Recap

- To implement Bayes Classifier we need class conditional densities.

Recap

- To implement Bayes Classifier we need class conditional densities.
- We have considered maximum likelihood estimation for parameters of a density and seen many examples.

Recap

- To implement Bayes Classifier we need class conditional densities.
- We have considered maximum likelihood estimation for parameters of a density and seen many examples.
- We have briefly looked at Bayesian estimation of parameters.

Recap

- To implement Bayes Classifier we need class conditional densities.
- We have considered maximum likelihood estimation for parameters of a density and seen many examples.
- We have briefly looked at Bayesian estimation of parameters.
- In this class we discuss Bayesian estimation in more detail.

Bayesian Estimation (Recap)

- We think of the parameter as a random variable.

Bayesian Estimation (Recap)

- We think of the parameter as a random variable.
- We capture our lack of knowledge about the value of a parameter through a probability density over the parameter space.

Bayesian Estimation (Recap)

- We think of the parameter as a random variable.
- We capture our lack of knowledge about the value of a parameter through a probability density over the parameter space.
- We call this the **prior** density of the parameter.

Bayesian Estimation (Recap)

- We think of the parameter as a random variable.
- We capture our lack of knowledge about the value of a parameter through a probability density over the parameter space.
- We call this the **prior** density of the parameter.
- Any information we may have about the value of parameter can be incorporated into this.

Bayesian Estimation (Recap)

- We think of the parameter as a random variable.
- We capture our lack of knowledge about the value of a parameter through a probability density over the parameter space.
- We call this the **prior** density of the parameter.
- Any information we may have about the value of parameter can be incorporated into this.
- We then view the role of data as transforming our prior density into a **posterior** density for the parameter.

Bayesian Parameter Estimation

- As earlier, let θ be the parameter and let $\mathcal{D}$ be the data

Bayesian Parameter Estimation

- As earlier, let θ be the parameter and let $\mathcal{D}$ be the data
- Recall that

$$\mathcal{D} = \{x_1, \cdots, x_n\}$$

is the set of *iid* data and each x_i has density $f(x_i \mid \theta)$ (which is the assumed model).

Bayesian Parameter Estimation

- As earlier, let θ be the parameter and let $\mathcal{D}$ be the data
- Recall that

$$\mathcal{D} = \{x_1, \, \cdots, \, x_n\}$$

  is the set of *iid* data and each x_i has density $f(x_i \mid \theta)$ (which is the assumed model).
- Let $f(\theta)$ be the prior density of the parameter and let $f(\theta \mid \mathcal{D})$ be the posterior density.

- Now, using Bayes theorem we get

$$f(\theta \mid \mathcal{D}) = \frac{f(\mathcal{D} \mid \theta) f(\theta)}{\int \; f(\mathcal{D} \mid \theta) f(\theta) \, d\theta}$$

where $f(\mathcal{D} \mid \theta) = \prod_i \; f(x_i \mid \theta)$ is the data likelihood that we considered earlier.

- A form for the prior density, that results in the same form of density for the posterior is called **conjugate prior**.

- A form for the prior density, that results in the same form of density for the posterior is called **conjugate prior**.
- Posterior density depends on product of prior and data likelihood.

- A form for the prior density, that results in the same form of density for the posterior is called **conjugate prior**.
- Posterior density depends on product of prior and data likelihood.
- The form of data likelihood depends on the form assumed for $f(x \mid \theta)$.

- A form for the prior density, that results in the same form of density for the posterior is called **conjugate prior**.
- Posterior density depends on product of prior and data likelihood.
- The form of data likelihood depends on the form assumed for $f(x \mid \theta)$.
- Hence the conjugate prior is determined by the the form of $f(x \mid \theta)$ (and hence that of data likelihood).

- When we use conjugate prior, the prior and posterior would belong to the same class of densities.

- When we use conjugate prior, the prior and posterior would belong to the same class of densities.
- Hence calculating posterior would be like updating parameter values.

- When we use conjugate prior, the prior and posterior would belong to the same class of densities.
- Hence calculating posterior would be like updating parameter values.
- We consider a few examples of Bayesian estimation now.

- How do we use the final posterior density for implementing the classifier?
- There are many possibilities for this.
- We finally need the class conditional densities for implementing the Bayes classifier.
- So, one method is: can we find density of x based on the data (so that the density is not dependent on any unknown parameter).

