Bayes Theorem & Science of Uncertainty

Mathematics for Machine Learning

Lecturer: Matthew Wicker

Material Covered

Models: Linear models, basis expansion, logistic regression, neural networks, Prob. densities

Techniques: Least squares estimation, forward AD, reverse AD, computational graphs, gradient descent, convergence, convexity, Lipschitz continuity, Maximum likelihood, maximum a posteriori, LOTUS, change of variables, expectation identities

Settings: Regression, Classification, Density Estimation

This lecture: Bayes theorem, Bayesian terminology, equating coefficients, kinds of uncertainty

Everyday probabilistic reasoning

Your young cousin is enthusiastic about birds and wants you to help them identify a cool bird they have just seen. They describe it as being white with an orange/red-ish beak

Everyday probabilistic reasoning

Your young cousin is enthusiastic about birds and wants you to help them identify a cool bird they have just seen. They describe it as being white with an orange/red-ish beak

P(Gull | Obs.)	P(Pigeon|Obs.)

Everyday probabilistic reasoning

Your young cousin is enthusiastic about birds and wants you to help them identify a cool bird they have just seen. They describe it as being white with an orange/red-ish beak

$P(\text{Gull} \mid \text{Obs.})$	$P(\text{Pigeon} \mid \text{Obs.})$
0.7	0.3

Everyday probabilistic reasoning

Your young cousin is enthusiastic about birds and wants you to help them identify a cool bird they have just seen. They describe it as being white with an orange/red-ish beak

In reality it is much closer to being flipped!

$P(\text{Gull} \mid \text{Obs.})$	$P(\text{Pigeon} \mid \text{Obs.})$
~~0.7~~ 0.3	~~0.3~~ 0.7

How can we compute these probabilities?

$P(\text{Gull} \mid \text{Obs.})$	$P(\text{Pigeon} \mid \text{Obs.})$
~~0.7~~ 0.3	~~0.3~~ 0.7

$$p(\text{Gull} \mid \text{white, orange})$$

Event (A)

Observation (B)

How can we compute these probabilities?

P(Gull | Obs.)	P(Pigeon|Obs.)
~~0.7~~ 0.3	~~0.3~~ 0.7

$$P(\text{Pigeon}|\text{white, orange})$$

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

Bayes theorem

How can we compute these probabilities?

$$P(\text{Pigeon}|\text{white, orange})$$

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

Bayes theorem

Names for Bayesian Probabilities

Posterior distribution

Likelihood

Prior distribution

$$P(\underbrace{\text{Pigeon}}_{\text{event/data}}|\underbrace{\text{white, orange}}_{\text{observation}}) = \frac{P(\text{white, orange}|\text{Pigeon})P(\text{Pigeon})}{P(\text{white, orange})}$$

Model Evidence/Marginal Likelihood

Names for Bayesian Probabilities

Posterior distribution

Likelihood

Prior distribution

$$P(\text{Pigeon}|\text{white, orange}) \propto P(\text{white, orange}|\text{Pigeon})P(\text{Pigeon})$$

Prior: evidence should not determine belief in a vacuum

Posterior distribution

Likelihood

Prior distribution

$$P(\text{Pigeon}|\text{white, orange}) \propto P(\text{white, orange}|\text{Pigeon})P(\text{Pigeon})$$

The prior distribution in this equation serves the critical role of reminding us that data should not fully determine belief

How can we incorporate this prior knowledge?

$$P(\text{Pigeon})$$

If I said I saw a bird (absent any other data) what is the probability that it is a pigeon?

Population of Pigeons: 3,000,000 (0.97)
Population of Gulls: 100,000 (0.03)

How can we incorporate this prior knowledge?

Likelihood (0.01)

0.97

$$P(\text{Pigeon}|\text{white, orange}) \propto P(\text{white, orange}|\text{Pigeon})P(\text{Pigeon})$$

$$P(\text{Gull}|\text{white, orange}) \propto P(\text{white, orange}|\text{Gull})P(\text{Gull})$$

0.90 (black-backed gull 0.1)

0.03

How can we incorporate this prior knowledge?

