Recap

- We have been considering Maximum Likelihood estimation.
- The ML estimate is the maximizer of likelihood (or log likelihood) function

$$\hat{\theta}_{ML} = \arg\max_{\theta} \prod_{i=1}^{n} f(x_i|\theta) = \arg\max_{\theta} \sum_{i=1}^{n} \ln(f(x_i|\theta))$$

- We saw that finding ML estimate is same as finding a distribution, f_θ, from the model class which is closest to to the empirical distribution of the data, f_{data}, in the sense of minimizing KL divergence

Recap

- We have considered the EM algorithm for ML estimation of a mixture density model.
- Mixture densities are useful models in many applications.
- The EM algorithm is a general method useful in many other situations as well.

Recap

- The general situation where EM is useful is as follows.
- The data that we have is ‘incomplete’
- This is because of some ‘hidden’ or ‘missing’ data.
- Given the complete data, ML estimation is easy.
- We ‘design’ what the missing data is. (We have a probability model for the complete data).

Recap

- For mixture density estimation, the given data, x_i is incomplete data.
- (x_i, z_i) constitutes the complete data where z_i indicates which mixture component x_i came from.
- We saw how to derive EM algorithm for this case.

The EM Algorithm

- The EM algorithm is an efficient iterative procedure for ML estimation in all such situations.
- The algorithm has two steps: 'Expectation' and 'Maximization'

Notation

- x_i, $i = 1, \cdots, n$, is the incomplete data and (x_i, z_i), $i = 1, \cdots, n$, is the complete data.
- $f(x, z \mid \theta)$ is the density for the complete data.
- The complete data log likelihood is

$$l(\theta \mid \mathcal{D}^c) = \ln(f(\mathbf{x}, \mathbf{z} \mid \theta)) = \ln\left(\prod_{i=1}^{n} f(x_i, z_i \mid \theta)\right)$$

- The two steps of EM algorithm are as follows:

E-step : Compute $Q(\theta,\ \theta^{(k)})$ which is expectation of the complete data loglikelihood w.r.t. the conditional distribution of hidden variables conditioned on incomplete data and current value of θ as $\theta^{(k)}$.

$$\begin{aligned} Q(\theta,\ \theta^{(k)}) &= E_{\mathbf{Z}|\mathbf{x},\theta^{(k)}} \ln(f(\mathbf{x}, \mathbf{Z} \mid \theta)) \\ &= \int \ln(f(\mathbf{x}, \mathbf{z} \mid \theta))\ f(\mathbf{z}|\mathbf{x}, \theta^{(k)}) d\mathbf{z} \end{aligned}$$

M-step : Compute next value of θ as $\theta^{(k+1)}$ by maximizing $Q(\theta,\ \theta^{(k)})$ over θ.

$$\theta^{(k+1)} = \arg\max_{\theta}\ Q(\theta,\ \theta^{(k)})$$

- Next question is: why does this procedure work?

Convergence of EM

- Our overall objective is to find ML estimate for θ.
- We want to maximize the log likelihood $\ln(f(\mathbf{x} \mid \theta))$.
- EM algorithm is an iterative technique for finding the maximum.
- We will now show that each iteration of the EM algorithm improves the log likelihood.
- This does not completely prove that the EM algorithm converges.
- However, under fairly general conditions, EM algorithm can be proved to converge to a local maximum of the log likelihood function.

- We have
$$f(\mathbf{z},\ \mathbf{x}|\theta') = f(\mathbf{x}|\theta')\ f(\mathbf{z} \mid \mathbf{x}, \theta')$$
- Using a better notation we write this as,
$$f_{\mathbf{ZX}}(\mathbf{z},\ \mathbf{x}|\theta') = f_{\mathbf{X}}(\mathbf{x}|\theta')\ f_{\mathbf{Z}|\mathbf{X}}(\mathbf{z} \mid \mathbf{x}, \theta')$$
- By taking log on both sides we can write this as
$$\ln(f_{\mathbf{X}}(\mathbf{x} \mid \theta')) = \ln(f_{\mathbf{ZX}}(\mathbf{z},\ \mathbf{x} \mid \theta'))\ -\ \ln(f_{\mathbf{Z}|\mathbf{X}}(\mathbf{z} \mid \mathbf{x}, \theta'))$$

$$\ln(f_{\mathbf{X}}(\mathbf{x} \mid \theta')) = \ln(f_{\mathbf{ZX}}(\mathbf{z},\ \mathbf{x} \mid \theta'))\ -\ \ln(f_{\mathbf{Z}|\mathbf{X}}(\mathbf{z} \mid \mathbf{x}, \theta'))$$

