Recap

- To implement Bayes Classifier we need class conditional densities.

Recap

- To implement Bayes Classifier we need class conditional densities.
- Two main approaches to estimating densities – Parametric and non-parametric

Recap

- To implement Bayes Classifier we need class conditional densities.
- Two main approaches to estimating densities – Parametric and non-parametric
- In the parametric method we assume that the form of the density is known and estimate the parameters.

Recap

- To implement Bayes Classifier we need class conditional densities.
- Two main approaches to estimating densities – Parametric and non-parametric
- In the parametric method we assume that the form of the density is known and estimate the parameters.
- Maximum likelihood method is a general procedure for obtaining consistent estimators for parameters.

Recap

- Maximum Likelihood (ML) estimate is the maximizer of the likelihood (or log likelihhod) function.

Recap

- Maximum Likelihood (ML) estimate is the maximizer of the likelihood (or log likelihhod) function.
- For most standard density models, one can analytically derive ML estimates.

Recap

- Maximum Likelihood (ML) estimate is the maximizer of the likelihood (or log likelihhod) function.
- For most standard density models, one can analytically derive ML estimates.
- We have seen some examples of obtaining ML estimates.

Recap

- Maximum Likelihood (ML) estimate is the maximizer of the likelihood (or log likelihhod) function.
- For most standard density models, one can analytically derive ML estimates.
- We have seen some examples of obtaining ML estimates.
- We now see more examples of ML estimates.

Example

- Suppose the assumed density for x is exponential

Example

- Suppose the assumed density for x is exponential

$$f(x \mid \lambda) = \lambda \exp(-\lambda x), \; x \geq 0$$

Example

- Suppose the assumed density for x is exponential

$$f(x \mid \lambda) = \lambda \exp(-\lambda x), \; x \geq 0$$

- Given *iid* data, $\mathcal{D} = \{x_1, \cdots, x_n\}$, we need to estimate λ.

Example

- Suppose the assumed density for x is exponential

$$f(x \mid \lambda) = \lambda \exp(-\lambda x), \; x \geq 0$$

- Given *iid* data, $\mathcal{D} = \{x_1, \cdots, x_n\}$, we need to estimate λ.
- The likelihood function is

$$L(\lambda \mid \mathcal{D}) = \prod_{i=1}^{n} \lambda \exp(-\lambda x_i)$$

Example

- The log likelihood function is

$$l(\lambda \mid \mathcal{D}) = \sum_{i=1}^{n} \left(\ln(\lambda) - \lambda x_i\right)$$

Example

- The log likelihood function is

$$l(\lambda \mid \mathcal{D}) = \sum_{i=1}^{n} \left(\ln(\lambda) - \lambda x_i\right)$$

- Differentiating w.r.t. λ and equating to zero, we get

$$\frac{n}{\lambda} - \sum_{i=1}^{n} x_i = 0$$

- This gives us the final ML estimate as

$$\hat{\lambda} = \frac{n}{\sum_{i=1}^{n} x_i}$$

- This gives us the final ML estimate as

$$\hat{\lambda} = \frac{n}{\sum_{i=1}^{n} x_i}$$

- The final estimate is intuitively clear. (Note that $Ex = \frac{1}{\lambda}$).

Another Example

- Consider the multidimensional Gaussian density

$$f(x \,|\, \theta) = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \; \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu) \right)$$

where $x \in \Re^d$ and $\theta = (\mu, \Sigma)$ are the parameters.

Another Example

- Consider the multidimensional Gaussian density

$$f(x\,|\,\theta) = \frac{1}{\sqrt{(2\pi)^d|\Sigma|}}\ \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)$$

  where $x \in \Re^d$ and $\theta = (\mu, \Sigma)$ are the parameters.

- For a random vector x having the above joint density, $\mu \in \Re^d$ is the mean vector (i.e., $Ex = \mu$) and the $d \times d$ matrix Σ is the covariance matrix defined by

$$\Sigma = E(x-\mu)(x-\mu)^T$$

- To find the ML estimate for the parameters, we have to maximise the log likelihood.

- To find the ML estimate for the parameters, we have to maximise the log likelihood.
- Recall that the log likelihood function is defined by

$$l(\theta \mid \mathcal{D}) = \sum_{i=1}^{n} \ln(f(x_i \mid \theta))$$

where $\mathcal{D} = \{x_1, \cdots, x_n\}$ constitutes the *iid* data from which we are estimating the parameters of the density.