- Having obtained $f(\theta \mid \mathcal{D})$, we have

$$\begin{aligned} f(x \mid \mathcal{D}) &= \int f(x, \theta \mid \mathcal{D})\, d\theta \\ &= \int f(x \mid \theta) f(\theta \mid \mathcal{D})\, d\theta \end{aligned}$$

- Depending on the form of posterior, we may be able to get a closed form expression for the density as needed.

- Another possibility is to use some specific value of θ based on the posterior density.
- We can take mode of the posterior density as the parameter value.
- Called MAP estimate. (Maximum Aposteriori Probability)
- Or, we can take the mean of the posterior density as the parameter value.
- Both these are also often used.

Example

- Consider estimating mean of a Gaussian density (with the variance assumed known).

Example

- Consider estimating mean of a Gaussian density (with the variance assumed known).
- Hence the class conditional density model is

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

where we assume that σ is known. Here μ is the unknown parameter.

- The likelihood is now given by

$$f(\mathcal{D} \mid \mu) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right)$$

- The likelihood is now given by

$$f(\mathcal{D} \mid \mu) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right)$$

- As a function of μ this has an exponential of a quadratic in μ.

- The likelihood is now given by

$$f(\mathcal{D} \mid \mu) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right)$$

- As a function of μ this has an exponential of a quadratic in μ.
- Hence, If the prior is normal (which has an exponential of a quadratic in μ) the product would once again be a normal density.

- The likelihood is now given by

$$f(\mathcal{D} \mid \mu) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right)$$

- As a function of μ this has an exponential of a quadratic in μ.
- Hence, If the prior is normal (which has an exponential of a quadratic in μ) the product would once again be a normal density.
- Thus, the conjugate prior here is normal density.

- Let us take the prior as $f(\mu) = \mathcal{N}(\mu_0,\ \sigma_0)$.

- Let us take the prior as $f(\mu) = \mathcal{N}(\mu_0,\ \sigma_0)$.
- Now the posterior density for μ can be written as

$$f(\mu \mid \mathcal{D}) = \frac{f(\mathcal{D} \mid \mu) f(\mu)}{\int\ f(\mathcal{D} \mid \mu) f(\mu)\, d\mu}$$

- Let us take the prior as $f(\mu) = \mathcal{N}(\mu_0,\ \sigma_0)$.
- Now the posterior density for μ can be written as

$$f(\mu \mid \mathcal{D}) = \frac{f(\mathcal{D} \mid \mu) f(\mu)}{\int f(\mathcal{D} \mid \mu) f(\mu)\, d\mu}$$

- By substituting for $f(\mathcal{D} \mid \mu)$ and $f(\mu)$ we get

$$f(\mu|\mathcal{D}) \propto \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \ - \ \frac{1}{2\sigma_0^2} (\mu - \mu_0)^2 \right)$$

$$f(\mu \mid \mathcal{D}) \propto \exp\left(-\frac{1}{2\sigma^2}\,\sum_{i=1}^{n}(x_i-\mu)^2 \;-\; \frac{1}{2\sigma_0^2}\,(\mu-\mu_0)^2\right)$$

$$f(\mu \mid \mathcal{D}) \propto \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \; - \; \frac{1}{2\sigma_0^2} (\mu - \mu_0)^2 \right)$$

- Hence we get $f(\mu \mid \mathcal{D}) \propto \exp(-\frac{1}{2}A)$ where

$$A = \frac{1}{\sigma^2} \sum_{i=1}^{n} x_i^2 + \mu^2 \left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2} \right) - 2\mu \left(\sum_{i=1}^{n} \frac{x_i}{\sigma^2} + \frac{\mu_0}{\sigma_0^2} \right)$$

$$f(\mu \mid \mathcal{D}) \propto \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \; - \; \frac{1}{2\sigma_0^2} (\mu - \mu_0)^2 \right)$$

- Hence we get $f(\mu \mid \mathcal{D}) \propto \exp(-\frac{1}{2}A)$ where

$$A = \frac{1}{\sigma^2} \sum_{i=1}^{n} x_i^2 + \mu^2 \left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2} \right) - 2\mu \left(\sum_{i=1}^{n} \frac{x_i}{\sigma^2} + \frac{\mu_0}{\sigma_0^2} \right)$$

- As expected, the posterior is also Gaussian.

