GENERAL ASSEMBLY

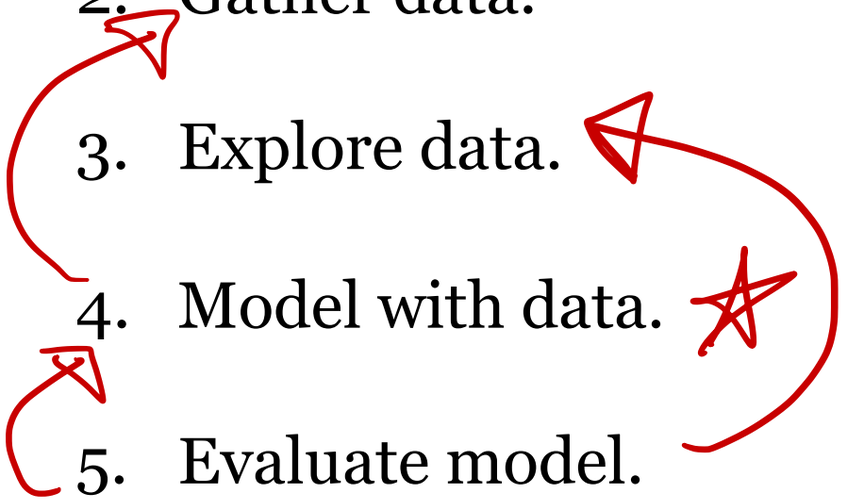
NAÏVE BAYES

Matt Brems

Data Science Immersive, GA DC

DATA SCIENCE PROCESS

1. Define problem.
2. Gather data.
3. Explore data.
4. Model with data.
5. Evaluate model.
6. Answer problem.



LEARNING OBJECTIVES

- By the end of this lesson, students should be able to:
    - **Intuitively explain** how Bayes' Theorem can be used as a modeling tactic.
    - **Implement** Naive Bayes in scikit-learn.
    - **Discuss** assumptions, advantages, and disadvantages of Naive Bayes as a classifier.

CONDITIONAL PROBABILITY

- Recall that we use $P(A)$ to refer to the probability that A occurs, where A is some event.

"P of A"

- If we want to describe the probability that A occurs given that we know something else to be true, we use $P(A|B)$.

"P of A given B"

A = roll a 2

B = roll an even

P(A) = 1/6

P(A|B) = 1/3

- Note that $P(A|B)$ is usually not the same as $P(B|A)$!

↳ 1/3

↳ 1

BAYES' THEOREM

- Bayes' Theorem (Bayes' Rule) relates $P(A|B)$ to $P(B|A)$.

$$P(A \text{ and } B) = P(B \text{ and } A)$$

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

multiplication rule for probabilities

divided both sides by $P(B)$

Bayes' Theorem

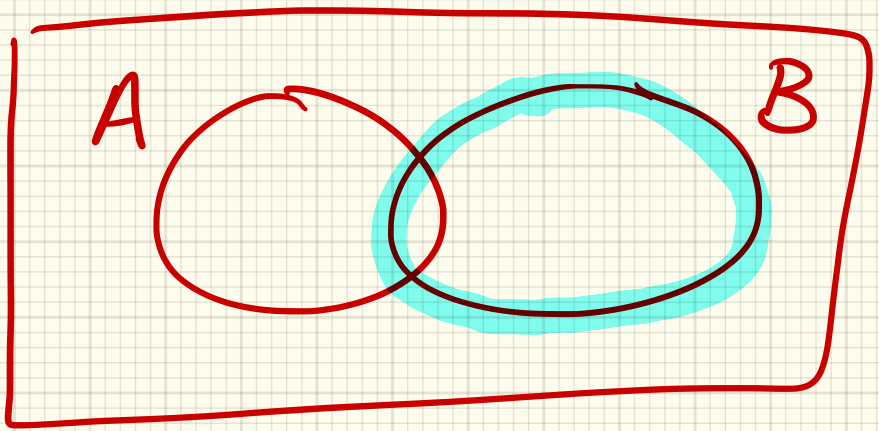
$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

BREAKING DOWN BAYES' THEOREM

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

- $P(A)$ is the probability that A occurs given no supplemental information.
  ↳ prior prob. of A
- $P(B|A)$ is the probability of B given that A is true.

- $P(B)$ is the probability that B occurs given no supplemental information.
  - $P(B)$ what we scale $P(B|A)P(A)$ by to ensure we are only looking at A within the context of B occurring.



S = sample space

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)} = \frac{P(B|A)P(A)}{P(B)}$$

APPLYING BAYES' THEOREM TO SPAM CLASSIFICATION

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

- Bayes' Theorem is really neatly set up as a classification model.
- We can estimate the probability – `.predict_proba()` – that an observation falls into a specific class, then classify that observation – `.predict()` – accordingly!