- Now take expectation with respect to the conditional distribution of $\mathbf{Z}$ conditioned on $\mathbf{x}$ and $\theta^{(k)}$.
- It is simple to see

$$E_{\mathbf{Z}|\mathbf{X},\,\theta^{(k)}}\ \ln(f_{\mathbf{X}}(\mathbf{x} \mid \theta')) = \ln(f_{\mathbf{X}}(\mathbf{x} \mid \theta'))$$

- By definition, we have

$$E_{\mathbf{Z}|\mathbf{X},\,\theta^{(k)}}\ \ln(f_{\mathbf{ZX}}(\mathbf{z},\ \mathbf{x} \mid \theta')) = Q(\theta',\ \theta^{(k)})$$

- Let

$$E_{\mathbf{Z}|\mathbf{X},\,\theta^{(k)}}\ \ln(f_{\mathbf{Z}|\mathbf{X}}(\mathbf{z} \mid \mathbf{x}, \theta')) = R(\theta',\ \theta^{(k)})$$

- Putting all these together we get

$$\ln(f_{\mathbf{X}}(\mathbf{x} \mid \theta')) = Q(\theta',\ \theta^{(k)})\ -\ R(\theta',\ \theta^{(k)}),\ \forall \theta', \forall k$$

- Note that

$$\begin{aligned} R(\theta',\ \theta^{(k)}) &= E_{\mathbf{Z}|\mathbf{X},\ \theta^{(k)}}\ \ln(f_{\mathbf{Z}|\mathbf{X}}(\mathbf{z} \mid \mathbf{x}, \theta')) \\ &= \int \ln(f_{\mathbf{Z}|\mathbf{X}}(\mathbf{z} \mid \mathbf{x}, \theta'))\ f_{\mathbf{Z}|\mathbf{X}}(\mathbf{z} \mid \mathbf{x}, \theta^{(k)})\ d\mathbf{z} \end{aligned}$$

- We show that: $R(\theta',\ \theta) \leq R(\theta,\ \theta),\ \forall \theta, \theta'$.
- This would imply that log likelihood improves in each iteration.

- We first show the follwing: for any two densities $p(z)$ and $q(z)$, we have

$$\int \ln(p(z))\, p(z)\, dz \;\geq\; \int \ln(q(z))\, p(z)\, dz$$

  Note that this is same as saying that KL divergence is non-negative. Recall

$$KL(p||q) = -\int p(z) \ln\left(\frac{q(z)}{p(z)}\right)\; dz$$

- For this we use Jensen's inequality: For any random variable Y and any convex function g,

$$E[g(Y)] \;\geq\; g\left(E[Y]\right)$$

- Take $Y = h(Z)$ and let $p(z)$ be density of Z. Then, for convex g,

$$\int g(h(z))\, p(z)\, dz \;\geq\; g\left(\int h(z)\, p(z)\, dz\right)$$

- Take $h(z) = q(z)/p(z)$ and $g(x) = -\ln(x)$

$$\begin{aligned}
\int -\ln\left(\frac{q(z)}{p(z)}\right) p(z)\, dz \;&\geq\; -\ln\left(\int \frac{q(z)}{p(z)}\, p(z)\, dz\right) \\
&= \; -\ln\left(\int q(z)\, dz\right) \\
&= \; 0
\end{aligned}$$

- This gives us

$$\int \ln(p(z))\, p(z)\, dz \;\geq\; \int \ln(q(z))\, p(z)\, dz$$

- We have

$$R(\theta,\ \theta') \;=\; \int \ln(f_{\mathbf{Z}|\mathbf{X}}(\mathbf{z} \mid \mathbf{x}, \theta))\, f_{\mathbf{Z}|\mathbf{X}}(\mathbf{z} \mid \mathbf{x}, \theta')\, d\mathbf{z}$$

