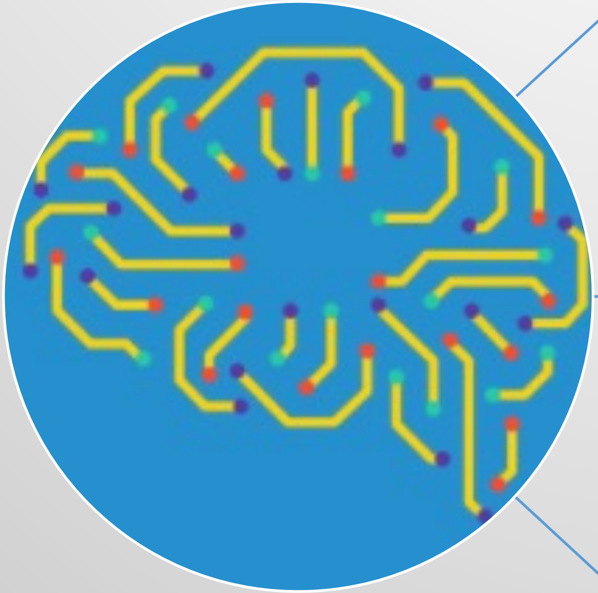
Natural Language Processing

Dr. Karen Mazidi

Part Five:
Machine Learning



Topics

- ML basics
- Important probability distributions

Quizzes

- Q: Python ML

Homework

- Portfolio: Reading ACL Papers

Machine Learning

- NLP techniques:
- Rules-based
- Statistical
- Traditional ML
- Deep Learning

