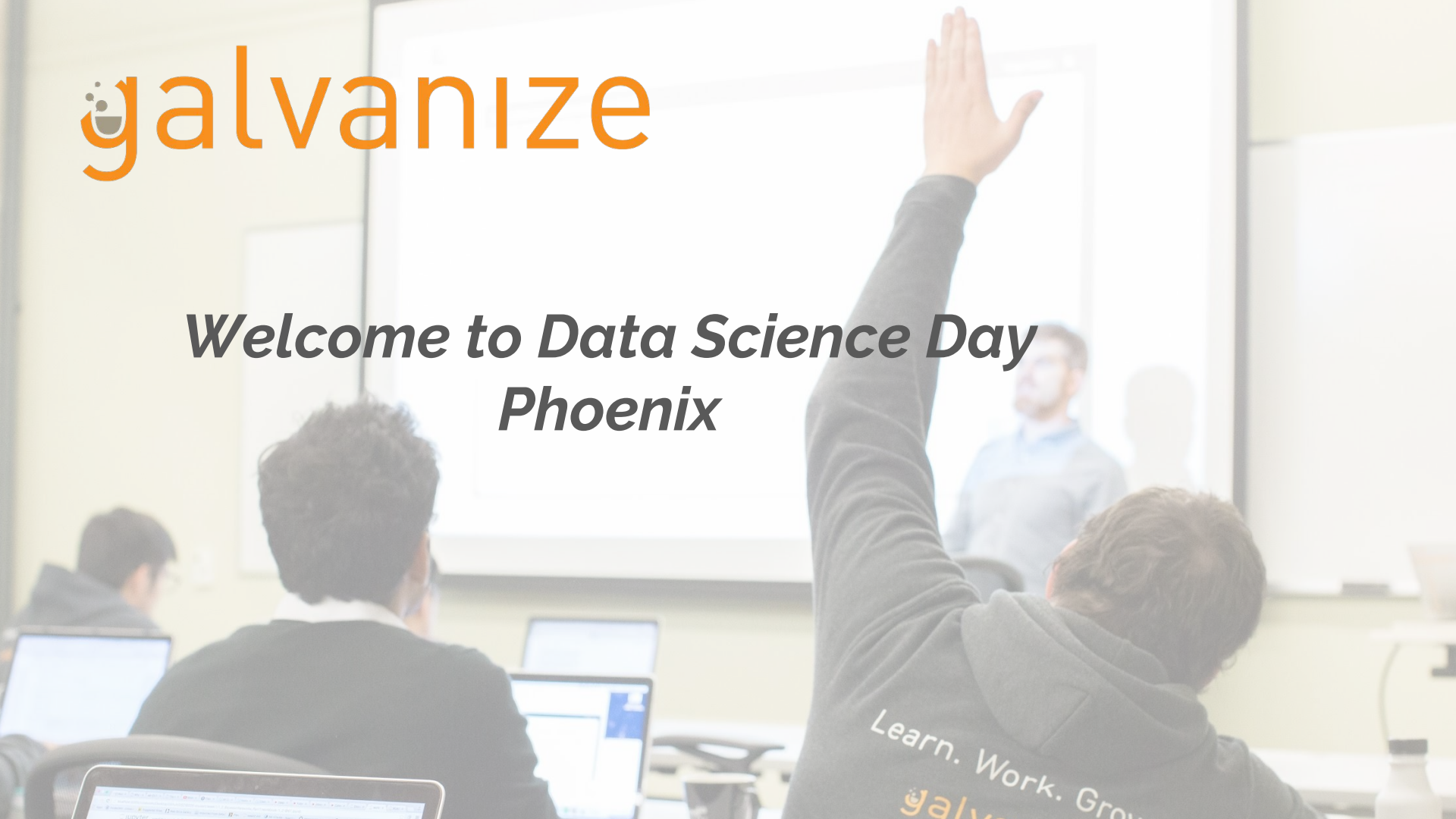
galvanize

Welcome to Data Science Day Phoenix



Naive Bayes Algorithm(s): A Sneak Peek

(Concentrated, Distilled, and generally all over the place)

