Scene of the Crime

- A murder has occurred!

Scene of the Crime

- A murder has occurred!
  - Suppose police arrive at a crime scene and encounter a murder

Scene of the Crime

- A murder has occurred!
  - Suppose police arrive at a crime scene and encounter a murder
  - They happen to find DNA evidence at the scene, possibly belonging to the murderer...

Scene of the Crime

- A murder has occurred!
  - Suppose police arrive at a crime scene and encounter a murder
  - They happen to find DNA evidence at the scene, possibly belonging to the murderer...
  - The police reference this to their database containing 100,000 people who have previously committed a crime

Scene of the Crime

- A murder has occurred!
  - Suppose police arrive at a crime scene and encounter a murder
  - They happen to find DNA evidence at the scene, possibly belonging to the murderer...
  - The police reference this to their database containing 100,000 people who have previously committed a crime
  - A match is found! Analysts say that the test gives a false positive 1 out of every 1,000,000 times.

Scene of the Crime

- A murder has occurred!
  - Suppose police arrive at a crime scene and encounter a murder
  - They happen to find DNA evidence at the scene, possibly belonging to the murderer...
  - The police reference this to their database containing 100,000 people who have previously committed a crime
  - A match is found! Analysts say that the test gives a false positive 1 out of every 1,000,000 times.
  - How confident are you that the DNA match belongs to the right person (let's call them OJ)? 99.9999% sure?

Prosecution in a Court of Law

- The police appoint a prosecutor to accuse OJ of murder in a court of law...

Prosecution in a Court of Law

- The police appoint a prosecutor to accuse OJ of murder in a court of law...
  - The prosecutor makes a hard fought argument claiming that the DNA test is correct 99.9999% of the time. As such, the culprit is beyond a reasonable doubt, OJ.

Prosecution in a Court of Law

- The police appoint a prosecutor to accuse OJ of murder in a court of law...
  - The prosecutor makes a hard fought argument claiming that the DNA test is correct 99.9999% of the time. As such, the culprit is beyond a reasonable doubt, OJ.
  - If you are on the jury, do you believe this to be true? Why or why not? What are some things that could change your mind?

Prosecution in a Court of Law

- The police appoint a prosecutor to accuse OJ of murder in a court of law...
  - The prosecutor makes a hard fought argument claiming that the DNA test is correct 99.9999% of the time. As such, the culprit is beyond a reasonable doubt, OJ.
  - If you are on the jury, do you believe this to be true? Why or why not? What are some things that could change your mind?
  - The problem has colloquially been termed *prosecutor's fallacy* and is a misuse of conditional probability. Let's take a closer look while the jury deliberates.

Conditional Probability

- We express the probability that an event A happens given that another event B has occurred as $\mathbb{P}(A|B)$.

Conditional Probability

- We express the probability that an event A happens given that another event B has occurred as $\mathbb{P}(A|B)$.
- How do we calculate $\mathbb{P}(A|B)$?
  - $\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$

Conditional Probability

- We express the probability that an event A happens given that another event B has occurred as $\mathbb{P}(A|B)$.
- How do we calculate $\mathbb{P}(A|B)$?
  - $\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$
- Is $\mathbb{P}(A|B) = P(B|A)$?

Conditional Probability

- We express the probability that an event A happens given that another event B has occurred as $\mathbb{P}(A|B)$.
- How do we calculate $\mathbb{P}(A|B)$?
  - $\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$
- Is $\mathbb{P}(A|B) = P(B|A)$?
  - If true, this would say $\frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B \cap A)}{\mathbb{P}(A)}$.

Conditional Probability

- We express the probability that an event A happens given that another event B has occurred as $\mathbb{P}(A|B)$.
- How do we calculate $\mathbb{P}(A|B)$?
    - $\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$
- Is $\mathbb{P}(A|B) = P(B|A)$?
    - If true, this would say $\frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B \cap A)}{\mathbb{P}(A)}$.
    - Is $\mathbb{P}(A \cap B) = \mathbb{P}(B \cap A)$? What about $\mathbb{P}(B) = \mathbb{P}(A)$?