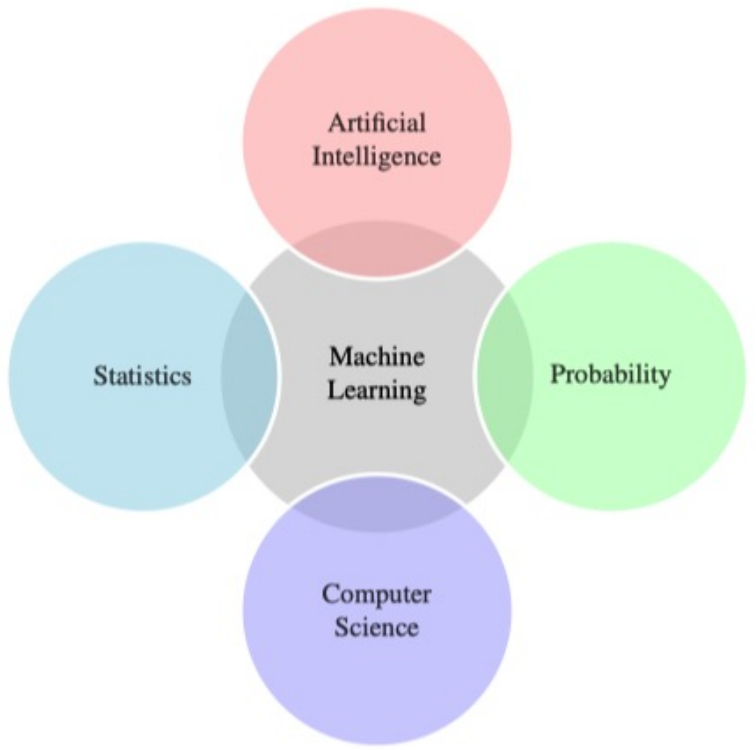
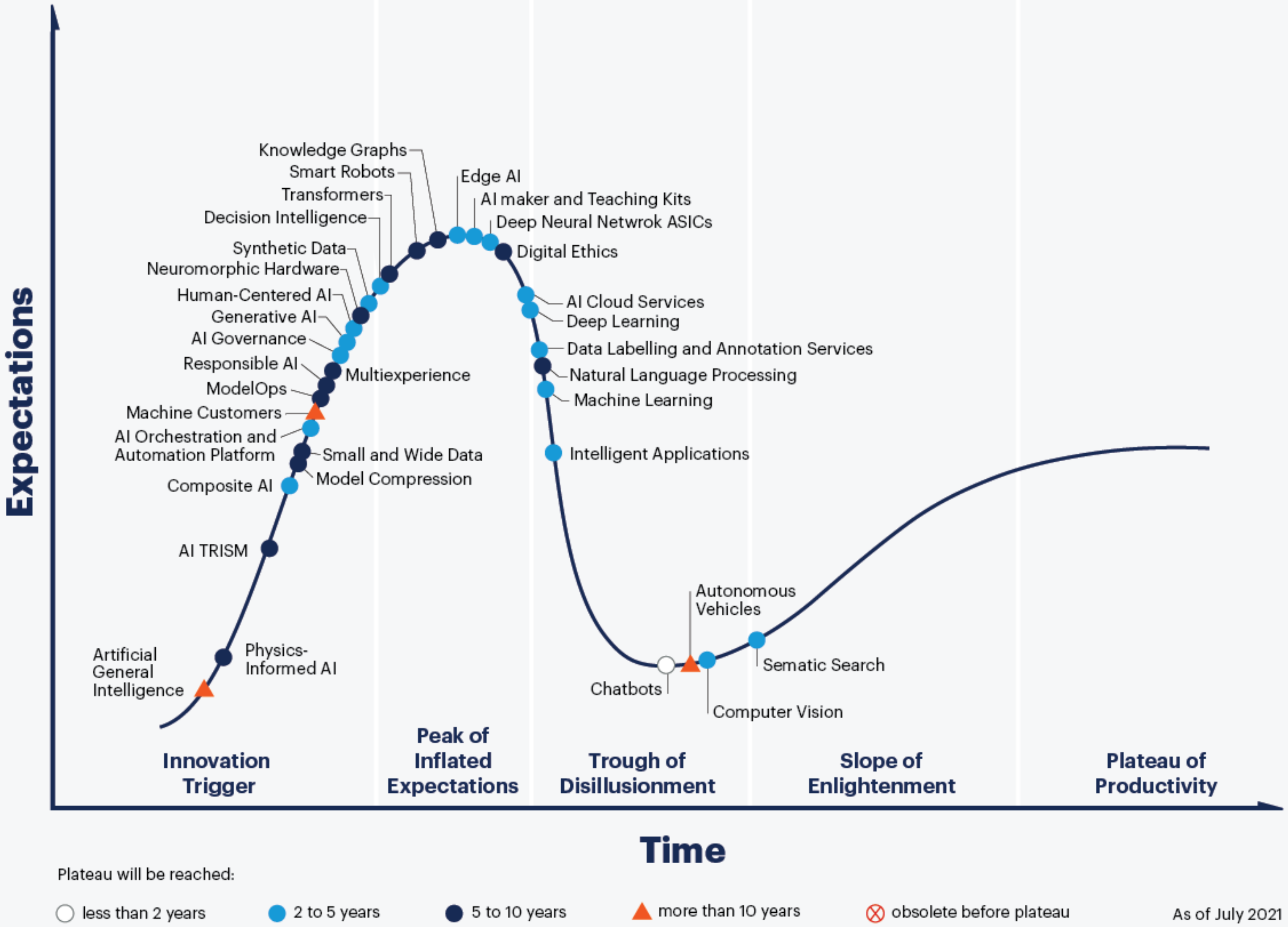


Figure 17.1: Fields Related to Machine Learning

Machine learning

- Machine Learning trains computers to accurately recognize patterns in data for purposes of:
  - data analysis, ex: sentiment analysis
  - prediction: classification
  - Action selection by autonomous agents: Siri