Likelihood (0.01)

0.97

0.097

$$P(\text{Pigeon}|\text{white, orange}) \propto P(\text{white, orange}|\text{Pigeon})P(\text{Pigeon})$$

$$P(\text{Gull}|\text{white, orange}) \propto P(\text{white, orange}|\text{Gull})P(\text{Gull})$$

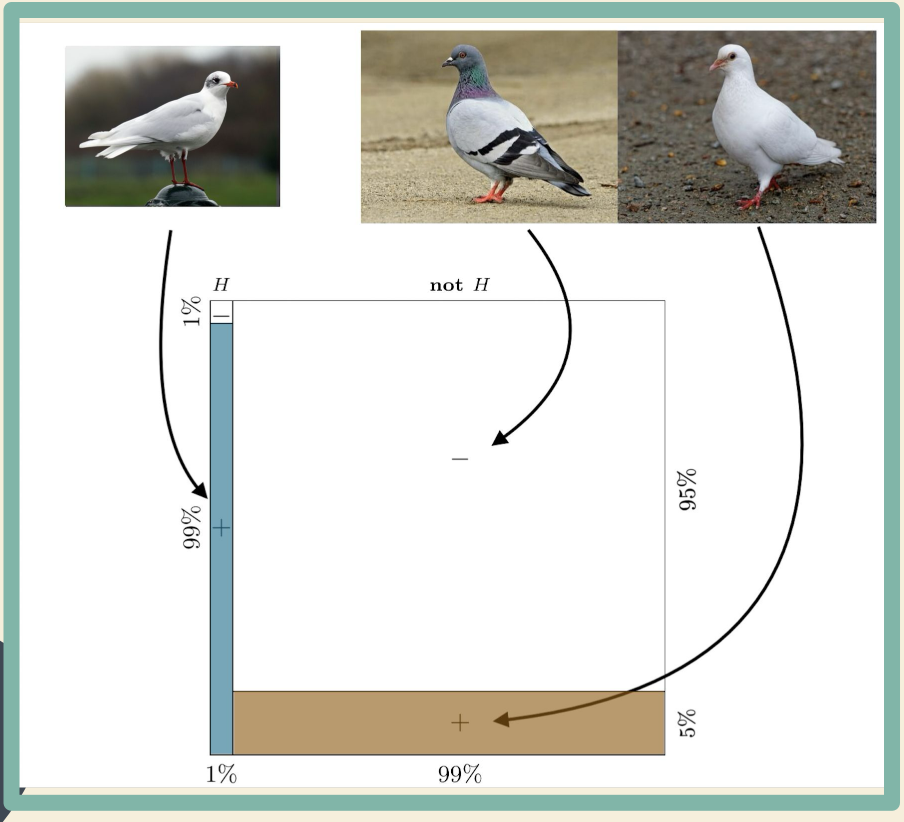
0.027

0.90 (black-backed gull 0.1)

0.03

So it is 3 times more likely according to Bayes that the bird was a pigeon!

Geometric intuition for Bayes



Moving to density estimation: Gaussian density estimation

Voltage we observe

Source voltage sent

Observational noise

$$V = S + \epsilon$$

We motivate Bayesian density estimation with an analog communication where we are trying to decode a source voltage over a noisy channel.

Moving to density estimation: Gaussian density estimation

Find $p(S|V) = \frac{p(V|S)p(S)}{p(V)}$

Prior

$$p(s) = \mathcal{N}(s; 0, 1)$$

We assume knowledge about the signal we are going to receive (e.g., that they are sending us a zero)

$$V = S + \epsilon$$

We motivate Bayesian density estimation with an analog communication where we are trying to decode a source voltage over a noisy channel.

Moving to density estimation: Gaussian density estimation

prior dist.: $$p(s) = \mathcal{N}(s; 0, 1)$$

prob of obs. given what intention was

likelihood: $$p(v|s) = \mathcal{N}(v; s, \sigma^2)$$

Now, we model the likelihood with the observational noise, and we can compute the likelihood of observing a value on the other end of the wire

$$V = S + \epsilon$$

We motivate Bayesian density estimation with an analog communication where we are trying to decode a source voltage over a noisy channel.