- Suppose $f(\mu \,|\, \mathcal{D})$ is $\mathcal{N}(\mu_n,\ \sigma_n)$. Then

$$f(\mu \,|\, \mathcal{D}) \propto \ \exp\left(-\frac{1}{2}\left[\frac{\mu^2}{\sigma_n^2} + \frac{\mu_n^2}{\sigma_n^2} - 2\mu\frac{\mu_n}{\sigma_n^2}\right]\right)$$

- Suppose $f(\mu \mid \mathcal{D})$ is $\mathcal{N}(\mu_n,\ \sigma_n)$. Then

$$f(\mu \mid \mathcal{D}) \propto \ \exp\left(-\frac{1}{2}\left[\frac{\mu^2}{\sigma_n^2} + \frac{\mu_n^2}{\sigma_n^2} - 2\mu\frac{\mu_n}{\sigma_n^2}\right]\right)$$

- Now, comparing with the earlier expression, we get

- Suppose $f(\mu \mid \mathcal{D})$ is $\mathcal{N}(\mu_n,\ \sigma_n)$. Then

$$f(\mu \mid \mathcal{D}) \propto \exp\left(-\frac{1}{2}\left[\frac{\mu^2}{\sigma_n^2} + \frac{\mu_n^2}{\sigma_n^2} - 2\mu\frac{\mu_n}{\sigma_n^2}\right]\right)$$

- Now, comparing with the earlier expression, we get

$$\frac{1}{\sigma_n^2} = \frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}$$

$$\frac{\mu_n}{\sigma_n^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n} x_i + \frac{\mu_0}{\sigma_0^2}$$

- Solving these, we get

$$\sigma_n^2 = \frac{\sigma^2 \sigma_0^2}{\sigma^2 + n\sigma_0^2}$$

$$\mu_n = \frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2}\bar{\mu}_n + \frac{\sigma^2}{n\sigma_0^2 + \sigma^2}\mu_0$$

where $\bar{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$ is the ML estimate for μ.

- Solving these, we get

$$\sigma_n^2 = \frac{\sigma^2\sigma_0^2}{\sigma^2 + n\sigma_0^2}$$

$$\mu_n = \frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2}\bar{\mu}_n + \frac{\sigma^2}{n\sigma_0^2 + \sigma^2}\mu_0$$

where $\bar{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$ is the ML estimate for μ.

- The μ_n and σ_n completely specify the posterior density (after we have seen n examples).

$$
\begin{aligned}
\sigma_n^2 &= \frac{\sigma^2\sigma_0^2}{\sigma^2 + n\sigma_0^2} \\
\mu_n &= \frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2}\bar{\mu}_n + \frac{\sigma^2}{n\sigma_0^2 + \sigma^2}\mu_0
\end{aligned}
$$

$$\sigma_n^2 = \frac{\sigma^2 \sigma_0^2}{\sigma^2 + n\sigma_0^2}$$

$$\mu_n = \frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2}\bar{\mu}_n + \frac{\sigma^2}{n\sigma_0^2 + \sigma^2}\mu_0$$

- μ_n is a convex combination of $\bar{\mu}_n$ and μ_0. Both prior and data have a role to play.

$$\sigma_n^2 = \frac{\sigma^2\sigma_0^2}{\sigma^2 + n\sigma_0^2}$$

$$\mu_n = \frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2}\bar{\mu}_n + \frac{\sigma^2}{n\sigma_0^2 + \sigma^2}\mu_0$$

- μ_n is a convex combination of $\bar{\mu}_n$ and μ_0. Both prior and data have a role to play.
- For large n, $\mu_n \approx \bar{\mu}_n$ and σ_n becomes very small.

$$\sigma_n^2 = \frac{\sigma^2\sigma_0^2}{\sigma^2 + n\sigma_0^2}$$

$$\mu_n = \frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2}\bar{\mu}_n + \frac{\sigma^2}{n\sigma_0^2 + \sigma^2}\mu_0$$

- μ_n is a convex combination of $\bar{\mu}_n$ and μ_0. Both prior and data have a role to play.
- For large n, $\mu_n \approx \bar{\mu}_n$ and σ_n becomes very small.
- As n becomes very large Bayesian estimate is essentially same as ML estimate.