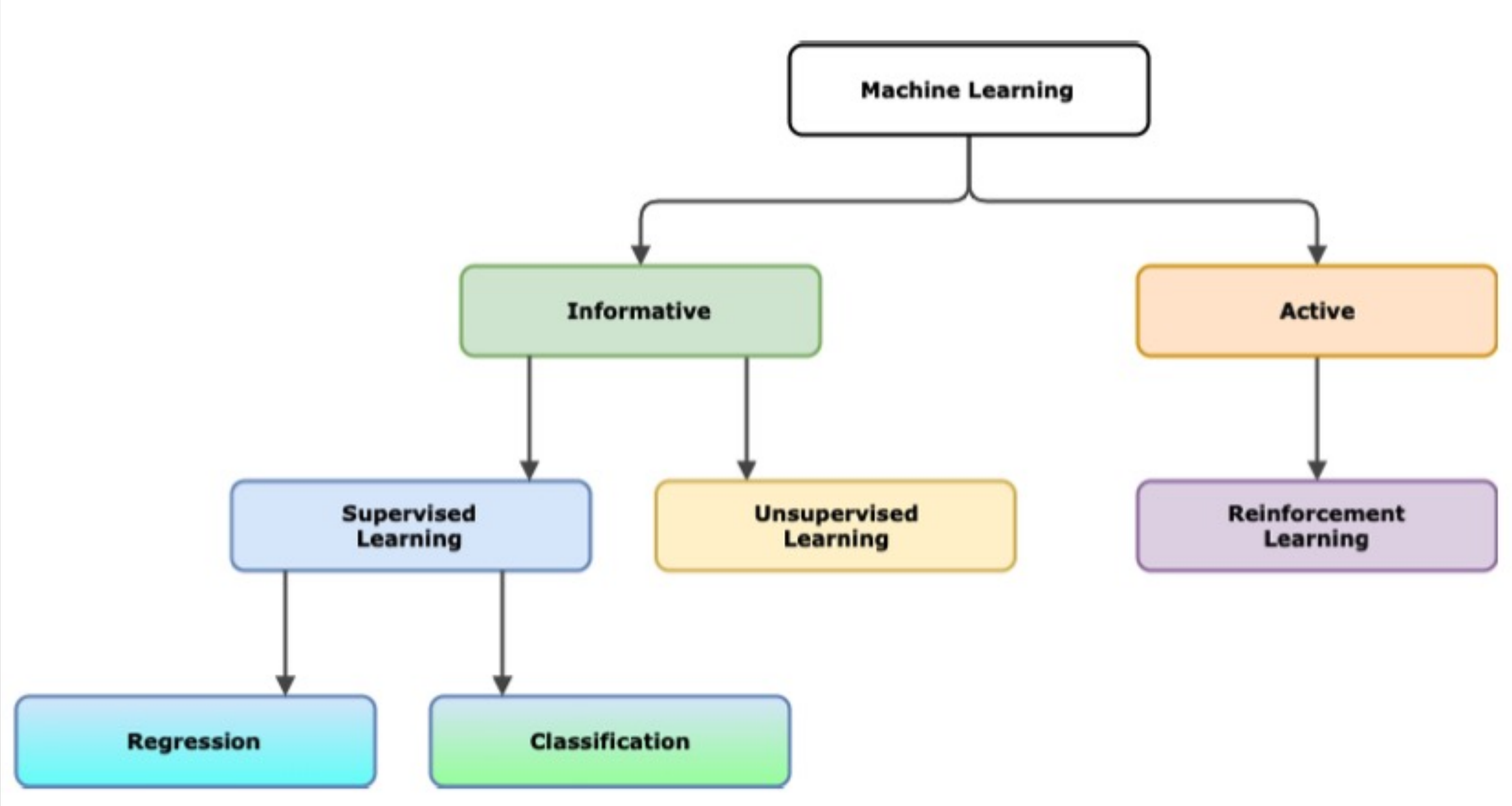
Hype Cycle for Artificial Intelligence, 2021



gartner.com

Source: Gartner

Machine learning scenarios



Terminology

GPA	Hours	SAT	Class
3.2	15	1450	Junior
3.8	21	1420	Sophomore
2.5	9	1367	Freshman

- Rows: example, instance, observation
- Columns: feature, attribute
- Supervised learning: predictor, target
- Data:
- Quantitative real numbers
- Qualitative, categorical data (aka factors)

Probability in NLP data

- Documents consist of words
- Words are random variables in documents
- Classification:
  - Example: P(sarcasm | really)