Conditional Probability

- We express the probability that an event A happens given that another event B has occurred as $\mathbb{P}(A|B)$.
- How do we calculate $\mathbb{P}(A|B)$?
  - $\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$
- Is $\mathbb{P}(A|B) = P(B|A)$?
  - If true, this would say $\frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B \cap A)}{\mathbb{P}(A)}$.
  - Is $\mathbb{P}(A \cap B) = \mathbb{P}(B \cap A)$? What about $\mathbb{P}(B) = \mathbb{P}(A)$?
- Very rarely are these equal to one another! Even so, the above formula is very useful. Keep it in mind as we move forward.

Prosecutor's fallacy

- Let D denote the event of a DNA match and I denote the event where a person is innocent. The prosecutor reports $\mathbb{P}(D|I)$ as $\mathbb{P}(I|D)$!
- To illustrate, $\mathbb{P}(D|I) = .000001$ and $\mathbb{P}(I|D) =$?

Prosecutor's fallacy

- Let D denote the event of a DNA match and I denote the event where a person is innocent. The prosecutor reports $\mathbb{P}(D|I)$ as $\mathbb{P}(I|D)$!
- To illustrate, $\mathbb{P}(D|I) = .000001$ and $\mathbb{P}(I|D) = ?$
  - If suspects in the police database are innocent with probability $\frac{1}{10}$, then $\mathbb{P}(D) = .000001 * \mathbb{P}(I) + 1 * \mathbb{P}(G) = .9000001$ so $\mathbb{P}(I|D) = \frac{\mathbb{P}(D \cap I)}{.9000001}$
  - If suspects have a uniform probability of being innocent and $\mathbb{P}(I) = \frac{99999}{100000}$. Then, $\mathbb{P}(D) = .000001 * .99999 + 1 * .00001 = .0000109999$. So, $\mathbb{P}(I|D) = \frac{\mathbb{P}(D \cap I)}{0.0000055}$.

Prosecutor's fallacy

- Let D denote the event of a DNA match and I denote the event where a person is innocent. The prosecutor reports $\mathbb{P}(D|I)$ as $\mathbb{P}(I|D)$!
- To illustrate, $\mathbb{P}(D|I) = .000001$ and $\mathbb{P}(I|D) = ?$
  - If suspects in the police database are innocent with probability $\frac{1}{10}$, then $\mathbb{P}(D) = .000001 * \mathbb{P}(I) + 1 * \mathbb{P}(G) = .9000001$ so $\mathbb{P}(I|D) = \frac{\mathbb{P}(D \cap I)}{.9000001}$
  - If suspects have a uniform probability of being innocent and $\mathbb{P}(I) = \frac{99999}{100000}$. Then, $\mathbb{P}(D) = .000001 * .99999 + 1 * .00001 = .0000109999$. So, $\mathbb{P}(I|D) = \frac{\mathbb{P}(D \cap I)}{0.0000055}$.
    - In other words, *prior* beliefs can significantly impact results. Is there a systematic way to update one's beliefs when given new information?

Prosecutor's fallacy

- Let D denote the event of a DNA match and I denote the event where a person is innocent. The prosecutor reports $\mathbb{P}(D|I)$ as $\mathbb{P}(I|D)$!
- To illustrate, $\mathbb{P}(D|I) = .000001$ and $\mathbb{P}(I|D) = ?$
  - If suspects in the police database are innocent with probability $\frac{1}{10}$, then $\mathbb{P}(D) = .000001 * \mathbb{P}(I) + 1 * \mathbb{P}(G) = .9000001$ so $\mathbb{P}(I|D) = \frac{\mathbb{P}(D \cap I)}{.9000001}$
  - If suspects have a uniform probability of being innocent and $\mathbb{P}(I) = \frac{99999}{100000}$. Then, $\mathbb{P}(D) = .000001 * .99999 + 1 * .00001 = .0000109999$. So, $\mathbb{P}(I|D) = \frac{\mathbb{P}(D \cap I)}{0.0000055}$.
    - In other words, *prior* beliefs can significantly impact results. Is there a systematic way to update one's beliefs when given new information?
    - Can anyone calculate $\mathbb{P}(D \cap I)$? Do we know this information?