$$
\begin{aligned}
\sigma_n^2 &= \frac{\sigma^2 \sigma_0^2}{\sigma^2 + n\sigma_0^2} \\
\mu_n &= \frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2}\bar{\mu}_n + \frac{\sigma^2}{n\sigma_0^2 + \sigma^2}\mu_0
\end{aligned}
$$

$$
\begin{aligned}
\sigma_n^2 &= \frac{\sigma^2\sigma_0^2}{\sigma^2 + n\sigma_0^2} \\
\mu_n &= \frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2}\bar{\mu}_n + \frac{\sigma^2}{n\sigma_0^2 + \sigma^2}\mu_0
\end{aligned}
$$

- ‘Large n’ means $n\sigma_0^2 >> \sigma^2$.

$$
\begin{aligned}
\sigma_n^2 &= \frac{\sigma^2 \sigma_0^2}{\sigma^2 + n\sigma_0^2} \\
\mu_n &= \frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2}\bar{\mu}_n + \frac{\sigma^2}{n\sigma_0^2 + \sigma^2}\mu_0
\end{aligned}
$$

- ‘Large n’ means $n\sigma_0^2 >> \sigma^2$.
- We can say: μ_0 is our initial guess on μ and σ_0 determines the level of uncertainity in this guess.

- The Bayesian estimate is the whole posterior density.

- The Bayesian estimate is the whole posterior density.
- As explained earlier, we can use mean or mode of posterior.

- The Bayesian estimate is the whole posterior density.
- As explained earlier, we can use mean or mode of posterior.
- Thus we can take the class conditional density to be Gaussian with mean μ_n and variance σ^2.

- The Bayesian estimate is the whole posterior density.
- As explained earlier, we can use mean or mode of posterior.
- Thus we can take the class conditional density to be Gaussian with mean μ_n and variance σ^2.
- We can also calculate $f(x \mid \mathcal{D})$.

- We have

$$
\begin{aligned}
f(x \mid \mathcal{D}) &= \int_{-\infty}^{\infty} f(x \mid \mu) f(\mu \mid \mathcal{D})\, d\mu \\
&= \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \\
&\qquad \frac{1}{\sigma_n\sqrt{2\pi}} \exp\left(-\frac{(\mu-\mu_n)^2}{2\sigma_n^2}\right)\, d\mu
\end{aligned}
$$

- We can show that

$$f(x \mid \mathcal{D}) = \frac{1}{\sqrt{2\pi(\sigma_n^2 + \sigma^2)}} \exp\left(-\frac{(x-\mu_n)^2}{2(\sigma^2 + \sigma_n^2)}\right)$$

- We can show that

$$f(x \mid \mathcal{D}) = \frac{1}{\sqrt{2\pi(\sigma_n^2 + \sigma^2)}} \exp\left(-\frac{(x-\mu_n)^2}{2(\sigma^2 + \sigma_n^2)}\right)$$

- This is Gaussian with mean μ_n but with variance $\sigma^2 + \sigma_n^2$.

- We can show that

$$f(x \mid \mathcal{D}) = \frac{1}{\sqrt{2\pi(\sigma_n^2 + \sigma^2)}} \exp\left(-\frac{(x-\mu_n)^2}{2(\sigma^2 + \sigma_n^2)}\right)$$

- This is Gaussian with mean μ_n but with variance $\sigma^2 + \sigma_n^2$.
- This is the class conditional density we can use.

- We can show that

$$f(x \mid \mathcal{D}) = \frac{1}{\sqrt{2\pi(\sigma_n^2 + \sigma^2)}} \exp\left(-\frac{(x-\mu_n)^2}{2(\sigma^2 + \sigma_n^2)}\right)$$

- This is Gaussian with mean μ_n but with variance $\sigma^2 + \sigma_n^2$.
- This is the class conditional density we can use.
- Naturally takes care of the sample size in estimation.

Another Example

- Consider estimating a Bernoulli density with parameter p.