Probability distributions

- Most important for NLP:
  - Uniform
  - Binomial and Beta
  - Multinomial and Dirichlet
  - Gaussian

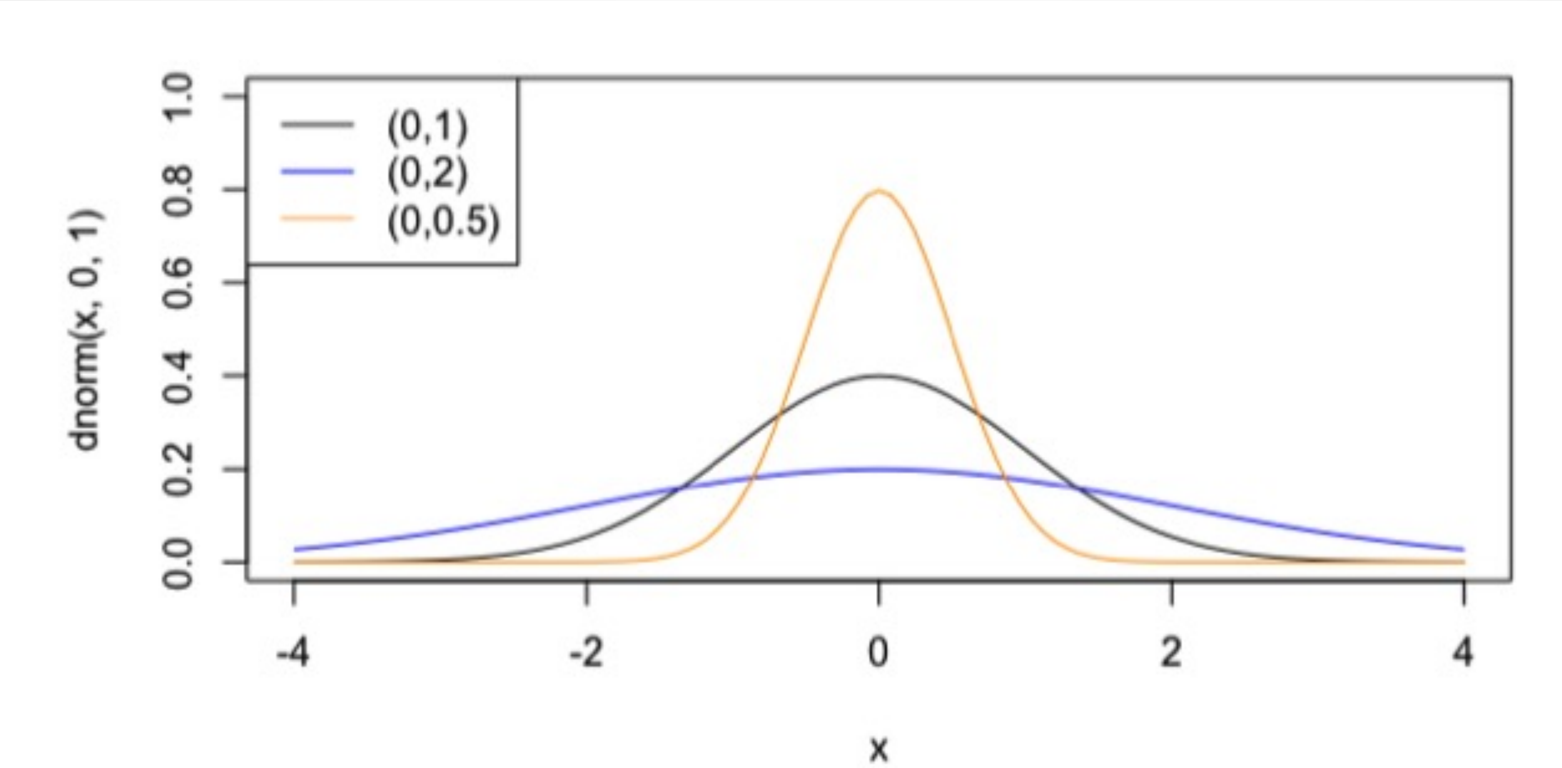
Gaussian

- Normal distribution for quantitative variables
- Defined by mean (mu) and variance (sigma squared)