- Hence we have

$$R(\theta,\ \theta') \leq R(\theta,\ \theta)$$

- To show that EM algorithm improves loglikelihood in each iteration we need to show

$$\ln(f_{\mathbf{X}}(\mathbf{x} \mid \theta^{(k+1)})) \;-\; \ln(f_{\mathbf{X}}(\mathbf{x} \mid \theta^{(k)})) > 0$$

- We showed earlier

$$\ln(f_{\mathbf{X}}(\mathbf{x} \mid \theta')) = Q(\theta',\, \theta^{(k)}) \;-\; R(\theta',\, \theta^{(k)}),\ \forall \theta', \forall k$$

- Using this

$$\begin{aligned}
\ln(f_{\mathbf{X}}(\mathbf{x} \mid \theta^{(k)})) &= Q(\theta^{(k)},\, \theta^{(k)}) \;-\; R(\theta^{(k)},\, \theta^{(k)}) \\
\ln(f_{\mathbf{X}}(\mathbf{x} \mid \theta^{(k+1)})) &= Q(\theta^{(k+1)},\, \theta^{(k)}) \;-\; R(\theta^{(k+1)},\, \theta^{(k)})
\end{aligned}$$

- Hence we have

$$\begin{aligned}
\ln(f_{\mathbf{X}}(\mathbf{x} \mid \theta^{(k+1)})) \;-\; \ln(f_{\mathbf{X}}(\mathbf{x} \mid \theta^{(k)})) &= \\
\left[Q(\theta^{(k+1)},\, \theta^{(k)}) \;-\; Q(\theta^{(k)},\, \theta^{(k)})\right] &+ \\
\left[R(\theta^{(k)},\, \theta^{(k)}) \;-\; R(\theta^{(k+1)},\, \theta^{(k)})\right]&
\end{aligned}$$

- The M-step in the EM algorithm would ensure

$$\left[Q(\theta^{(k+1)},\ \theta^{(k)})\ -\ Q(\theta^{(k)},\ \theta^{(k)})\right] > 0$$

(Recall $\theta^{(k+1)} = \arg\max_\theta Q(\theta, \theta^{(k)})$).

- By Jensen's inequality,

$$\left[R(\theta^{(k)},\ \theta^{(k)})\ -\ R(\theta^{(k+1)},\ \theta^{(k)})\right] \geq 0$$

- Thus we have

$$\ln(f_{\mathbf{X}}(\mathbf{x} \mid \theta^{(k+1)}))\ -\ \ln(f_{\mathbf{X}}(\mathbf{x} \mid \theta^{(k)})) > 0$$

showing that each iteration improves log likelihood.

- Hence each iteration of EM algorithm is like a gradient ascent step on log likelihood.
- Note that we only needed $Q(\theta^{(k+1)}, \theta^{(k)}) > Q(\theta^{(k)}, \theta^{(k)})$
- Hence, the M-step need not do full maximization. Gradient-based algorithm also would do.
- This analysis does not show convergence to a maximum.
- As mentioned earlier, the EM algorithm converges to a (local) maximum under fairly general conditions.
- Though this is a convergence only to local maximum of log likelihood, in practice it is quite good for estimating mixture densities.
- Since convergence is to a local maximum, initial conditions play a role.

- We started the analysis with

$$\ln(f_{\mathbf{X}}(\mathbf{x} \mid \theta')) = Q(\theta',\, \theta^{(k)}) \;-\; R(\theta',\, \theta^{(k)}),\;\forall \theta', \forall k$$

- We also showed that

$$R(\theta,\, \theta') \leq R(\theta,\, \theta),\;\forall \theta, \theta'$$

- Hence, we have

$$\ln(f_{\mathbf{X}}(\mathbf{x} \mid \theta')) \geq Q(\theta',\, \theta^{(k)}) \;-\; R(\theta^{(k)},\, \theta^{(k)}),\;\forall \theta', \forall k$$

- What we have is

$$\begin{aligned}\ln(f_{\mathbf{X}}(\mathbf{x} \mid \theta')) &\geq Q(\theta',\ \theta^{(k)}) - R(\theta^{(k)},\ \theta^{(k)}) \\ &= \int \ln(f_{\mathbf{ZX}}(\mathbf{z},\ \mathbf{x} \mid \theta')) f_{\mathbf{Z}|\mathbf{X}}(\mathbf{z} \mid \mathbf{x}, \theta^{(k)})\ d\mathbf{z} - \\ &\qquad \int \ln(f_{\mathbf{Z}|\mathbf{X}}(\mathbf{z} \mid \mathbf{x}, \theta^{(k)})) f_{\mathbf{Z}|\mathbf{X}}(\mathbf{z} \mid \mathbf{x}, \theta^{(k)})\ d\mathbf{z}\end{aligned}$$

- Hence $Q(\theta', \theta^{(k)})$ is a lower bound on the log likelihood at θ' and hence maximizing it would 'push-up' log likelihhod.
- Maximizing a lower bound to maximize log likelihood is also a general technique.