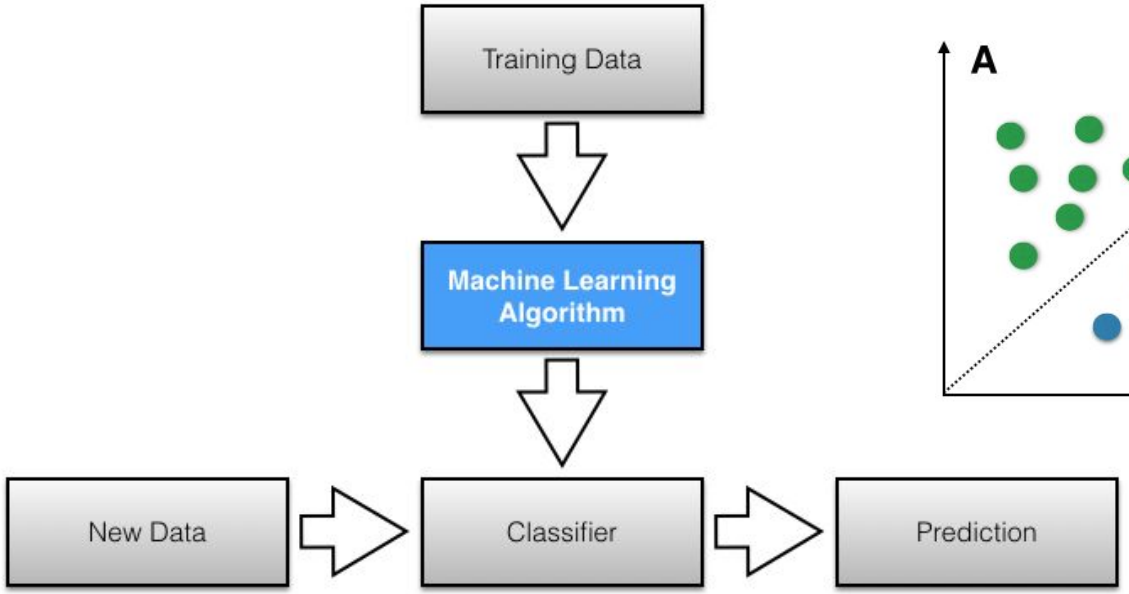
Aniket Majumdar
Sr. Data Scientist

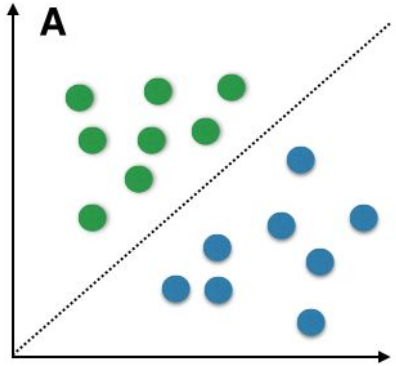
- Intro to Thomas Bayes
- The Classification Problem
- Naive Bayes Framework
- An implementation

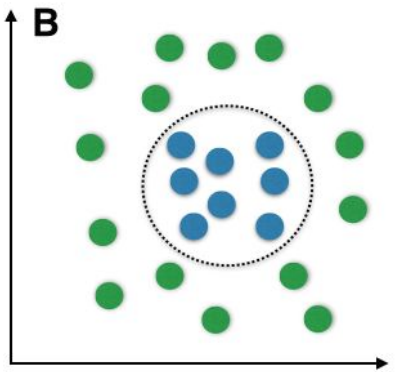
galvanize

https://www.britannica.com/biography/Thomas-Bayes

The CLASSIFICATION Problem







- Linear Classifier based on Bayes' Theorem
- Simple but Efficient -- Train and Predict quickly
- Found to outperform other classifiers, especially for small sample sizes
    - More features than observations
- Useful in online settings (continually receiving new data)
- Some successful use cases:
    - Diagnosis of diseases
    - Classification of RNA sequences in taxonomic studies
    - Spam filtering in e-mails