Another Example

- Consider estimating a Bernoulli density with parameter p.

$$f(x \mid p) = p^x (1-p)^{1-x}, \quad x \in \{0,\ 1\}$$

Another Example

- Consider estimating a Bernoulli density with parameter p.

$$f(x \mid p) = p^x (1-p)^{1-x}, \quad x \in \{0,\ 1\}$$

- The likelihood is given by

$$f(\mathcal{D} \mid p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^{\sum x_i}(1-p)^{n-\sum x_i}$$

- Hence the conjugate prior should have the form

$$f(p) \propto p^a (1-p)^b, \quad p \in [0,\ 1]$$

- Hence the conjugate prior should have the form

$$f(p) \propto p^a (1-p)^b, \quad p \in [0,\ 1]$$

- Such a density is Beta density. It is given by (with parameters a, b)

$$f(p) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\, p^{a-1}(1-p)^{b-1},\ p \in [0,\ 1], \quad a, b \geq 1$$

- Hence the conjugate prior should have the form

$$f(p) \propto p^a(1-p)^b, \quad p \in [0,\ 1]$$

- Such a density is Beta density. It is given by

$$f(p) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\, p^{a-1}(1-p)^{b-1},\ p \in [0,\ 1], \quad a, b \geq 1$$

Where $\Gamma(z)$ is the gamma function given by

$$\Gamma(z) = \int_0^\infty x^{z-1}e^{-x}\ dx$$

- The Beta($a,\ b$) density is

$$f(p) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\, p^{a-1}(1-p)^{b-1},\ p \in [0,\ 1],\ \ a,b \geq 1$$

- The Beta($a,\ b$) density is

$$f(p) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\, p^{a-1}(1-p)^{b-1},\ p \in [0,\ 1],\ \ a, b \geq 1$$

- This is an important density over $[0,\ 1]$.

- The Beta($a,\ b$) density is

$$f(p) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\, p^{a-1}(1-p)^{b-1},\ p \in [0,\ 1],\ \ a, b \geq 1$$

- This is an important density over $[0,\ 1]$.
- When $a = b = 1$ it reduces to the uniform density.

- The Beta($a,\ b$) density is

$$f(p) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\, p^{a-1}(1-p)^{b-1},\ p \in [0,\ 1],\ \ a,b \geq 1$$

- This is an important density over $[0,\ 1]$.
- When $a = b = 1$ it reduces to the uniform density.
- To show that this is a density we need to show

$$\Gamma(a)\Gamma(b) = \Gamma(a+b)\int_0^1 p^{a-1}(1-p)^{b-1}\, dp$$

$$\Gamma(a)\Gamma(b) = \int_0^\infty x^{a-1}\, e^{-x}\, dx \int_0^\infty y^{b-1}\, e^{-y}\, dy$$

$$
\begin{aligned}
\Gamma(a)\Gamma(b) &= \int_0^\infty x^{a-1}\, e^{-x}\, dx \int_0^\infty y^{b-1}\, e^{-y}\, dy \\
&= \int_0^\infty \left[\int_0^\infty e^{-(x+y)}\, x^{a-1}\, y^{b-1}\, dy \right] dx
\end{aligned}
$$

$$
\begin{aligned}
\Gamma(a)\Gamma(b) &= \int_0^\infty x^{a-1}\, e^{-x}\, dx \int_0^\infty y^{b-1}\, e^{-y}\, dy \\
&= \int_0^\infty \left[\int_0^\infty e^{-(x+y)}\, x^{a-1}\, y^{b-1}\, dy \right] dx
\end{aligned}
$$

We now change the variable in the inner integral from y to t as: $t = x + y$.

$$
\begin{aligned}
\Gamma(a)\Gamma(b) &= \int_0^\infty x^{a-1}\, e^{-x}\, dx \int_0^\infty y^{b-1}\, e^{-y}\, dy \\
&= \int_0^\infty \left[\int_0^\infty e^{-(x+y)}\, x^{a-1}\, y^{b-1}\, dy \right] dx \\
&= \int_0^\infty \left[\int_x^\infty e^{-t}\, x^{a-1}\, (t-x)^{b-1}\, dt \right] dx
\end{aligned}
$$

$$
\begin{aligned}
\Gamma(a)\Gamma(b) &= \int_0^\infty x^{a-1}\, e^{-x}\, dx \int_0^\infty y^{b-1}\, e^{-y}\, dy \\
&= \int_0^\infty \left[\int_0^\infty e^{-(x+y)}\, x^{a-1}\, y^{b-1}\, dy\right] dx \\
&= \int_0^\infty \left[\int_x^\infty e^{-t}\, x^{a-1}\, (t-x)^{b-1}\, dt\right] dx
\end{aligned}
$$