Gaussian density estimation

$$\sigma = 0.63$$

$$v_0 = 0.25$$

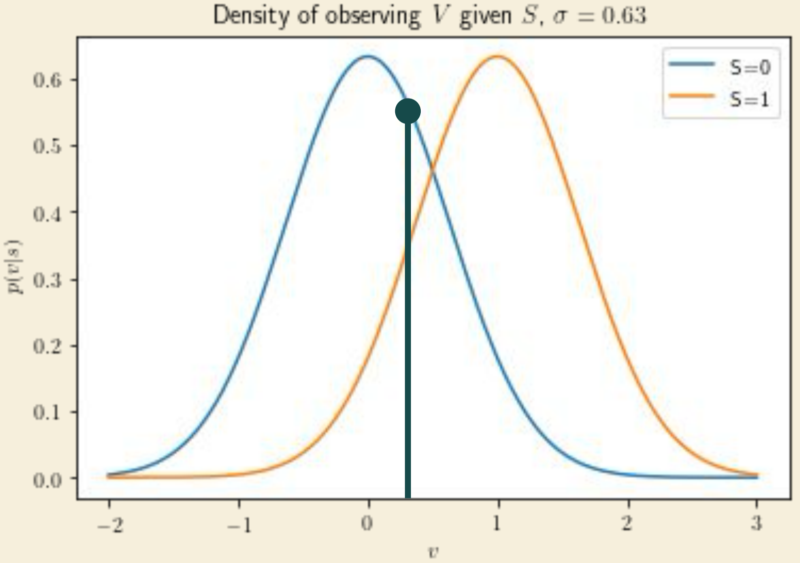
If 0.25 is our observation, but is an extremely noisy channel, how sure are we that the sender intended a 0 or a 1?

Gaussian density estimation

$$\sigma = 0.63$$

$$v_0 = 0.25$$

p(v|s) →



Properties of Gaussians

We have talked about some properties of Gaussians in prior lectures:

$$X \sim \mathcal{N}(\mu_X, \sigma_X^2)$$

$$Y = X + c$$

$$Y \sim \mathcal{N}(\mu_X + c, \sigma_X^2)$$

$Var(X) = \sigma^2_X$

$Var(Y) = Var(X + c)$

$= E[(X + c - E(X + c))^2]$

$= E[(X + c - E(X) - E(c))^2]$

$= E[(X + c - E(X) - c)^2]$

$= E[(X - E(X))^2]$

$= \sigma^2_X$

By linearity of expectation

Properties of Gaussians

We have talked about some properties of Gaussians in prior lectures:

$$X \sim \mathcal{N}(\mu_X, \sigma_X^2)$$

X, Y need to be indep.

$$Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$$

$$Z = X + Y$$

add μ & σ^2

$$Z \sim \mathcal{N}(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$$

Sum of Gaussians is Gaussian

Properties of Gaussians

We have talked about some properties of Gaussians in prior lectures:

$$X \sim \mathcal{N}(\mu_X, \sigma_X^2)$$

$$Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$$

$$Z \sim XY$$

Product of Gaussians, not Gaussian!

Properties of Gaussians

We have talked about some properties of Gaussians in prior lectures:

$$(X, Y) \sim \mathcal{N}(\mu_X, \sigma_X^2)\mathcal{N}(\mu_Y, \sigma_Y^2)$$

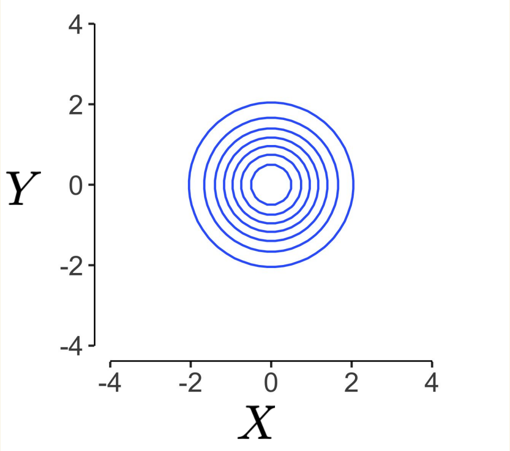
The product of (independent) Gaussian densities, is!