Bayes' Theorem

- Named after Thomas Bayes (1701-1761) who gave a formula allowing for a convenient way to update one's beliefs based on new evidence.

Bayes' Theorem

- Named after Thomas Bayes (1701-1761) who gave a formula allowing for a convenient way to update one's beliefs based on new evidence.
- This formula is given by:

$$\mathbb{P}(I|D) = \frac{\mathbb{P}(D|I)\mathbb{P}(I)}{\mathbb{P}(D)}$$

.

  - $\mathbb{P}(D)$ and $\mathbb{P}(I)$ are the probabilities of observing D and I independently. They are known as the *marginals* of D and I in the joint distribution $\mathbb{P}(D, I)$.

Bayes' Theorem

- Named after Thomas Bayes (1701-1761) who gave a formula allowing for a convenient way to update one's beliefs based on new evidence.
- This formula is given by:

$$\mathbb{P}(I|D) = \frac{\mathbb{P}(D|I)\mathbb{P}(I)}{\mathbb{P}(D)}$$

.

  - $\mathbb{P}(D)$ and $\mathbb{P}(I)$ are the probabilities of observing D and I independently. They are known as the *marginals* of D and I in the joint distribution $\mathbb{P}(D, I)$.
- Recall, our answer in calculating $\mathbb{P}(I|D)$ depended on our *prior* beliefs regarding the distribution of innocent people contained in the database.

Bayes' Theorem

- Named after Thomas Bayes (1701-1761) who gave a formula allowing for a convenient way to update one's beliefs based on new evidence.
- This formula is given by:

$$\mathbb{P}(I|D) = \frac{\mathbb{P}(D|I)\mathbb{P}(I)}{\mathbb{P}(D)}$$

.

  - $\mathbb{P}(D)$ and $\mathbb{P}(I)$ are the probabilities of observing D and I independently. They are known as the *marginals* of D and I in the joint distribution $\mathbb{P}(D, I)$.
- Recall, our answer in calculating $\mathbb{P}(I|D)$ depended on our *prior* beliefs regarding the distribution of innocent people contained in the database.
- We typically call:
  - $\mathbb{P}(D|I)$ the *class-conditional* or *likelihood*

Bayes' Theorem

- Named after Thomas Bayes (1701-1761) who gave a formula allowing for a convenient way to update one's beliefs based on new evidence.
- This formula is given by:

$$\mathbb{P}(I|D) = \frac{\mathbb{P}(D|I)\mathbb{P}(I)}{\mathbb{P}(D)}$$

.

  - $\mathbb{P}(D)$ and $\mathbb{P}(I)$ are the probabilities of observing D and I independently. They are known as the *marginals* of D and I in the joint distribution $\mathbb{P}(D, I)$.
- Recall, our answer in calculating $\mathbb{P}(I|D)$ depended on our *prior* beliefs regarding the distribution of innocent people contained in the database.
- We typically call:
  - $\mathbb{P}(D|I)$ the *class-conditional* or *likelihood*
  - $\mathbb{P}(I)$ the *prior*

Bayes' Theorem

- Named after Thomas Bayes (1701-1761) who gave a formula allowing for a convenient way to update one's beliefs based on new evidence.
- This formula is given by:

$$\mathbb{P}(I|D) = \frac{\mathbb{P}(D|I)\mathbb{P}(I)}{\mathbb{P}(D)}$$

.