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

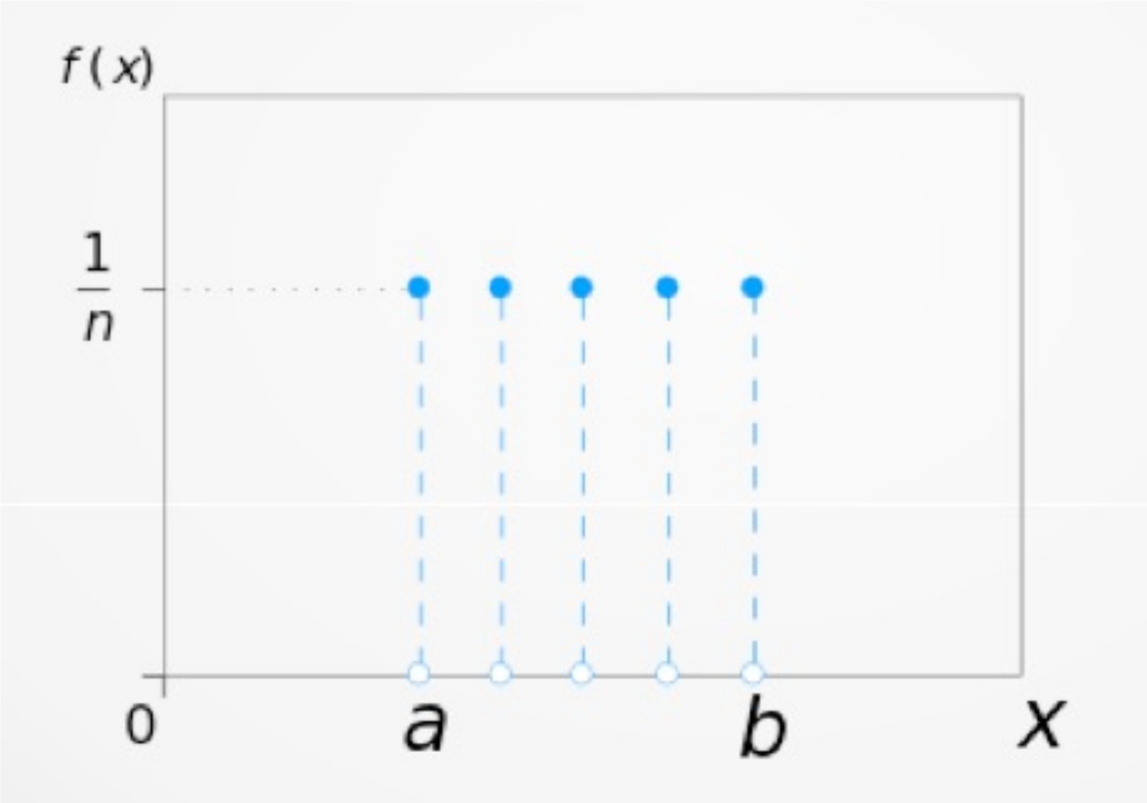
Gaussian

- Same means, different variance



Uniform distribution

- Sometimes used as a prior distribution
- Every word has an equal prior distribution



Binomial and Beta distributions

- Binary variables:
  - Sarcastic or not
  - Subjective or objective
  - Word present or not

Bernoulli distribution

- Parameter mu is the expected value

$$Bernoulli(x|\mu) = \mu^x(1-\mu)^{1-x}$$

- Example:
  - p(sarcasm) = 0.2
  - 0.2^1 * 0.8^0

Binomial distribution

- The sum of outcomes of multiple Bernoulli events
- N is number of trials, k is number in positive class
- Each trial is independent; each has two outcomes 0, 1
- 100 word vocabulary, P(word in document) = 0.2
- What is the chance that word x will be in document?