The log likelihood function is given by

$$l(\theta|\mathcal{D}) = \sum_{i=1}^{n} \left(-\frac{1}{2}\ln((2\pi)^d|\Sigma|) - \frac{1}{2}(x_i - \mu)^T\Sigma^{-1}(x_i - \mu) \right)$$

where $\theta = (\mu, \Sigma)$ constitute the parameters to be estimated.

The log likelihood function is given by

$$l(\theta|\mathcal{D}) = \sum_{i=1}^{n} \left(-\frac{1}{2} \ln((2\pi)^d|\Sigma|) - \frac{1}{2}(x_i - \mu)^T \Sigma^{-1} (x_i - \mu) \right)$$

where $\theta = (\mu, \Sigma)$ constitute the parameters to be estimated.

- To find the ML estimates, we have to equate the partial derivatives of l (with respect to the parameters) to zero and solve.

- Now, $\frac{\partial l}{\partial \mu} = 0$ gives us

$$\sum_{i=1}^{n} \Sigma^{-1}(x_i - \mu) = 0$$

- Now, $\frac{\partial l}{\partial \mu} = 0$ gives us

$$\sum_{i=1}^{n} \Sigma^{-1}(x_i - \mu) = 0$$

which gives us the ML estimate for μ as

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

- Now, $\frac{\partial l}{\partial \mu} = 0$ gives us

$$\sum_{i=1}^{n} \Sigma^{-1}(x_i - \mu) = 0$$

which gives us the ML estimate for μ as

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Thus, even in the multidimensional case, the ML estimate for mean is the sample mean.

- Finding the partial derivative with respect to Σ is algebraically involved.

- Finding the partial derivative with respect to Σ is algebraically involved.
- However, one can show that the ML estimate for Σ is

$$\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\mu})(x_i - \hat{\mu})^T$$

- Finding the partial derivative with respect to Σ is algebraically involved.
- However, one can show that the ML estimate for Σ is

$$\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\mu})(x_i - \hat{\mu})^T$$

- Again, the final ML estimate is intuitively obvious.

- Finding the partial derivative with respect to Σ is algebraically involved.
- However, one can show that the ML estimate for Σ is

$$\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\mu})(x_i - \hat{\mu})^T$$

- Again, the final ML estimate is intuitively obvious. (Recall that $\Sigma = E(x - \mu)(x - \mu)^T$).

One more example

- Suppose we have a discrete random variable, say, z, that takes values $a_1, \cdots, a_M$ with probabilities $p_1, \cdots, p_M$.

One more example

- Suppose we have a discrete random variable, say, z, that takes values $a_1, \cdots, a_M$ with probabilities $p_1, \cdots, p_M$.
- Given data in the form of *iid* realizations of this random variable, we want to estimate the parameters p_i.

One more example

- Suppose we have a discrete random variable, say, z, that takes values $a_1, \cdots, a_M$ with probabilities $p_1, \cdots, p_M$.
- Given data in the form of *iid* realizations of this random variable, we want to estimate the parameters p_i.
- Note that the parameters satisfy: $p_i \geq 0$ and $\sum_i p_i = 1$.

- For our estimation, we represent the discrete random variable, z by an M-dimensional vector random variable $x = [x^1, \cdots, x^M]^T$.

- For our estimation, we represent the discrete random variable, z by an M-dimensional vector random variable $x = [x^1, \cdots, x^M]^T$.
- The idea is that if z takes value a_i then we will represent it by x whose i^{th} component is one and all others are zero.

- For our estimation, we represent the discrete random variable, z by an M-dimensional vector random variable $x = [x^1, \cdots, x^M]^T$.
- The idea is that if z takes value a_i then we will represent it by x whose i^{th} component is one and all others are zero.
- So, the random vector x actually takes only M possible values, namely, $[1,\ 0,\ \cdots,\ 0]^T$, $[0,\ 1,\ 0,\ \cdots, 0]^T$ etc.

- For our estimation, we represent the discrete random variable, z by an M-dimensional vector random variable $x = [x^1, \cdots, x^M]^T$.
- The idea is that if z takes value a_i then we will represent it by x whose i^{th} component is one and all others are zero.
- So, the random vector x actually takes only M possible values, namely, $[1,\ 0,\ \cdots,\ 0]^T$, $[0,\ 1,\ 0,\ \cdots, 0]^T$ etc.
- This is sometimes called '1 of M' representation for a discrete random variable taking M values.

- Thus, $x = [x^1, \cdots, x^M]^T$ satisfies: $x^i \in \{0, 1\}$ and $\sum_i x^i = 1$.

- Thus, $x = [x^1, \cdots, x^M]^T$ satisfies: $x^i \in \{0, 1\}$ and $\sum_i x^i = 1$.
- Also now we have $p_i =$ Prob$[x^i = 1]$.