  - $\mathbb{P}(D)$ and $\mathbb{P}(I)$ are the probabilities of observing D and I independently. They are known as the *marginals* of D and I in the joint distribution $\mathbb{P}(D, I)$.
- Recall, our answer in calculating $\mathbb{P}(I|D)$ depended on our *prior* beliefs regarding the distribution of innocent people contained in the database.
- We typically call:
  - $\mathbb{P}(D|I)$ the *class-conditional* or *likelihood*
  - $\mathbb{P}(I)$ the *prior*
  - $\mathbb{P}(D)$ the evidence

Bayes' Theorem

- Named after Thomas Bayes (1701-1761) who gave a formula allowing for a convenient way to update one's beliefs based on new evidence.
- This formula is given by:

$$\mathbb{P}(I|D) = \frac{\mathbb{P}(D|I)\mathbb{P}(I)}{\mathbb{P}(D)}$$

.

  - $\mathbb{P}(D)$ and $\mathbb{P}(I)$ are the probabilities of observing D and I independently. They are known as the *marginals* of D and I in the joint distribution $\mathbb{P}(D, I)$.
- Recall, our answer in calculating $\mathbb{P}(I|D)$ depended on our *prior* beliefs regarding the distribution of innocent people contained in the database.
- We typically call:
  - $\mathbb{P}(D|I)$ the *class-conditional* or *likelihood*
  - $\mathbb{P}(I)$ the *prior*
  - $\mathbb{P}(D)$ the evidence
  - $\mathbb{P}(I|D)$ the *posterior*

Bayes' Theorem in Court

- $\mathbb{P}(D|I) = .000001$ is our class-conditional

Bayes' Theorem in Court

- $\mathbb{P}(D|I) = .000001$ is our class-conditional
- $\mathbb{P}(I)$ is the prior (take the two examples where $P(I) = .1$ and $P(I) = 0.99999$)

Bayes' Theorem in Court

- $\mathbb{P}(D|I) = .000001$ is our class-conditional
- $\mathbb{P}(I)$ is the prior (take the two examples where $P(I) = .1$ and $P(I) = 0.99999$)
- $\mathbb{P}(D)$ is the evidence given by the DNA test:
  - $\mathbb{P}(D) = \mathbb{P}(D|I)\mathbb{P}(I) + \mathbb{P}(D|G)\mathbb{P}(G)$

Bayes' Theorem in Court

- $\mathbb{P}(D|I) = .000001$ is our class-conditional
- $\mathbb{P}(I)$ is the prior (take the two examples where $P(I) = .1$ and $P(I) = 0.99999$)
- $\mathbb{P}(D)$ is the evidence given by the DNA test:
  - $\mathbb{P}(D) = \mathbb{P}(D|I)\mathbb{P}(I) + \mathbb{P}(D|G)\mathbb{P}(G)$
  - Case 1: $\mathbb{P}(D) = .000001 * .1 + 1 * .9 = .9000001$

Bayes' Theorem in Court

- $\mathbb{P}(D|I) = .000001$ is our class-conditional
- $\mathbb{P}(I)$ is the prior (take the two examples where $P(I) = .1$ and $P(I) = 0.99999$)
- $\mathbb{P}(D)$ is the evidence given by the DNA test:
  - $\mathbb{P}(D) = \mathbb{P}(D|I)\mathbb{P}(I) + \mathbb{P}(D|G)\mathbb{P}(G)$
  - Case 1: $\mathbb{P}(D) = .000001 * .1 + 1 * .9 = .9000001$
  - Case 2: $\mathbb{P}(D) = .000001 * .99999 + 1 * .00001 = .0000109999$

Bayes' Theorem in Court

- $\mathbb{P}(D|I) = .000001$ is our class-conditional
- $\mathbb{P}(I)$ is the prior (take the two examples where $P(I) = .1$ and $P(I) = 0.99999$)
- $\mathbb{P}(D)$ is the evidence given by the DNA test:
  - $\mathbb{P}(D) = \mathbb{P}(D|I)\mathbb{P}(I) + \mathbb{P}(D|G)\mathbb{P}(G)$
  - Case 1: $\mathbb{P}(D) = .000001 * .1 + 1 * .9 = .9000001$
  - Case 2: $\mathbb{P}(D) = .000001 * .99999 + 1 * .00001 = .0000109999$
- And so, $\mathbb{P}(I|D) = \frac{\mathbb{P}(D|I)\mathbb{P}(I)}{\mathbb{P}(D)}$:

Bayes' Theorem in Court

- $\mathbb{P}(D|I) = .000001$ is our class-conditional
- $\mathbb{P}(I)$ is the prior (take the two examples where $P(I) = .1$ and $P(I) = 0.99999$)
- $\mathbb{P}(D)$ is the evidence given by the DNA test:
  - $\mathbb{P}(D) = \mathbb{P}(D|I)\mathbb{P}(I) + \mathbb{P}(D|G)\mathbb{P}(G)$
  - Case 1: $\mathbb{P}(D) = .000001 * .1 + 1 * .9 = .9000001$
  - Case 2: $\mathbb{P}(D) = .000001 * .99999 + 1 * .00001 = .0000109999$
- And so, $\mathbb{P}(I|D) = \frac{\mathbb{P}(D|I)\mathbb{P}(I)}{\mathbb{P}(D)}$:
  - Case 1: $\frac{(.000001*.1)}{.9000001} < 1\%$

Bayes' Theorem in Court

- $\mathbb{P}(D|I) = .000001$ is our class-conditional
- $\mathbb{P}(I)$ is the prior (take the two examples where $P(I) = .1$ and $P(I) = 0.99999$)
- $\mathbb{P}(D)$ is the evidence given by the DNA test:
  - $\mathbb{P}(D) = \mathbb{P}(D|I)\mathbb{P}(I) + \mathbb{P}(D|G)\mathbb{P}(G)$
  - Case 1: $\mathbb{P}(D) = .000001 * .1 + 1 * .9 = .9000001$
  - Case 2: $\mathbb{P}(D) = .000001 * .99999 + 1 * .00001 = .0000109999$
- And so, $\mathbb{P}(I|D) = \frac{\mathbb{P}(D|I)\mathbb{P}(I)}{\mathbb{P}(D)}$:
  - Case 1: $\frac{(.000001 * .1)}{.9000001} < 1\%$
  - Case 2: $\frac{(.000001 * 0.99999)}{.0000109999} \approx 90.1\%$

The Prior Distribution

As seen in the last slide, our prior beliefs, $\mathbb{P}(I)$, can significantly impact those of our posterior. How do you get a “good” estimate on the prior?

The Prior Distribution

As seen in the last slide, our prior beliefs, $\mathbb{P}(I)$, can significantly impact those of our posterior. How do you get a "good" estimate on the prior?

- People frequently just ask an expert to provide them one

The Prior Distribution

As seen in the last slide, our prior beliefs, $\mathbb{P}(I)$, can significantly impact those of our posterior. How do you get a "good" estimate on the prior?

- People frequently just ask an expert to provide them one
- Estimate it from the data: $\hat{p}_1(I) = \frac{\text{\# of convicted murderers}}{\text{\# of people}}$

The Prior Distribution

As seen in the last slide, our prior beliefs, $\mathbb{P}(I)$, can significantly impact those of our posterior. How do you get a "good" estimate on the prior?

- People frequently just ask an expert to provide them one
- Estimate it from the data: $\hat{p}_1(I) = \frac{\text{\# of convicted murderers}}{\text{\# of people}}$
  - Is this a good estimate? What about $\hat{p}_2(I) = \frac{\text{\# of people with knives}}{\text{\# of people}}$?
- Also known as the *class prior* or *predictor prior* probability

Classification

Can we use Bayes' rule to classify a suspect as guilty or innoncent? Is there a way to let a computer be both judge and jury? Yes!

Classification

Can we use Bayes' rule to classify a suspect as guilty or innoncent? Is there a way to let a computer be both judge and jury? Yes!

- One such way is by comparing posterior probabilities:
    - For the suspect O.J.:

Classification

Can we use Bayes' rule to classify a suspect as guilty or innoncent? Is there a way to let a computer be both judge and jury? Yes!