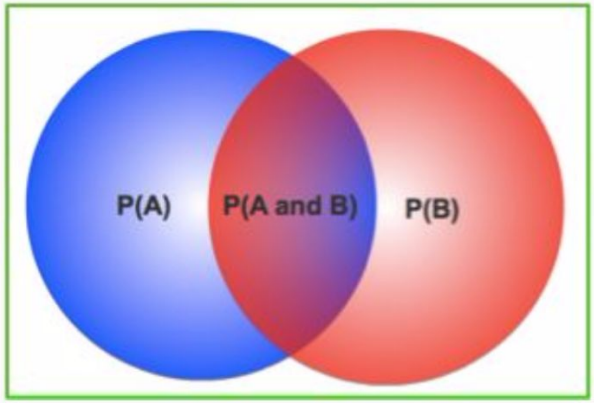
- Bayes' Theorem:
  - Example
- Derive Naive Bayes classifier
  - *Conditional* Independence
  - MAP Estimation
- Apply Naive Bayes to xoxo classification
- Understand nuances of Naive Bayes

- Bayes' Theorem:
  - Example -- 6% of the population have a certain type of disease. Suppose that 85% of the screening tests on the people with the disease show positive results, and 10% of the tests on the people without the disease show positive results (aka *false positives*). Given a person with a positive test result, what is the probability that this person actually has the disease?
  - A: 'Has disease' B: 'Test +'
  - $P(A) = 0.06,\ P(B|A) = 0.85,\ P(B \mid\sim A) = 0.1,\ P(A|B) = ?$

galvanize

Conditional Probability

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



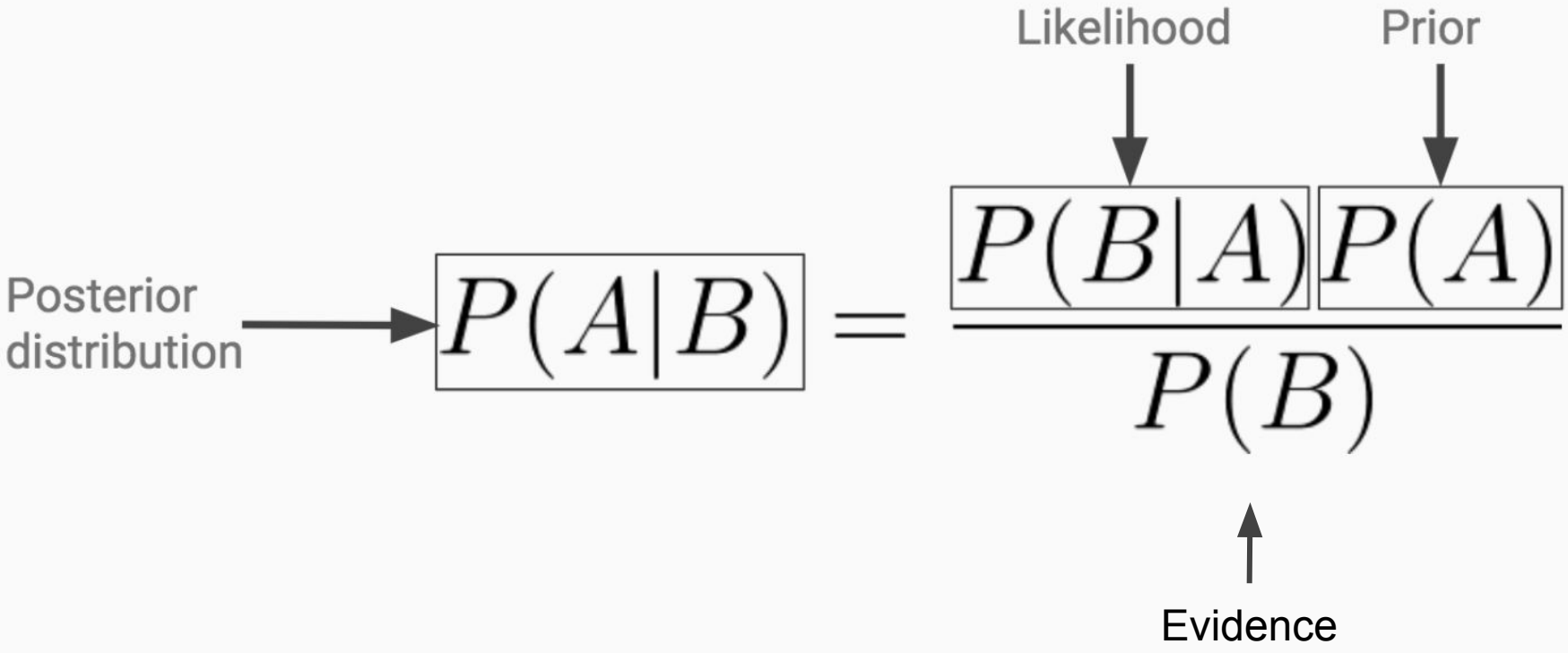
$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