- Thus, $x = [x^1, \cdots, x^M]^T$ satisfies: $x^i \in \{0, 1\}$ and $\sum_i x^i = 1$.
- Also now we have $p_i = \text{Prob}[x^i = 1]$.
- Now the mass function for x can be written as

$$f(x \mid p) = \prod_{i=1}^{M} p_i^{x^i},$$

$$x = [x^1, \cdots, x^M]^T, \; x^i \in \{0, 1\}, \; \sum_i x^i = 1$$

- Thus, $x = [x^1, \cdots, x^M]^T$ satisfies:
  $x^i \in \{0, 1\}$ and $\sum_i x^i = 1$.
- Also now we have $p_i = \text{Prob}[x^i = 1]$.
- Now the mass function for x can be written as

$$f(x \mid p) = \prod_{i=1}^{M} p_i^{x^i},$$

$$x = [x^1, \cdots, x^M]^T, \ x^i \in \{0, 1\}, \ \sum_i x^i = 1$$

- Here, $p = (p_1, \cdots, p_M)^T$ is the parameter vector.

- Now the problem of estimating the parameters, p_i, becomes the following.

- Now the problem of estimating the parameters, p_i, becomes the following.
- We are given *iid* data

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

where $x_i = [x_i^1, \cdots, x_i^M]^T$ with $x_i^j \in \{0, 1\}$ and $\sum_j x_i^j = 1, \forall i$.

- Now the problem of estimating the parameters, p_i, becomes the following.
- We are given *iid* data

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

where $x_i = [x_i^1, \cdots, x_i^M]^T$ with $x_i^j \in \{0, 1\}$ and $\sum_j x_i^j = 1, \forall i$.

- We know the probability mass function of x and we need to derive ML estimates for parameters p_i.

- The log likelihood function is given by

- The log likelihood function is given by

$$l(p \mid \mathcal{D}) = \sum_{i=1}^{n} \ln(f(x_i \mid p))$$

- The log likelihood function is given by

$$
\begin{aligned}
l(p \mid \mathcal{D}) &= \sum_{i=1}^{n} \ln(f(x_i \mid p)) \\
&= \sum_{i=1}^{n} \ln\left(\prod_{j=1}^{M} p_j^{x_i^j}\right)
\end{aligned}
$$

- The log likelihood function is given by

$$
\begin{aligned}
l(p \mid \mathcal{D}) &= \sum_{i=1}^{n} \ln(f(x_i \mid p)) \\
&= \sum_{i=1}^{n} \ln\left(\prod_{j=1}^{M} p_j^{x_i^j}\right) \\
&= \sum_{i=1}^{n} \sum_{j=1}^{M} x_i^j \ln(p_j)
\end{aligned}
$$

- We now want to find values for $p_i, \ i = 1, \cdots, M$, to maximize $l(p \,|\, \mathcal{D})$.

- We now want to find values for $p_i, \ i = 1, \cdots, M$, to maximize $l(p \,|\, \mathcal{D})$.
- But this is not an unconstrained maximization.

- We now want to find values for $p_i, \; i = 1, \cdots, M$, to maximize $l(p \mid \mathcal{D})$.
- But this is not an unconstrained maximization.
- We need to maximize l over only those p_i that satisfy $p_i \geq 0$ and $\sum_i p_i = 1$.

- We now want to find values for $p_i, \; i = 1, \cdots, M$, to maximize $l(p \mid \mathcal{D})$.
- But this is not an unconstrained maximization.
- We need to maximize l over only those p_i that satisfy $p_i \geq 0$ and $\sum_i p_i = 1$.
- Hence ML estimation of the parameters here becomes a constrained optimization problem as follows.

The constrained optimization problem is

$$\max_{p_i} \qquad l(p \mid \mathcal{D}) = \sum_{i=1}^{n} \sum_{j=1}^{M} x_i^j \ln(p_j)$$

subject to $\displaystyle \sum_{i=1}^{M} p_i = 1$

The constrained optimization problem is

$$\max_{p_i} \qquad l(p \mid \mathcal{D}) = \sum_{i=1}^{n} \sum_{j=1}^{M} x_i^j \ln(p_j)$$

subject to $\displaystyle \sum_{i=1}^{M} p_i = 1$

- We can solve this by the method of lagrange multipliers. (We have not explicitly included the non-negativity constraint).

- The lagrangian for this problem is given by

$$\sum_{i=1}^{n} \sum_{s=1}^{M} x_i^s \ln(p_s) + \lambda \left(1 - \sum_{s=1}^{M} p_s\right)$$

where λ is the Lagrange multiplier.