Now we interchange the order of integration.

$$
\begin{aligned}
\Gamma(a)\Gamma(b) &= \int_0^\infty x^{a-1}\, e^{-x}\, dx \int_0^\infty y^{b-1}\, e^{-y}\, dy \\
&= \int_0^\infty \left[\int_0^\infty e^{-(x+y)}\, x^{a-1}\, y^{b-1}\, dy \right] dx \\
&= \int_0^\infty \left[\int_x^\infty e^{-t}\, x^{a-1}\, (t-x)^{b-1}\, dt \right] dx \\
&= \int_0^\infty \left[\int_0^t e^{-t}\, x^{a-1}\, (t-x)^{b-1}\, dx \right] dt
\end{aligned}
$$

$$\Gamma(a)\Gamma(b) \;=\; \int_0^\infty \left[\int_0^t e^{-t}\, x^{a-1}\, (t-x)^{b-1}\, dx\right]\, dt$$

$$\Gamma(a)\Gamma(b) \; = \; \int_0^\infty \left[\int_0^t e^{-t}\, x^{a-1}\, (t-x)^{b-1}\, dx \right] dt$$

Now, in the inner integral we change the variable from x to u as: $x = tu$. (When x goes from 0 to t, u goes from 0 to 1. Also, $dx = t du$).

$$
\begin{aligned}
\Gamma(a)\Gamma(b) &= \int_0^\infty \left[\int_0^t e^{-t}\, x^{a-1}\, (t-x)^{b-1}\, dx\right]\, dt \\
&= \int_0^\infty \left[\int_0^1 e^{-t}\, t^{a-1}\, u^{a-1}\, t^{b-1}\, (1-u)^{b-1}\, t\, du\right]\, dt
\end{aligned}
$$

$$
\begin{aligned}
\Gamma(a)\Gamma(b) &= \int_0^\infty \left[\int_0^t e^{-t}\, x^{a-1}\, (t-x)^{b-1}\, dx\right] \, dt \\
&= \int_0^\infty \left[\int_0^1 e^{-t}\, t^{a-1}\, u^{a-1}\, t^{b-1}\, (1-u)^{b-1}\, t\, du\right] \, dt \\
&= \int_0^\infty \left[\int_0^1 e^{-t}\, t^{a+b-1}\, u^{a-1}\, (1-u)^{b-1}\, du\right] \, dt
\end{aligned}
$$

$$
\begin{aligned}
\Gamma(a)\Gamma(b) &= \int_0^\infty \left[\int_0^t e^{-t}\, x^{a-1}\, (t-x)^{b-1}\, dx\right] dt \\
&= \int_0^\infty \left[\int_0^1 e^{-t}\, t^{a-1}\, u^{a-1}\, t^{b-1}\, (1-u)^{b-1}\, t\, du\right] dt \\
&= \int_0^\infty \left[\int_0^1 e^{-t}\, t^{a+b-1}\, u^{a-1}\, (1-u)^{b-1}\, du\right] dt \\
&= \int_0^\infty e^{-t}\, t^{a+b-1}\, dt \int_0^1 u^{a-1}\, (1-u)^{b-1}\, du
\end{aligned}
$$

Thus we get

$$\begin{aligned}\Gamma(a)\Gamma(b) &= \int_0^\infty e^{-t}\, t^{a+b-1}\, dt \int_0^1 u^{a-1}\,(1-u)^{b-1}\, du \\ &= \Gamma(a+b) \int_0^1 u^{a-1}\,(1-u)^{b-1}\, du\end{aligned}$$

This completes the proof that the Beta density as given is indeed a density

- The Beta($a,\ b$) density is given by

$$f(p) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\, p^{a-1}(1-p)^{b-1},\ p \in [0,\ 1],\ \ a, b \geq 1$$

- The Beta($a,\ b$) density is given by

$$f(p) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\, p^{a-1}(1-p)^{b-1},\ p \in [0,\ 1],\ \ a, b \geq 1$$

- By differentiating we can easily show that its mode is at $\frac{a-1}{a+b-2}$.

