Naïve Bayesians

Back to Basics Series

20 Feb 2021

Goal

Developing the Bayesian muscle to solve a wide range of problems

Naïve Bayesian Philosophy

Intuitive (Visual) Understanding of the Bayesian Reasoning

Ability to model real world problems in a Bayesian Setting

Starting from Simple Probabilistic modelling

Adapting it in a a Bayesian setting
And moving towards ML models

PyMC3

Fluency in the Calculus of Bayesian Stats & ML model

Season 2: Back to Basics

- Ep 1 — Bayes Theorem
- Ep 2, Ep 3 — Problems with Binomial Likelihoods
- Ep 4 — Disease Detection
- Ep 5 — Naive Bayes Classification
- Ep 6 — Gaussian Naive Bayes Classification
- Ep 7 — German Tank Problem
- Ep 8 — Waiting Times (Continuous Distributions)

Back to Basics

		Canonical Problem	Applications
Ep 1	Bayes Theorem	There are 2 boxes from which cookies can be taken from. Box A and Box B. Box A contains 10 chocolate cookies, Box B contains 5 ginger cookies. Given that you get a chocolate cookie which box was it taken from?	The Shy Librarian Problem Naive Bayes algorithm
Ep 2 Ep 3	Problems with Binomial Likelihoods	You have 2 coins C1 and C2. p(heads for C1) = .7 & P(heads for C2) = 0.6 You flip the coin 10 times. What is the probability that the given coin you picked is C1 given you have 7 heads and 3 tails?	A/B Testing
Ep 4	Disease Detection	A particular disease affects 1% of the population. There is an imperfect test for this disease: The test gives a positive result for 90% of people who have the disease, and 5% of the people who are disease-free. Given a positive test result – what is the probability of having the disease?	COVID Tests (PCR & Antibody)! Fraud Detection
Ep 5	Naive Bayes Classification	Given these words occur in this text what's the probability it's spam?	Any Classification Problem
Ep 6	Gaussian Naive Bayes Classification	Given the weights and heights of basketball players, what's the probability that person a is a basketball player given weight = w and height = h?	

S

		Canonical Problem	Applications
Ep 7	German Tank Problem	Suppose tanks were given a serial number based on the order in which they were manufactured. Given that you've observed a tank with serial number "10", how many tanks were actually manufactured in total?	?
Ep 8	Waiting Times (Continuous Distributions)	Time between receiving a spam email, t_i was recorded on a random day. How long do you have to wait to get your next spam email?	Planning Trials Estimating Queues

Bayes Rule

Posterior Likelihood Prior

$$P(\theta \mid D) = \frac{P(D \mid \theta)\, P(\theta)}{P(D)}$$

Normalising Constant

Bayes Rule

Posterior Likelihood Prior

$$P(\theta_i \mid D) = \frac{P(D \mid \theta_i)\, P(\theta_i)}{\sum_{all\ j} P(D \mid \theta_j)\, P(\theta_j)}$$

Normalising Constant

Canonical Problem Simplified

Time between receiving a spam email, t_i was recorded on a random day. How long do you have to wait to get your next spam email?

9.9
12
5.1
12
4.5
8.2
9.3
11

mins

Canonical Problem Simplified

Time between receiving a spam email, t_i was recorded on a random day.
How long do you have to wait to get your next spam email?

9.9
12
5.1
12
4.5
8.2
9.3
11

mins

$$\sum t_i = 72 \text{ mins}$$

$$n = 8$$

How long do you have to wait to get your next spam email?

Before

$\sum t_i = 72$ mins, $n = 8$

Expected value:

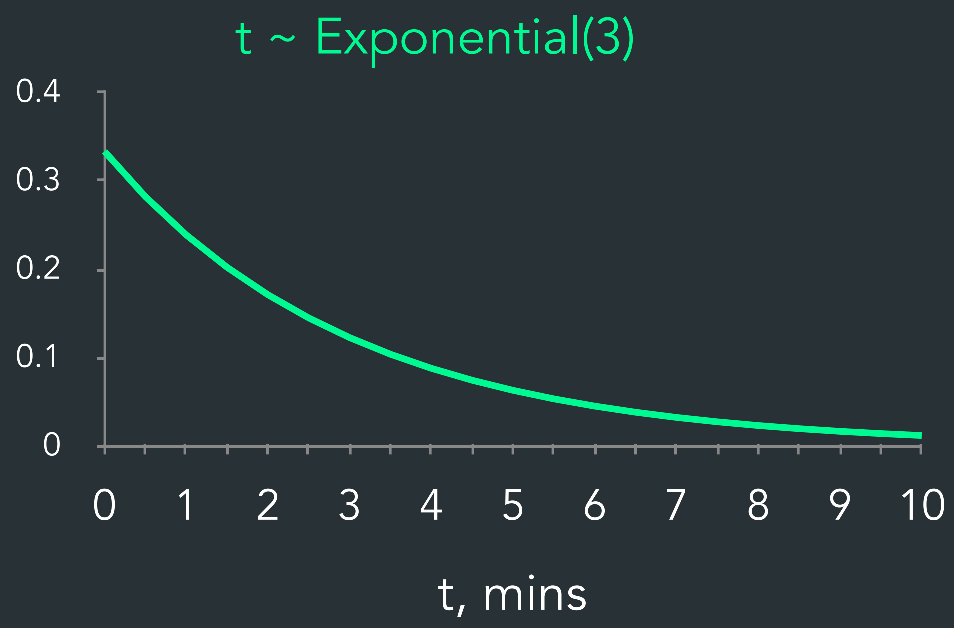
$72/8 = 9$ mins

- How confident are you?
- What if you had more data?
- What if you knew that it was more likely for the data to tampered?

Exponential Distribution

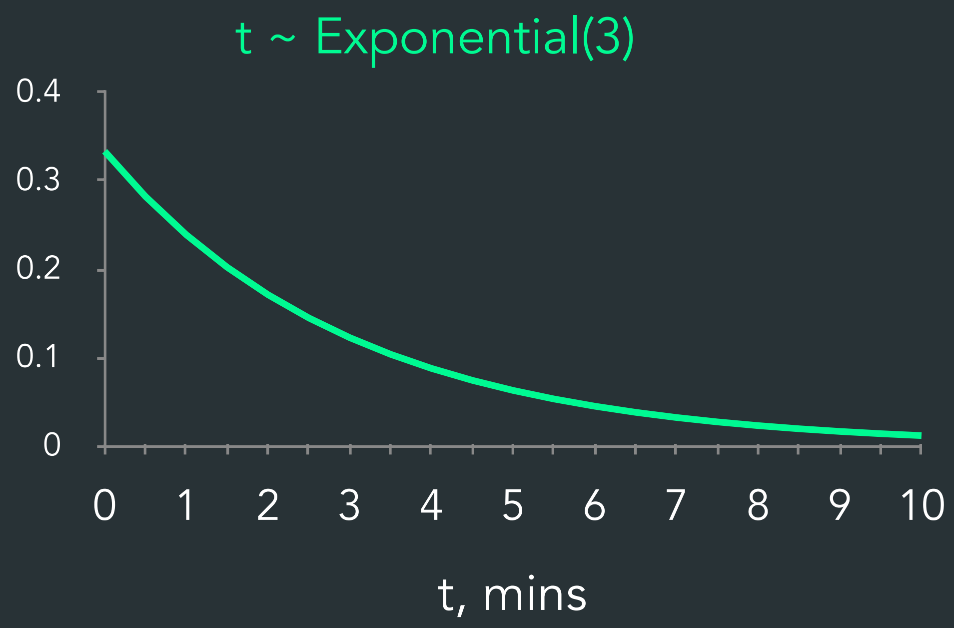
- t ~ Exponential($1/\mu$)
  - $t > 0$
  - $\mu > 0$