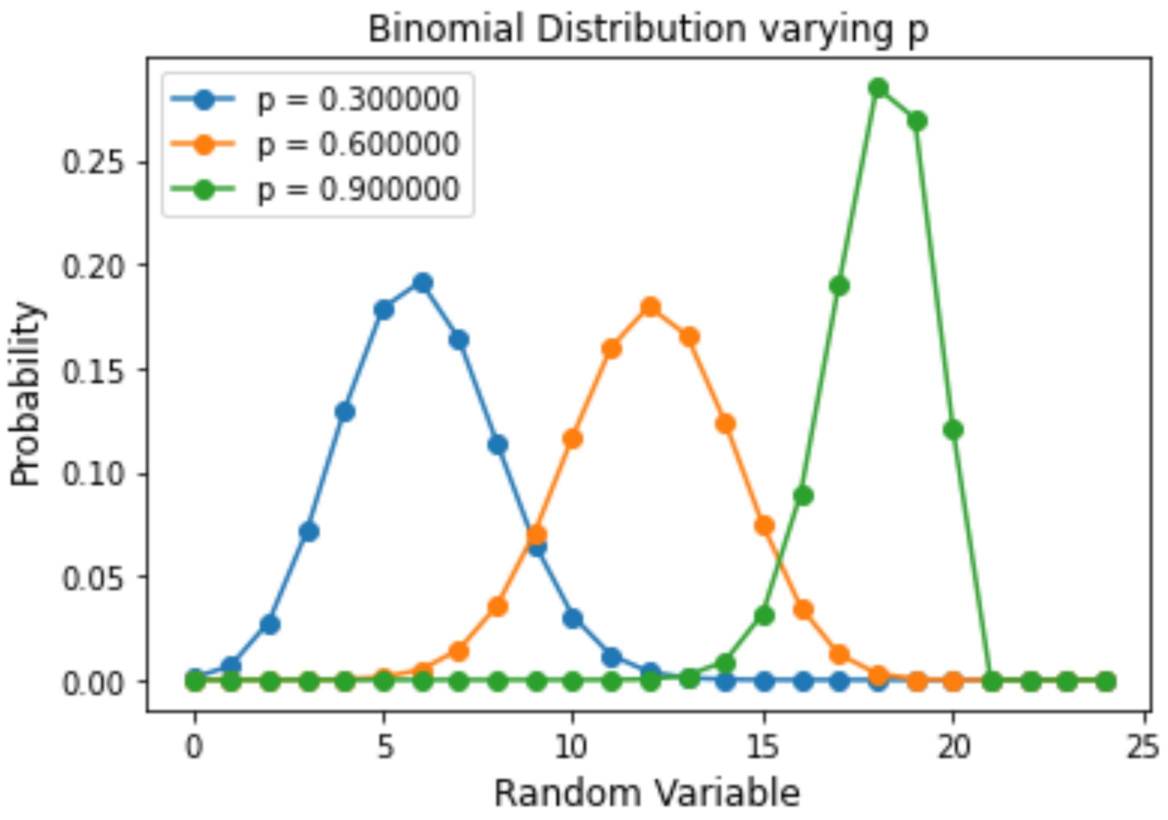
$$Binomial(k|N,\mu) = \binom{N}{k} \mu^k (1-\mu)^{N-k}$$

Let's let k=20 for our 100 trials. Will the outcome of the binomial be 0.2?

$$Binomial(20|100, 0.2) = \binom{100}{20} 0.2^{20} (1-0.2)^{80} = 0.09930021$$

Binomial distribution

- Shape controlled by p (mu)