Product of Gaussian RVs not Gaussian
Product of Gaussian densities is Gaussian

Properties of Gaussians

We have talked about some properties of Gaussians in prior lectures:

$$(X, Y) \sim \mathcal{N}(\mu_X, \sigma_X^2)\mathcal{N}(\mu_Y, \sigma_Y^2)$$



Method: Equating coefficients

$$p(s) = \mathcal{N}(s; 0, 1)$$
$$p(v|s) = \mathcal{N}(v; s, \sigma^2)$$

$$V = S + \epsilon$$

When we recall our model from before, we can now see that our posterior is proportional to a Gaussian, but what is its form?

$$\underbrace{p(s|v)}_{\therefore\ \text{Gaussian}} \propto \underbrace{p(v|s)}_{\text{Gaussian}}\underbrace{p(s)}_{\text{Gaussian}}$$

Method: Equating coefficients

$$\frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{2\sigma^2}\right)}$$

$$p(s|v) \propto p(v|s)p(s)$$

$$= \mathcal{N}(v; s, \sigma^2)\mathcal{N}(s; 0, 1)$$

$\mu = s$ $\qquad$ $\mu = 0$

cut end of consts as looking at proportional relationship

$$= \exp\left(-\frac{(v-s)^2}{2\sigma^2}\right)\exp\left(-\frac{s^2}{2}\right)$$

$\sigma^2 = 1$

expand squared term

$$= \exp\left(-\frac{v^2}{2\sigma^2} + \frac{sv}{\sigma^2} - \frac{s^2}{2\sigma^2} - \frac{s^2}{2}\right)$$

Method: Equating coefficients

$$p(s|v) \propto p(v|s)p(s)$$

$$= \mathcal{N}(v; s, \sigma^2)\mathcal{N}(s; 0, 1)$$

$$= \exp\left(-\frac{v^2}{2\sigma^2} + \frac{sv}{\sigma^2} - \frac{s^2}{2\sigma^2} - \frac{s^2}{2}\right)$$

$$\propto \exp\left(-\frac{1+\sigma^2}{2\sigma^2}s^2 + \frac{v}{\sigma^2}s\right)$$

$\frac{v}{\sigma^2}s$

$\underbrace{\frac{v^2}{2\sigma^2}}_{\text{const.}} + \frac{sv}{\sigma^2} \underbrace{- \frac{s^2}{2\sigma^2} - \frac{s^2}{2}}$

$\frac{sv}{\sigma^2} \rightarrow \frac{v}{\sigma^2}s$

$-s^2\left(\frac{1}{2\sigma^2} + \frac{1}{2}\right)$

$= -s^2\left(\frac{1}{2\sigma^2} + \frac{\sigma^2}{2\sigma^2}\right)$

$= -s^2\left(\frac{1+\sigma^2}{2\sigma^2}\right)$

Method: Equating coefficients

$$p(\boxed{s}|v) \propto p(v|s)p(s)$$

This is a Gaussian density
So get in form of Gaussian

$$\propto \exp\left(-\frac{1+\sigma^2}{2\sigma^2}\boxed{s^2} + \frac{v}{\sigma^2}\boxed{s}\right)$$

Gaussian with s as main RV

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

x as main RV

We want to find the coefficients of a Gaussian that make the above density look like the one we have just computed.

Method: Equating coefficients

$$p(\boxed{s}|v) \propto p(v|s)p(s)$$

$$\propto \exp\left(-\frac{1+\sigma^2}{2\sigma^2}\boxed{s^2}+\frac{v}{\sigma^2}\boxed{s}\right)$$

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

$$\propto \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Method: Equating coefficients

$$p(\boxed{s}|v) \propto p(v|s)p(s)$$

$$\propto \exp\left(-\frac{1+\sigma^2}{2\sigma^2}\boxed{s^2} + \frac{v}{\sigma^2}\boxed{s}\right)$$

$$\mathcal{N}(x;\mu,\sigma) \propto \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

$$= \exp\left(-\frac{x^2 - 2\mu x + \mu^2}{2\sigma^2}\right)$$

Method: Equating coefficients

This is normally distributed with RVs

$$p(\boxed{s}|v) \propto p(v|s)p(s)$$

$$\propto \exp\left(-\frac{1+\sigma^2}{2\sigma^2}\boxed{s^2} + \frac{v}{\sigma^2}\boxed{s}\right)$$