- The lagrangian for this problem is given by

$$\sum_{i=1}^{n} \sum_{s=1}^{M} x_i^s \ln(p_s) + \lambda \left(1 - \sum_{s=1}^{M} p_s \right)$$

where λ is the Lagrange multiplier.

- Now, we calculate the partial derivatives of the Lagrangian and equate them to zero to get the maximum.

- This gives us

$$\sum_{i=1}^{n} \frac{x_i^j}{p_j} - \lambda = 0, \; j = 1, \cdots, M$$

- This gives us

$$\sum_{i=1}^{n} \frac{x_i^j}{p_j} - \lambda = 0,\ j = 1, \cdots, M$$

Solving this, we get

$$p_j = \frac{1}{\lambda} \sum_{i=1}^{n} x_i^j,\ j = 1, \cdots, M$$

- Now using the constraint, $\sum_j p_j = 1$, we get value of λ as

- Now using the constraint, $\sum_j p_j = 1$, we get value of λ as

$$\lambda = \sum_{j=1}^{M} \sum_{i=1}^{n} x_i^j$$

- Now using the constraint, $\sum_j p_j = 1$, we get value of λ as

$$\begin{aligned} \lambda &= \sum_{j=1}^{M} \sum_{i=1}^{n} x_i^j \\ &= \sum_{i=1}^{n} \sum_{j=1}^{M} x_i^j \end{aligned}$$

- Now using the constraint, $\sum_j p_j = 1$, we get value of λ as

$$
\begin{aligned}
\lambda &= \sum_{j=1}^{M} \sum_{i=1}^{n} x_i^j \\
&= \sum_{i=1}^{n} \sum_{j=1}^{M} x_i^j \\
&= n
\end{aligned}
$$

- Now using the constraint, $\sum_j p_j = 1$, we get value of λ as

$$\begin{aligned}
\lambda &= \sum_{j=1}^{M} \sum_{i=1}^{n} x_i^j \\
&= \sum_{i=1}^{n} \sum_{j=1}^{M} x_i^j \\
&= n
\end{aligned}$$

where last step follows because $\sum_j x_i^j = 1, \forall i$.

- Thus, we get the final ML estimate for p_j as

$$\hat{p}_j = \frac{1}{n} \sum_{i=1}^{n} x_i^j$$

- Thus, we get the final ML estimate for p_j as

$$\hat{p}_j = \frac{1}{n} \sum_{i=1}^{n} x_i^j$$

- The final ML estimate for p_j is the fraction of times the j^{th} value occurs – intuitively clear.

- The distribution (or probability mass function) of any discrete random variable taking finitely many values, is specified by some M parameters like the p_i.

- The distribution (or probability mass function) of any discrete random variable taking finitely many values, is specified by some M parameters like the p_i.
- Hence, what we presented is a general procedure using which we can estimate the distribution of any discrete random variable.

- The distribution (or probability mass function) of any discrete random variable taking finitely many values, is specified by some M parameters like the p_i.
- Hence, what we presented is a general procedure using which we can estimate the distribution of any discrete random variable.
- Also, note that for discrete random variables, there is really no distinction between parametric and non-parametric ways of estimating the distribution.

- Features that take only finitely many values are important in some pattern classification problems.

- Features that take only finitely many values are important in some pattern classification problems.
- For example, search and ranking, document classification, spam filtering etc.

- Features that take only finitely many values are important in some pattern classification problems.
- For example, search and ranking, document classification, spam filtering etc.
- For example, for document classification, we can use ‘word count’ as the feature vector.

- Features that take only finitely many values are important in some pattern classification problems.
- For example, search and ranking, document classification, spam filtering etc.
- For example, for document classification, we can use ‘word count’ as the feature vector.
- Often called, ‘bag of words’ representation.

- In such cases, each feature is a discrete random variable.

- In such cases, each feature is a discrete random variable.
- We can estimate (marginal) distribution of feature using our procedure.

- In such cases, each feature is a discrete random variable.
- We can estimate (marginal) distribution of feature using our procedure.
- To implement Bayes classifier we need **joint** distribution of the feature vector.

- In such cases, each feature is a discrete random variable.
- We can estimate (marginal) distribution of feature using our procedure.
- To implement Bayes classifier we need **joint** distribution of the feature vector.
- We can, e.g., assume features are independent.