- In many situations involving missing data, hidden or latent variables etc. the EM algorithm is a popular method for maximum likelihood estimation of a model.
- It is most useful in learning mixture densities.
- It is also useful, for example, for learning probabilistic models such as HMMs, Graphical models etc.
- Identifiability is an issue in mixture estimation.

- Suppose our model is

$$f(x|\theta) = \lambda_1 f_1(x|\theta_1) + \lambda_2 f_2(x|\theta_2)$$

- If there exist $\lambda_i, \theta_i, \lambda'_i, \theta'_i$ such that

$$\lambda_1 f_1(x|\theta_1) + \lambda_2 f_2(x|\theta_2) = \lambda'_1 f_1(x|\theta'_1) + \lambda'_2 f_2(x|\theta'_2), \ \forall x$$

  then there is no unique solution.
- Even in the best case, we have uniqueness only upto permutations.
- But if there is a continuum of such solutions, it would be a serious problem in learning mixture desity models.

ML Estimation: Summary

- 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 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.
- 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.
- We call this the **prior** density of the parameter.
- Any information we may have about the value of parameter is to be captured through 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.
- Thus, our final inference about the parameter value is not **completely** governed by data alone; other knowledge we have also plays a role.
- The Bayesian approach is a generic approach for probabilistic modelling and inference.
- The Bayesian approach is characterized by thinking of probabilities as also capturing subjective beliefs or other knowledge about the unknown 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.
- 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.
- The posterior density is the final Bayesian estimate.

- 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.
- Otherwise we may be able to evaluate $f(x \mid \mathcal{D})$ at any x by sampling from 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.
- 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.

ML Vs MAP

- The posterior density is given by

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

- We have

$$\hat{\theta}_{\text{ML}} = \max_{\theta} f(\mathcal{D} \mid \theta)$$

$$\hat{\theta}_{\text{MAP}} = \max_{\theta} f(\mathcal{D} \mid \theta) f(\theta)$$

- If the prior is ‘flat’ both are same.

- 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.

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.
$$f(\theta \mid \mathcal{D}) = \frac{f(\mathcal{D} \mid \theta) f(\theta)}{Z}$$
- 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.
- Hence calculating posterior is essentially updating parameters of the density.
- We shall see some examples where this would become more clear.

Example

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

- We need to decide on the prior for μ.
- The posterior is

$$f(\mu \mid \mathcal{D}) = \frac{f(\mathcal{D} \mid \mu) f(\mu)}{Z}$$

- For conjugate prior we want $f(\mu)$ and $f(\mu \mid \mathcal{D})$ to have the same functional form.
- This depends on the form of data likelihood.

- 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.
- We want $f(\mu)$ such that when multiplied by $f(\mathcal{D} \mid \mu)$ we get the same form of function.
- 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 $f(\mu) = \mathcal{N}(\mu_0,\ \sigma_0^2)$.
- Quantities like μ_0, σ_0 are called *hyper-parameters*.
- 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

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

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

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

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

- As expected, the posterior is also Gaussian.

- Suppose $f(\mu \mid \mathcal{D})$ is $\mathcal{N}(\mu_n,\ \sigma_n^2)$. Then

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

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

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

- Now, comparing the expressions, we get

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

From these expressions we get

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

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

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

$$\mu_n = \sigma_n^2\left(\frac{\sigma_0^2 n\bar{\mu}_n + \sigma^2\mu_0^2}{\sigma^2\sigma_0^2}\right) \; = \frac{\sigma_0^2 n\bar{\mu}_n + \sigma^2\mu_0^2}{\sigma^2 + n\sigma_0^2}$$

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

- Thus we get

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

  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).

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

- 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.