APPLYING BAYES' THEOREM TO SPAM CLASSIFICATION

$$P(\text{spam}|\text{words in email}) = \frac{P(\text{words in email}|\text{spam})P(\text{spam})}{P(\text{words in email})}$$

$Y = \begin{cases} 1 & \text{if spam} \\ 0 & \text{else} \end{cases}$

given the words in my email, what is the prob. that my email is spam?

APPLYING BAYES' THEOREM TO SPAM CLASSIFICATION

$$P(\text{spam}|\text{words}) = \frac{P(w_1|\text{spam})P(w_2|w_1 \cap \text{spam})P(w_3|w_2 \cap w_1 \cap \text{spam}) \cdots P(\text{spam})}{P(w_1)P(w_2|w_1)P(w_3|w_2 \cap w_1) \cdots}$$

words = $w_1 \cap w_2 \cap w_3 \cap \cdots$ (and)

- This gets **really** complicated. Can we simplify this?

Yes! But we have to make an assumption.

$$P(w_{100}|w_{99} \cap w_{98} \cap w_{97} \cap \cdots w_1)$$

NAÏVE BAYES

- The Naïve Bayes classification algorithm is a:
  - classification modeling technique
  - that relies on Bayes Theorem
  - that makes one simplifying assumption.

- **We assume that our features are independent of one another.**

↳ words

APPLYING BAYES' THEOREM TO SPAM CLASSIFICATION

$$P(\text{spam}|\text{words}) = \frac{P(w_1|\text{spam})P(w_2|w_1 \cap \text{spam})P(w_3|w_2 \cap w_1 \cap \text{spam}) \cdots P(\text{spam})}{P(w_1)P(w_2|w_1)P(w_3|w_2 \cap w_1) \cdots}$$

$$P(\text{spam}|\text{words}) = \frac{P(w_1|\text{spam})P(w_2|\text{spam})P(w_3|\text{spam}) \cdots P(\text{spam})}{P(w_1)P(w_2)P(w_3) \cdots}$$

$$P(w_2|w_1) = P(w_2)$$

NAÏVE BAYES

- **Advantages** of making this assumption of feature independence:
  - Easier to calculate probabilities. → model is faster
  - Empirically, our classifications are surprisingly accurate.
    ↳ b/c model performs well on test data, we're willing to make this assumption
- **Disadvantages** of making this assumption of feature independence:
  - It's incredibly unrealistic, especially in the case of text data.
  - While our classifications are accurate, our predicted probabilities are usually quite bad.

use .predict()

.predict_proba() is less reliable

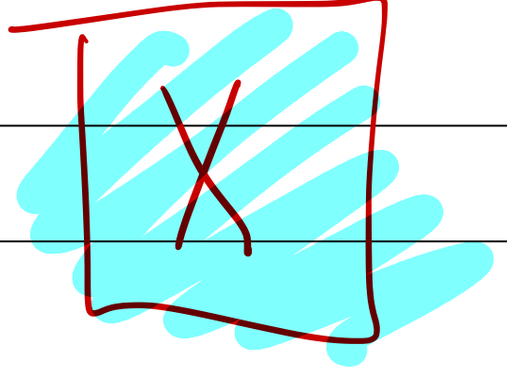
PROCESS OF NAÏVE BAYES

1. Decide which Naïve Bayes model to use.
    - BernoulliNB
    - MultinomialNB
    - GaussianNB

2. Decide what your priors will be.
    - Based on your data. *(default)*
    - Manually set.

3. `.fit()`, `.predict()`!

WHICH NAÏVE BAYES MODEL SHOULD I USE?



- BernoulliNB
  - ↳ if columns of X are 1/0, use Bernoulli NB
    - ↳ dummy variables
- MultinomialNB
  - ↳ if columns of X are integer counts, use Multinomial
    - ↳ CountVectorizer, Likert scale columns + TF-IDF Vectorizer
- GaussianNB
  - ↳ if columns of X are Normal
    - ↳ realistically — anything that isn't 1/0 or integer.

WHAT SHOULD MY PRIORS SHOULD BE?

Posterior

$$P(\text{spam}|\text{words in email}) = \frac{P(\text{words in email}|\text{spam})P(\text{spam})}{P(\text{words in email})}$$

Prior

- Estimated from data.

default

what you should do.

- Manually set.

only do this w/ subject-matter expertise

PROCESS OF NAÏVE BAYES

1. Decide which Naïve Bayes model to use.
    - BernoulliNB
    - MultinomialNB
    - GaussianNB

2. Decide what your priors will be.
    - Based on your data. *(default)*
    - Manually set.

3. `.fit()`, `.predict()`!

INTERVIEW QUESTION

- Suppose we want to detect whether Amazon reviews are spam or ham. How would you do this?

legit

1 0

1. Define problem. ✓
2. Gather data. .csv SQL API ← Google scrape* ask Amazon proxy Yelp eBay
3. Explore. missing outliers valid Y
4. Model

NB
- fast
- spam

Log Reg
- fast
- interpretable

(NNs CARTs)
↳ data?

KNN
↳ slow
↳ variance

5. Evaluate
6. Answer: minimize false positives $\frac{TN}{TN+FP}$ = specificity