- In such cases, each feature is a discrete random variable.
- We can estimate (marginal) distribution of feature using our procedure.
- To implement Bayes classifier we need **joint** distribution of the feature vector.
- We can, e.g., assume features are independent.
- Then, joint mass function is product of marginals.
- Often called, ‘naive Bayes’ classifier

ML Estimation

- ML estimates of parameters (of a density) are obtained as maximizers of the (log) likelihood function.

ML Estimation

- ML estimates of parameters (of a density) are obtained as maximizers of the (log) likelihood function.
- We have seen many examples of how we can analytically derive ML estimates.

ML Estimation

- ML estimates of parameters (of a density) are obtained as maximizers of the (log) likelihood function.
- We have seen many examples of how we can analytically derive ML estimates.
- ML estimates are easy to obtain for most standrad densities and it is a very useful method of estimation.

- ML method of estimation has some drawbacks.

- ML method of estimation has some drawbacks.
- ML estimates are consistent. Hence, given large number of samples we would get good estimates.

- ML method of estimation has some drawbacks.
- ML estimates are consistent. Hence, given large number of samples we would get good estimates.
- However, when sample size is small, ML estimates may be quite bad.

- ML method of estimation has some drawbacks.
- ML estimates are consistent. Hence, given large number of samples we would get good estimates.
- However, when sample size is small, ML estimates may be quite bad.
- Also, the method does not allow one to incorporate any additional knowledge one may have about the values of unknown parameters.

- ML method of estimation has some drawbacks.
- ML estimates are consistent. Hence, given large number of samples we would get good estimates.
- However, when sample size is small, ML estimates may be quite bad.
- Also, the method does not allow one to incorporate any additional knowledge one may have about the values of unknown parameters.
- The final estimated value of the parameter is determined by data alone.

Bayesian Estimation

- Bayesian estimation is the second parametric method of estimation that we consider in this course.

Bayesian Estimation

- Bayesian estimation is the second parametric method of estimation that we consider in this course.
- In ML estimation the parameters are taken to be constants that are unknown.

Bayesian Estimation

- Bayesian estimation is the second parametric method of estimation that we consider in this course.
- In ML estimation the parameters are taken to be constants that are unknown.
- In Bayesian estimation we think of the parameter itself as a random variable.

Bayesian Estimation

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

Bayesian Estimation

- 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

- 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

- 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. (We will see the details of this shortly).

Bayesian Approach

- We can think of the *prior* density of the parameter as capturing our **subjective beliefs** about the parameter value.

Bayesian Approach

- We can think of the *prior* density of the parameter as capturing our **subjective beliefs** about the parameter value.
- Thus, our final inference about the parameter value is not **completely** governed by data alone; other knowledge we have also plays a role.

Bayesian Approach

- Thus, our final inference about the parameter value is not **completely** governed by data alone; other knowledge we have also plays a role.
- Though we consider it only for parameter estimation of density functions, the Bayesian approach is to be viewed as a generic approach for probabilistic modelling and inference.

Bayesian Approach

- Thus, our final inference about the parameter value is not **completely** governed by data alone; other knowledge we have also plays a role.
- Though we consider it only for parameter estimation of density functions, the Bayesian approach is to be viewed as a generic approach for probabilistic modelling and inference.
- The Bayesian approach is characterized by thinking of probabilities as also capturing subjective beliefs.

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.

- 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.

- In the above expression for $f(\theta \mid \mathcal{D})$, the denominator is not a function of θ. It is a normalizing constant and when we do not need its details, we will denote it by Z.

- Essentially, the posterior density is taken as the final Bayesian estimate.

- Essentially, the posterior density is taken as the final Bayesian estimate.
- An important question: how does one represent the posterior (and the prior) density?

- Essentially, the posterior density is taken as the final Bayesian estimate.
- An important question: how does one represent the posterior (and the prior) density?
- It would be nice if these densities can be represented in some parametric form.

- Essentially, the posterior density is taken as the final Bayesian estimate.
- An important question: how does one represent the posterior (and the prior) density?
- It would be nice if these densities can be represented in some parametric form.
- For that, we would like the prior and posterior densities to have the same general parametric form.

- 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 a conjugate prior, both prior and posterior belong to the same family of densities.

- When we use a conjugate prior, both prior and posterior belong to the same family of densities.
- Hence calculating posterior is essentially updating parameters of the density.

- When we use a conjugate prior, both prior and posterior belong to the same family of densities.
- Hence calculating posterior is essentially updating parameters of the density.
- We shall see many examples where this would be more clear.

- How do we use the final posterior density for implementing the classifier?

- How do we use the final posterior density for implementing the classifier?
- There are many possibilities for this.

- 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.

- 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

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

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

- 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.

- 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.

- 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)

- 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.

- 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.