Definition of conditional probability: $P(A|B) = \frac{P(A \cap B)}{P(B)}$

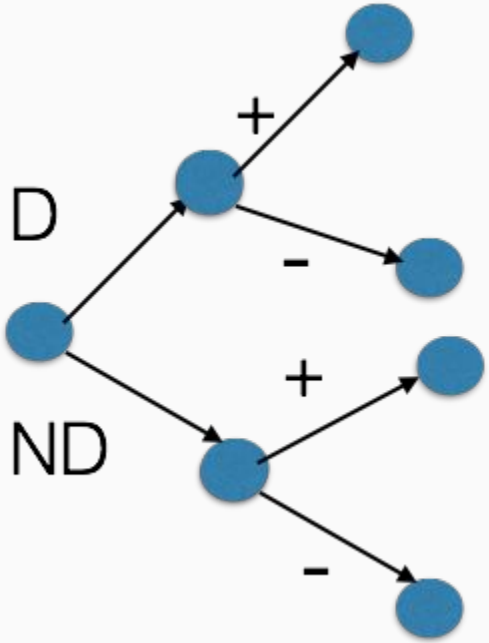
Property of joint probability: $P(B|A)P(A) = P(A \cap B)$

→ Bayes Theorem: $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$

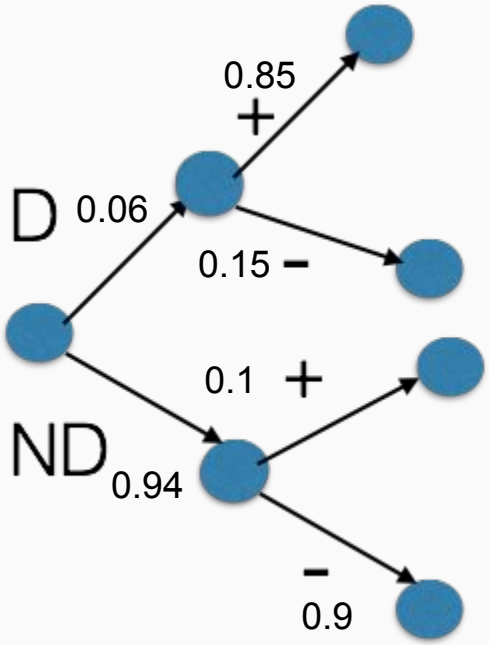
$$P(B) = P(B \text{ and } A) + P(B \text{ and } \sim A)$$