- The Beta($a,\ b$) density is given by

$$f(p) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\, p^{a-1}(1-p)^{b-1},\ p \in [0,\ 1],\ \ a, b \geq 1$$

- By differentiating we can easily show that its mode is at $\frac{a-1}{a+b-2}$.
- We can find its expected value as follows.

$$\text{mean} = \int_0^1 p \, \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \, p^{a-1} \, (1-p)^{b-1} \, dp$$

$$
\begin{aligned}
\text{mean} &= \int_0^1 p \, \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \, p^{a-1} \, (1-p)^{b-1} \, dp \\
&= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^1 p^{a} \, (1-p)^{b-1} \, dp
\end{aligned}
$$

$$
\begin{aligned}
\text{mean} &= \int_0^1 p \, \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \, p^{a-1} \, (1-p)^{b-1} \, dp \\
&= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^1 p^{a} \, (1-p)^{b-1} \, dp \\
&= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \, \frac{\Gamma(a+1)\Gamma(b)}{\Gamma(a+b+1)}
\end{aligned}
$$

$$
\begin{aligned}
\text{mean} &= \int_0^1 p \, \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \, p^{a-1} \, (1-p)^{b-1} \, dp \\
&= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^1 p^{a} \, (1-p)^{b-1} \, dp \\
&= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \, \frac{\Gamma(a+1)\Gamma(b)}{\Gamma(a+b+1)} \\
&= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \, \frac{a\Gamma(a)\Gamma(b)}{(a+b)\Gamma(a+b)} = \frac{a}{a+b}
\end{aligned}
$$

- Now getting back to Bayesian estimation of Bernoulli density, the posterior is given by

$$f(p \mid \mathcal{D}) \;=\; K\, f(\mathcal{D} \mid p) f(p)$$

- Now getting back to Bayesian estimation of Bernoulli density, the posterior is given by

$$\begin{aligned} f(p \mid \mathcal{D}) &= K\ f(\mathcal{D} \mid p) f(p) \\ &= K_1\ p^{\sum x_i}\ (1-p)^{n-\sum x_i}\ p^{a-1}\ (1-p)^{b-1} \end{aligned}$$

- Now getting back to Bayesian estimation of Bernoulli density, the posterior is given by

$$\begin{aligned} f(p \,|\, \mathcal{D}) &= K\ f(\mathcal{D} \,|\, p) f(p) \\ &= K_1\, p^{\sum x_i}\, (1-p)^{n-\sum x_i}\, p^{a-1}\, (1-p)^{b-1} \\ &= K_1\, p^{\sum x_i + a - 1}\, (1-p)^{n+b-\sum x_i - 1} \end{aligned}$$

- Now getting back to Bayesian estimation of Bernoulli density, the posterior is given by

$$\begin{aligned} f(p \mid \mathcal{D}) &= K\ f(\mathcal{D} \mid p) f(p) \\ &= K_1\ p^{\sum x_i}\ (1-p)^{n-\sum x_i}\ p^{a-1}\ (1-p)^{b-1} \\ &= K_1\ p^{\sum x_i + a - 1}\ (1-p)^{n+b-\sum x_i - 1} \end{aligned}$$

- Hence the posterior is Beta $\left(\sum x_i + a,\ n + b - \sum x_i\right)$ density

- Suppose we want the MAP estimate.

- Suppose we want the MAP estimate.
- Recall that mode of Beta($a,\ b$) is $\frac{a-1}{a+b-2}$.

- Suppose we want the MAP estimate.
- Recall that mode of Beta($a,\ b$) is $\frac{a-1}{a+b-2}$.
- Hence MAP estimate (mode of posterior density) is given by

$$\hat{p} = \frac{\sum_{i=1}^{n} x_i + a - 1}{n + a + b - 2}$$

- Suppose we want the MAP estimate.
- Recall that mode of Beta($a,\ b$) is $\frac{a-1}{a+b-2}$.
- Hence MAP estimate (mode of posterior density) is given by

$$\hat{p} = \frac{\sum_{i=1}^{n}\ x_i + a - 1}{n + a + b - 2}$$

- If $a = b = 1$ then this is same as ML estimate $\frac{1}{n} \sum x_i$.