Now they look very similar :)

$$\mathcal{N}(x;\mu,\sigma) \propto \exp\left(-\frac{x^2 - 2\mu x + \mu^2}{2\sigma^2}\right)$$

$$= \exp\left(\frac{-1}{2\sigma^2}x^2 + \frac{\mu x}{\sigma^2} - \cancel{\frac{\mu^2}{2\sigma^2}}\right)$$

nothing to do with x so get rid (as looking at proportionality)

Method: Equating coefficients

$$
\begin{aligned}
p(\boxed{s}|v) &\propto p(v|s)p(s) \\
&\propto \exp\left(-\frac{1+\sigma^2}{2\sigma^2}\boxed{s^2} + \frac{v}{\sigma^2}\boxed{s}\right)
\end{aligned}
$$

letting $\mu \to a$, $\sigma^2 \to b$ for standard normal

$$
= c \cdot \exp\left(-\frac{1}{2b}x^2 + \frac{a}{b}x\right)
$$

$\frac{1+\sigma^2}{2\sigma^2} = \frac{1}{2b} \to 2b = \frac{2\sigma^2}{1+\sigma^2}$

$b = \frac{\sigma^2}{1+\sigma^2}$

$\frac{v}{\sigma^2} = \frac{a}{b}$

$\hookrightarrow a = \frac{vb}{\sigma^2} = v\frac{\sigma^2}{(1+\sigma^2)\sigma^2}$

$= \frac{v}{1+\sigma^2}$

$$
\implies b = \frac{\sigma^2}{1+\sigma^2} \qquad\qquad \implies a = \frac{1}{1+\sigma^2}v
$$

Method: Equating coefficients

$$p(s|v) \propto p(v|s)p(s)$$

$$= c \cdot \exp\left(-\frac{1}{2b}x^2 + \frac{a}{b}x\right)$$

$$\implies b = \frac{\sigma^2}{1+\sigma^2} \qquad \implies a = \frac{1}{1+\sigma^2}v$$

$$\implies p(s|v) = \mathcal{N}(s; \frac{1}{1+\sigma^2}v, \frac{\sigma^2}{1+\sigma^2})$$

What about non-Gaussians? Conjugacy

The algebra of equating coefficients that we just carried out works in particular because we knew the form of the posterior distribution we were looking for. In general, this is not always the case.

When it is the case and we select our priors such that the posterior has a natural "closed-form" solution, we call the prior the conjugate prior.

post. ∝ likelihood × prior

likelihood × conj. prior → post. same dist. as conj. prior.

Beta-Bernoulli Model

Seller 1: 90 positive reviews, 10 negative reviews

Seller 2: 2 positive reviews, 0 negative reviews

Who should we go with?