Likelihood
Prior
Posterior distribution
$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$
Evidence

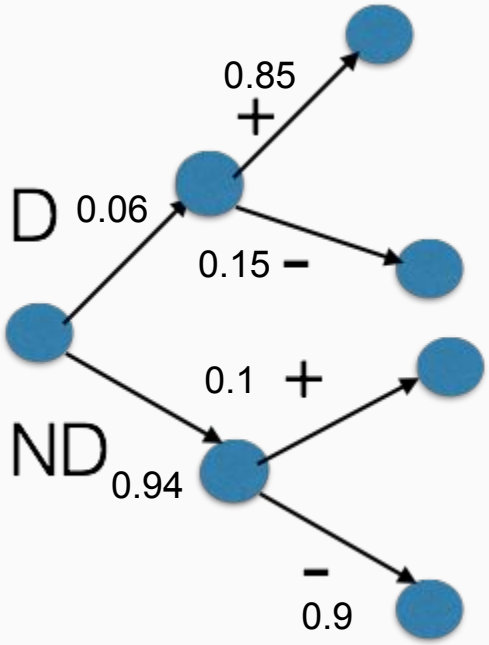
Review Example



- Looking for P(_|_) = ?



- Looking for P(D|+)



- Looking for P(D|+)
- P(D|+) = P(D & +) / P(+)
- P(+) = P(D & +) + P(ND & +)
- P(D|+) = 0.352

galvanize

Feature vector $X = (x_1, x_2, \ldots, x_n)$

Class variable $y = \{0,1\}$

$$P(y \mid x_1, \ldots, x_n) = \frac{P(y)P(x_1, \ldots x_n \mid y)}{P(x_1, \ldots, x_n)}$$

Bayes' Theorem

$$P(x_i|y, x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n) = P(x_i|y),$$

Conditional Independence

$$P(y \mid x_1, \ldots, x_n) = \frac{P(y)\prod_{i=1}^{n} P(x_i \mid y)}{P(x_1, \ldots, x_n)}$$

$x_1, x_2, \ldots, x_n$ *i.i.d.*

$$P(y \mid x_1, \ldots, x_n) \propto P(y)\prod_{i=1}^{n} P(x_i \mid y)$$

$$\Downarrow$$

$$\hat{y} = \arg\max_y P(y)\prod_{i=1}^{n} P(x_i \mid y),$$

Maximum A Priori Estimation (MAP)

P(y) = relative frequency of class y in Training data

Sci-kit Learn Python Library

- **Gaussian:** It is used in classification and it assumes that features follow a normal distribution.
- **Multinomial:** It is used for discrete counts. For example, let's say, we have a text classification problem. Here we can consider bernoulli trials which is one step further and instead of "word occurring in the document", we have "count how often word occurs in the document", you can think of it as "number of times outcome number x_i is observed over the *n* trials".
- **Bernoulli:** This model is useful if your feature vectors are binary (i.e. zeros and ones). One application would be text classification with 'bag of words' model where the 1s & 0s are "word occurs in the document" and "word does not occur in the document" respectively.

- If categorical variable has a category (in test data set), which was not observed in training data set, then model will assign a 0 (zero) probability and will be unable to make a prediction. This is often known as “Zero Frequency”. To solve this, we can use the smoothing technique. One of the simplest smoothing techniques is called *Laplace* estimation.
- On the other side naive Bayes is also known as a bad estimator, so the predicted probability outputs are not to be taken too seriously.
- Another limitation: the assumption of independent predictors. In real life, it is almost impossible that we get a set of predictors which are completely independent.

- Check Jupyter notebook

Differences Between Bayesians and Non-Bayesians
What is Fixed?

Frequentist:

- Data are a repeatable random sample - there is a frequency
- Underlying parameters remain constant during this repeatable process
- Parameters are fixed

Bayesian:

- Data are observed from the realized sample.
- Parameters are unknown and described probabilistically
- Data are fixed