- One such way is by comparing posterior probabilities:
    - For the suspect O.J.:
        - Is $\mathbb{P}(G|D) > \mathbb{P}(I|D)$?

Classification

Can we use Bayes' rule to classify a suspect as guilty or innoncent? Is there a way to let a computer be both judge and jury? Yes!

- One such way is by comparing posterior probabilities:
  - For the suspect O.J.:
    - Is $\mathbb{P}(G|D) > \mathbb{P}(I|D)$?

    Yes: Classify as guilty

Classification

Can we use Bayes' rule to classify a suspect as guilty or innoncent? Is there a way to let a computer be both judge and jury? Yes!

- One such way is by comparing posterior probabilities:
  - For the suspect O.J.:
    - Is $\mathbb{P}(G|D) > \mathbb{P}(I|D)$?

    Yes: Classify as guilty

    No: Classify as innocent
- We calculate this via Bayes' rule. In other words, take the maximum of:
  - $\frac{\mathbb{P}(D|G)\mathbb{P}(G)}{\mathbb{P}(D)}$ and $\frac{\mathbb{P}(D|I)\mathbb{P}(I)}{\mathbb{P}(D)}$

Classification

Can we use Bayes' rule to classify a suspect as guilty or innoncent? Is there a way to let a computer be both judge and jury? Yes!

- One such way is by comparing posterior probabilities:
  - For the suspect O.J.:
    - Is $\mathbb{P}(G|D) > \mathbb{P}(I|D)$?

    Yes: Classify as guilty

    No: Classify as innocent
- We calculate this via Bayes' rule. In other words, take the maximum of:
  - $\frac{\mathbb{P}(D|G)\mathbb{P}(G)}{\mathbb{P}(D)}$ and $\frac{\mathbb{P}(D|I)\mathbb{P}(I)}{\mathbb{P}(D)}$
  - Since these probabilities are positive, we don't actually need to calculate $\mathbb{P}(D)$.

Classification

Can we use Bayes' rule to classify a suspect as guilty or innoncent? Is there a way to let a computer be both judge and jury? Yes!

- One such way is by comparing posterior probabilities:
  - For the suspect O.J.:
    - Is $\mathbb{P}(G|D) > \mathbb{P}(I|D)$?

    Yes: Classify as guilty

    No: Classify as innocent
- We calculate this via Bayes' rule. In other words, take the maximum of:
  - $\frac{\mathbb{P}(D|G)\mathbb{P}(G)}{\mathbb{P}(D)}$ and $\frac{\mathbb{P}(D|I)\mathbb{P}(I)}{\mathbb{P}(D)}$
  - Since these probabilities are positive, we don't actually need to calculate $\mathbb{P}(D)$.
  - We then need only check the scores (no longer probabilities): $\mathbb{P}(D|G)\mathbb{P}(G)$ and $\mathbb{P}(D|I)\mathbb{P}(I)$

Classification

Can we use Bayes' rule to classify a suspect as guilty or innoncent? Is there a way to let a computer be both judge and jury? Yes!

- One such way is by comparing posterior probabilities:
  - For the suspect O.J.:
    - Is $\mathbb{P}(G|D) > \mathbb{P}(I|D)$?

    Yes: Classify as guilty

    No: Classify as innocent
- We calculate this via Bayes' rule. In other words, take the maximum of:
  - $\frac{\mathbb{P}(D|G)\mathbb{P}(G)}{\mathbb{P}(D)}$ and $\frac{\mathbb{P}(D|I)\mathbb{P}(I)}{\mathbb{P}(D)}$
  - Since these probabilities are positive, we don't actually need to calculate $\mathbb{P}(D)$.
  - We then need only check the scores (no longer probabilities): $\mathbb{P}(D|G)\mathbb{P}(G)$ and $\mathbb{P}(D|I)\mathbb{P}(I)$
  - Is this a good classification rule? Why or why not?

Monty Hall and Naive Bayes

Let's get started working with Bayes' rule and classification through a few exercises!