Beta-Bernoulli Model

Seller 1: 90 positive reviews, 10 negative reviews

Seller 2: 2 positive reviews, 0 negative reviews

$$p(x|\theta) = \theta^x(1-\theta)^x$$

likelihood we have a good experience

The seller's unknown reliability

This is what we want a random variable over

Beta-Bernoulli Model

Seller 1: 90 positive reviews, 10 negative reviews

Seller 2: 2 positive reviews, 0 negative reviews

$$p(\theta|\mathcal{D}) \propto p(\mathcal{D}|\theta)p(\theta)$$

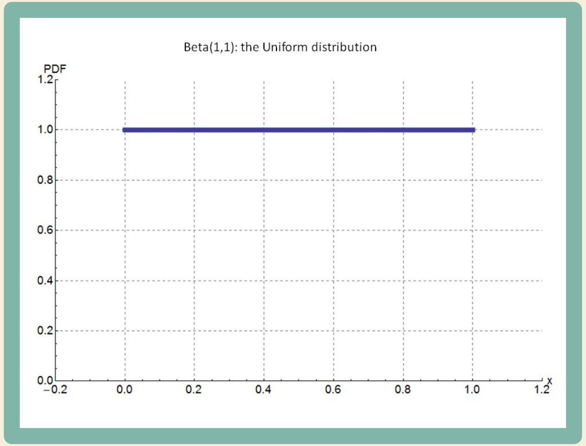
Beta-Bernoulli Model

Seller 1: 90 positive reviews, 10 negative reviews

Seller 2: 2 positive reviews, 0 negative reviews

$$p(\theta)$$

model good enough as uniform as making no assumptions



Beta prior w/ bern likelihood gives beta posterior

$$f(x;\alpha,\beta) = \text{constant} \cdot x^{\alpha-1}(1-x)^{\beta-1}$$

Beta-Bernoulli Model

Seller 1: 90 positive reviews, 10 negative reviews

Seller 2: 2 positive reviews, 0 negative reviews

$$p(\mathcal{D}|\theta) = \theta^{\#\text{ positive}}(1-\theta)^{\#\text{ negative}}$$

Beta-Bernoulli Model

Seller 1: 90 positive reviews, 10 negative reviews

Seller 2: 2 positive reviews, 0 negative reviews

$p(\theta|D) \propto p(D|\theta)\, p(\theta)$

$$p(\theta|\mathcal{D}) \propto (\theta^{90}(1-\theta)^{10})(\theta^{1}(1-\theta)^{1})$$

$$\propto \theta^{91}(1-\theta)^{11}$$

Beta: $\text{const} \times x^{\alpha-1}(1-x)^{\beta-1}$

$$p(\theta|\mathcal{D}) = \text{Beta}(91, 11)$$

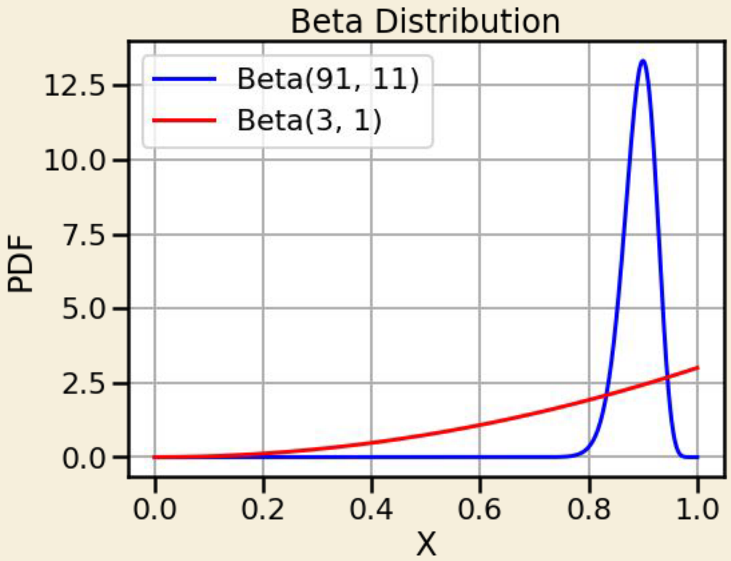
Beta-Bernoulli Model

Seller 1: 90 positive reviews, 10 negative reviews

Seller 2: 2 positive reviews, 0 negative reviews

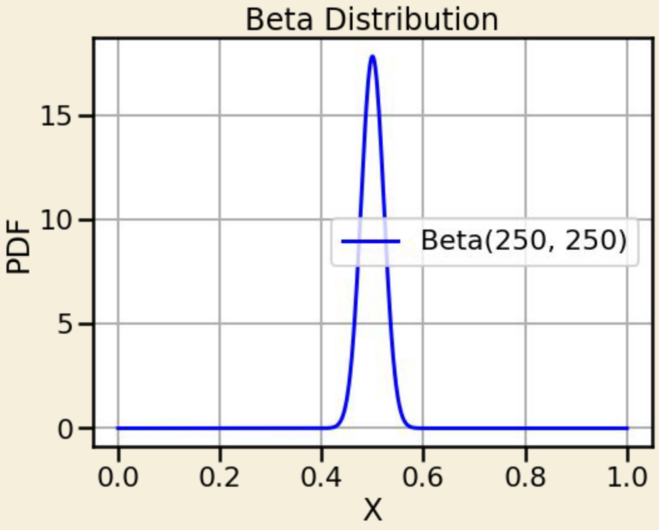
Seller 1 MAP: 0.892
(Variance: 0.0009)