- Suppose we want the MAP estimate.
- Recall that mode of Beta($a,\ b$) is $\frac{a-1}{a+b-2}$.
- Hence MAP estimate (mode of posterior density) is given by

$$\hat{p} = \frac{\sum_{i=1}^{n} x_i + a - 1}{n + a + b - 2}$$

- If $a = b = 1$ then this is same as ML estimate $\frac{1}{n} \sum x_i$.
- If $a = b = 1$ then prior is ‘flat’ and hence mode of posterior is maximum of likelihood.

- As earlier, we can compute $f(x \mid \mathcal{D})$ and use it as the class conditional density.

- As earlier, we can compute $f(x \mid \mathcal{D})$ and use it as the class conditional density.
- Since $x \in \{0,\ 1\}$, we need only $P(x = 1 \mid \mathcal{D})$.

- As earlier, we can compute $f(x \mid \mathcal{D})$ and use it as the class conditional density.
- Since $x \in \{0,\ 1\}$, we need only $P(x = 1 \mid \mathcal{D})$.

$$P[x = 1 \mid \mathcal{D}] \;=\; \int_0^1 P[x = 1 \mid p]\, f(p \mid \mathcal{D})\, dp$$

- As earlier, we can compute $f(x \mid \mathcal{D})$ and use it as the class conditional density.
- Since $x \in \{0,\ 1\}$, we need only $P(x = 1 \mid \mathcal{D})$.

$$\begin{aligned} P[x = 1 \mid \mathcal{D}] &= \int_0^1 P[x = 1 \mid p]\, f(p \mid \mathcal{D})\, dp \\ &= \int_0^1 p\, f(p \mid \mathcal{D})\, dp \end{aligned}$$

- As earlier, we can compute $f(x \mid \mathcal{D})$ and use it as the class conditional density.
- Since $x \in \{0,\ 1\}$, we need only $P(x = 1 \mid \mathcal{D})$.

$$\begin{aligned} P[x = 1 \mid \mathcal{D}] &= \int_0^1 P[x = 1 \mid p]\, f(p \mid \mathcal{D})\, dp \\ &= \int_0^1 p\, f(p \mid \mathcal{D})\, dp \\ &= \frac{\sum_{i=1}^{n} x_i + a}{n + a + b} \end{aligned}$$

- As earlier, we can compute $f(x \mid \mathcal{D})$ and use it as the class conditional density.
- Since $x \in \{0,\ 1\}$, we need only $P(x = 1 \mid \mathcal{D})$.

$$
\begin{aligned}
P[x = 1 \mid \mathcal{D}] &= \int_0^1 P[x = 1 \mid p]\, f(p \mid \mathcal{D})\, dp \\
&= \int_0^1 p\, f(p \mid \mathcal{D})\, dp \\
&= \frac{\sum_{i=1}^{n} x_i + a}{n + a + b}
\end{aligned}
$$

- This turns out to be simply the mean of the posterior.

- The ML estimate for p was

$$\hat{p}_{ML} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

- The ML estimate for p was

$$\hat{p}_{ML} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

- The Bayesian estimate is

$$\hat{p}_B = \frac{\sum_{i=1}^{n} x_i + a}{n + a + b}$$

- The ML estimate for p was

$$\hat{p}_{ML} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

- The Bayesian estimate is

$$\hat{p}_B = \frac{\sum_{i=1}^{n} x_i + a}{n + a + b}$$

- Choice of prior determines values of a, b.

$$\hat{p}_{ML} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

$$\hat{p}_B = \frac{\sum_{i=1}^{n} x_i + a}{n + a + b}$$

$$\hat{p}_{ML} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

$$\hat{p}_B = \frac{\sum_{i=1}^{n} x_i + a}{n + a + b}$$

- We can say we have started with $a + b$ ‘fictitous’ trials of which a were successes.

$$\hat{p}_{ML} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

$$\hat{p}_B = \frac{\sum_{i=1}^{n} x_i + a}{n + a + b}$$

- We can say we have started with $a+b$ 'fictitous' trials of which a were successes.
- This is how our 'prior beliefs' affect final estimate.

$$\hat{p}_{ML} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

$$\hat{p}_B = \frac{\sum_{i=1}^{n} x_i + a}{n + a + b}$$

- We can say we have started with $a + b$ 'fictitous' trials of which a were successes.
- This is how our 'prior beliefs' affect final estimate.
- As n becomes large, Bayes estimate is same as ML.