$$P(t \mid \mu) = (1/\mu)\, e^{(t/\mu)}$$



Exponential Distribution

- t ~ Exponential($1/\mu$)
  - $t > 0$
  - $\mu > 0$

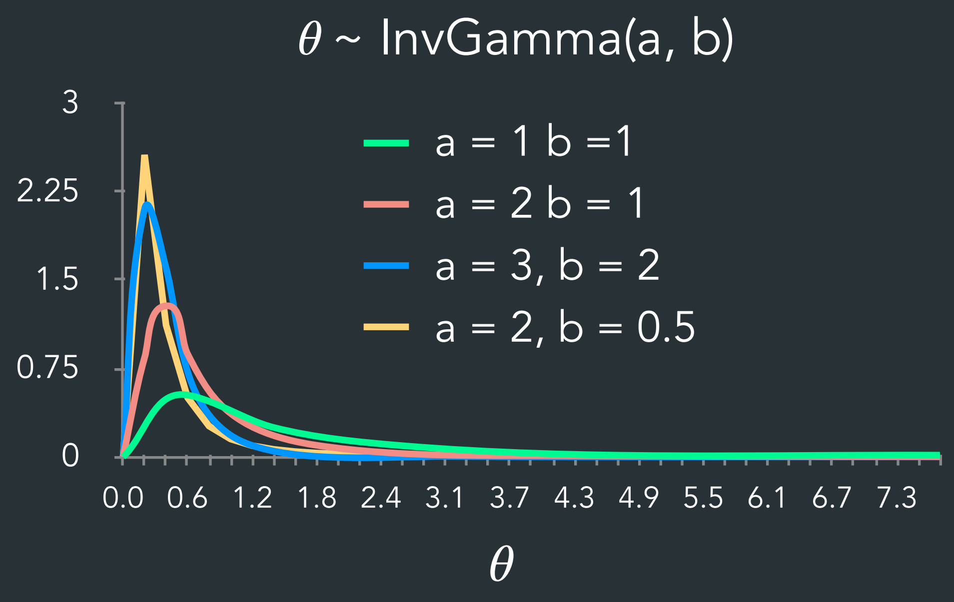


$$P(t \mid \mu) = (1/\mu)\, e^{(t/\mu)}$$

$P(t \mid \lambda) = \lambda e(\lambda t)$ Alternative Formulation

Inverse Gamma Distribution

- $\theta \sim \text{InvGamma}(a, b)$
  - $\theta > 0$
  - $a > 0, b > 0$



$$P(\theta) = c\ (1/\theta)^{a+1} e^{-b/x}$$

$$E(\theta) = \frac{b}{a - 1} \qquad \text{Mode} = \frac{b}{a + 1}$$

Bayes Rule

Posterior Likelihood Prior

$$P(\theta \mid D) = \frac{P(D \mid \theta)\, P(\theta)}{P(D)}$$

Normalising Constant

Conjugate Prior

Same Family

Posterior

Prior

$$P(\theta \mid D) = \frac{P(D \mid \theta)\, P(\theta)}{P(D)}$$

What is the probability of getting a head?

1. Let μ the average waiting time between 2 emails
   - Assume μ follows a InvGamma(2, 2) distribution

2. Let $t_1 \ldots t_N$ be the intervals between spam emails
   - Assume t are independent and identically distributed
   - $t_i \sim \text{Exponential}(1/\mu)$ for all i

$$P(\mu \mid t_1 \ldots t_N) = \frac{P(t_1 \ldots t_N \mid \mu)\, P(\mu)}{P(t_1 \ldots t_N)}$$

Conjugate prior proof

$$P(\mu \mid t_{1 \ldots} t_N) \propto P(t_{1 \ldots} t_N \mid \mu)\ P(\mu)$$

$$\propto (1/\mu)^N e^{-(\sum ti/\mu)} \quad (1/\mu)^{a+1} e^{-(b/\mu)}$$

$$\propto (1/\mu)^{a+N+1}\ e^{-(b+\sum ti/\mu)}$$

$$= (1/\mu)^{a'+1} e^{-b'/\mu}$$

$$= \mathrm{InvGamma}(a', b')$$

$$= \mathrm{InvGamma}(a+n, b+\sum t_i)$$

$$E(\mu) = \frac{b+\sum t_i}{a+n-1}$$

How long do you have to wait to get your next spam email?

Before

$\sum t_i = 72$ mins, $n = 8$

Expected value

- 72/8 = 9 mins

- How confident are you?
- What if you had more data?
- What if you knew that it was more likely for the data to tampered?

After

- Start with InvGamma(2, 2)
- Update prior belief with InvGamma(2+8, 2+72)
- Expected value: 74/(10-1) = 8.2