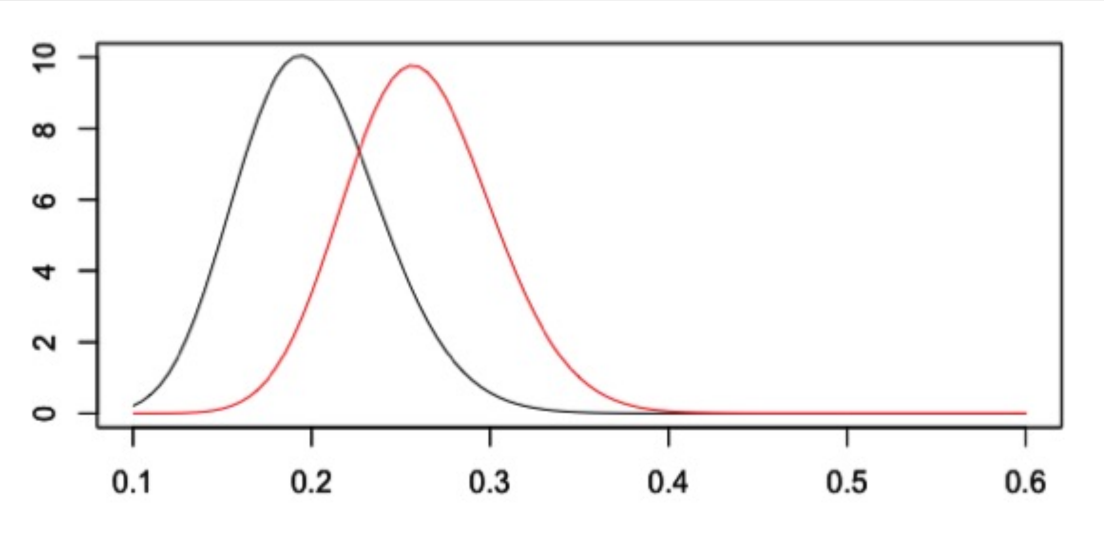


Beta distribution

- Beta is a distribution over binomials, the conjugate prior
- Gamma term is a constant ensuring integration to 1