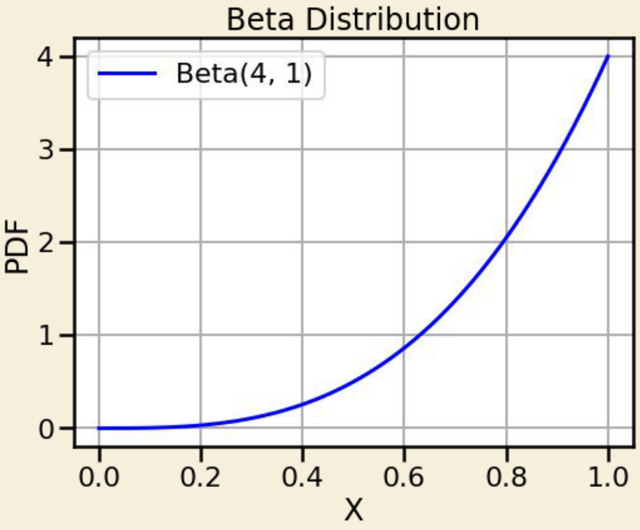
Seller 2 MAP: 0.75
(Variance: 0.05)



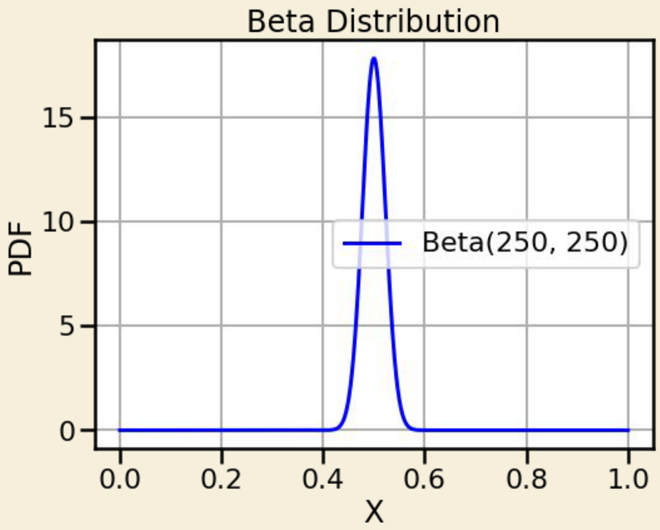
Initial look at *kinds* of uncertainty

Am I going to have a good experience with this seller?

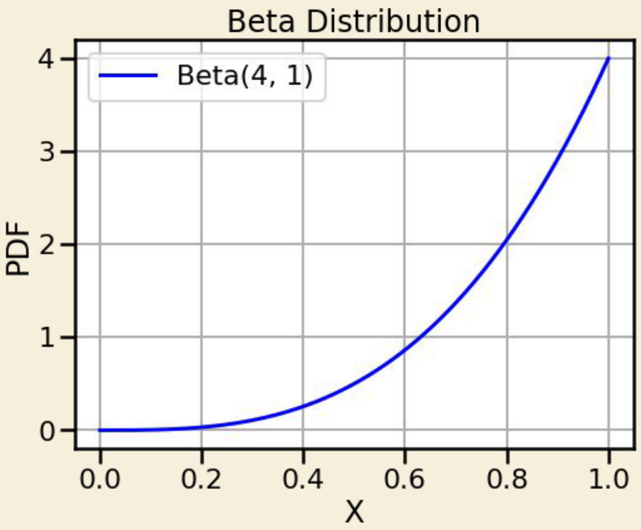




Initial look at *kinds* of uncertainty



Aleatoric uncertainty: Uncertainty attributable to the intrinsic noise of a system/the world



Epistemic uncertainty: Uncertainty attributable to lack of knowledge about the system

Subjective vs. Objective Bayes

How do we pick our prior distribution? Two schools of thought on this!

Objective Bayes: you should pick a prior that has as little influence on the analysis as possible! Picking your prior to be conjugate for algebraic nicety is not a principled reason.

Inverse Wishart distribution is the conjugate to a covariance matrix, but leads to bad inference properties

Subjective Bayes: Algebraic nicety is great. Additionally, we have all of this prior knowledge about systems we want to model and we want to incorporate that into my inference. It is a key strength of being Bayesian.

Next lecture: Bayes Theorem in Machine Learning