$$Beta(\mu|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1}$$

- Beta distribution for mu=0.2 and update a, b



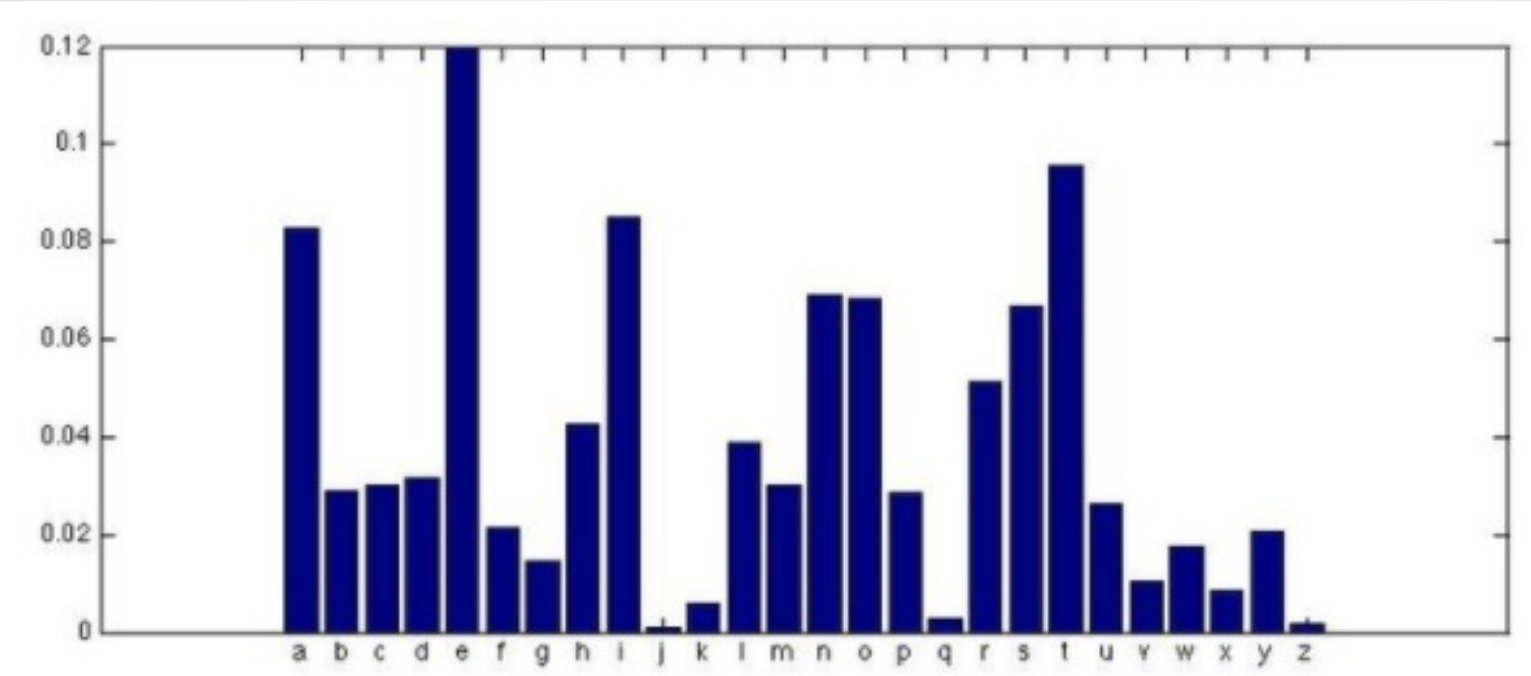
Multinomial distribution

- Categorical data with more than 2 classes, example: positive, negative, neutral sentiment
- N number of examples
- K number of classes
- The ms are the probability of each class

$$Multinomial(m_1, m_2, ..., m_k | N, \mu) = \left(\frac{N}{m_1!m_2!...m_k!}\right)\prod_{k=1}^{K}\mu_k^{m_k}$$

Multinomial distribution

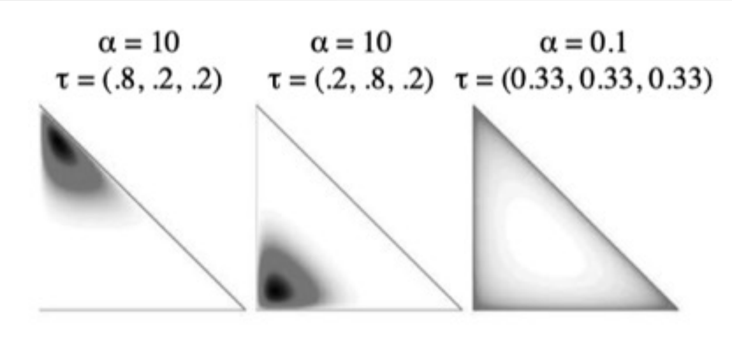
- Example: letters in a text



Dirichlet distribution

$$Dir(\mu|\alpha) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)...\Gamma(\alpha_k)} \prod_{k=1}^{K} \mu_k^{\alpha_k-1}$$

- Prior for a multinomial distribution
- Has k alpha parameters, one for each class
- Alpha_0 is the sum of all alphas



- Base measure, tau, is the expected value
- Smaller the alpha, the closer samples are to tau

Probability distributions in NLP

Text as a bag of words

Alice was beginning to get very tired of sitting by her sister on the bank, and of having nothing to do: once or twice she had peeped into the book her sister was reading, but it had no pictures or conversations in it, 'and what is the use of a book,' thought Alice 'without pictures or conversation?'

$P(\texttt{of}) = 3/66$ $P(\texttt{to}) = 2/66$ $P(\texttt{,}) = 4/66$

$P(\texttt{Alice}) = 2/66$ $P(\texttt{her}) = 2/66$ $P(\texttt{'}) = 4/66$

$P(\texttt{was}) = 2/66$ $P(\texttt{sister}) = 2/66$

Looking ahead

Traditional ML models commonly used for text data:

- Naive Bayes
- Logistic Regression
- Neural Networks

Essential points to note

- Recent advances in NLP have been driven largely by ML approaches
- Traditional ML algorithms used often in NLP:
    - Naïve Bayes
    - Logistic Regression
    - SVM
    - Neural Networks
- Deep learning is just a deep and large neural network
    - data hungry

To Do

- Quiz on ML Basics