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

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

- '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.
- If σ_0^2 is very large, then it is essentially same as MLE.

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

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

- The term inside the exp can be written as

$$
\begin{aligned}
& -\frac{(x-\mu)^2}{2\sigma^2} - \frac{(\mu-\mu_n)^2}{2\sigma_n^2} \\
= \; & -\frac{1}{2}\left\{\mu^2\left(\frac{1}{\sigma^2}+\frac{1}{\sigma_n^2}\right) - 2\mu\left(\frac{x}{\sigma^2}+\frac{\mu_n}{\sigma_n^2}\right) + \left(\frac{x^2}{\sigma^2}+\frac{\mu_n^2}{\sigma_n^2}\right)\right\} \\
= \; & \frac{-(\sigma_n^2+\sigma^2)}{2\sigma^2\sigma_n^2}\left[\mu^2 - 2\mu\frac{x\sigma_n^2+\mu_n\sigma^2}{\sigma^2+\sigma_n^2}\right] - \frac{1}{2}\frac{x^2\sigma_n^2+\sigma^2\mu_n^2}{\sigma^2\sigma_n^2}
\end{aligned}
$$

- Thus we have an integral of exp of quadratic w.r.t μ:

$$f(x \mid \mathcal{D}) = \int_{-\infty}^{\infty} K \exp(A)\, d\mu$$

  where

$$A = \frac{-(\sigma_n^2 + \sigma^2)}{2\sigma^2\sigma_n^2}\left[\mu^2 - 2\mu\frac{x\sigma_n^2 + \mu_n\sigma^2}{\sigma^2 + \sigma_n^2}\right] - \frac{1}{2}\frac{x^2\sigma_n^2 + \sigma^2\mu_n^2}{\sigma^2\sigma_n^2}$$

- Now we 'complete the square' in μ.

A Calculation Trick

$$
\begin{aligned}
I &= \int \exp\left(-\frac{1}{2K}\left[x^2 - 2bx + c\right]\right)\, dx \\
&= \int \exp\left(-\frac{1}{2K}\left[(x-b)^2 + c - b^2\right]\right)\, dx \\
&= \int \exp\left(-\frac{(x-b)^2}{2K}\right)\, \exp\left(-\frac{(c-b^2)}{2K}\right)\, dx \\
&= \exp\left(-\frac{(c-b^2)}{2K}\right)\sqrt{2\pi K}
\end{aligned}
$$

because

$$
\frac{1}{\sqrt{2\pi K}} \int \exp\left(-\frac{(x-b)^2}{2K}\right)\, dx = 1
$$

- We have the following inside exp in the integral w.r.t μ

$$\frac{-(\sigma_n^2+\sigma^2)}{2\sigma^2\sigma_n^2}\left[\mu^2-2\mu\frac{x\sigma_n^2+\mu_n\sigma^2}{\sigma^2+\sigma_n^2}\right]-\frac{1}{2}\frac{x^2\sigma_n^2+\sigma^2\mu_n^2}{\sigma^2\sigma_n^2}$$

- Now we 'complete the square' in μ.
- We end up with the remaining terms which can be seen to form a quadratic in x.

- This quadratic in x inside the exp is

$$
\begin{aligned}
& \frac{(\sigma_n^2+\sigma^2)}{2\sigma^2\sigma_n^2}\left(\frac{x\sigma_n^2+\mu_n\sigma^2}{\sigma^2+\sigma_n^2}\right)^2 - \frac{1}{2}\frac{x^2\sigma_n^2+\sigma^2\mu_n^2}{\sigma^2\sigma_n^2} \\
= \; & \frac{(x\sigma_n^2+\mu_n\sigma^2)^2-(\sigma^2+\sigma_n^2)(x^2\sigma_n^2+\sigma^2\mu_n^2)}{2\sigma^2\sigma_n^2(\sigma_n^2+\sigma^2)} \\
= \; & \frac{2x\mu_n\sigma^2\sigma_n^2 - x^2\sigma_n^2\sigma^2 - \mu_n^2\sigma_n^2\sigma^2}{2\sigma^2\sigma_n^2(\sigma_n^2+\sigma^2)} \\
= \; & -\frac{(x-\mu_n)^2}{2(\sigma_n^2+\sigma^2)}
\end{aligned}
$$

- Thus we showed that

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

- 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.

- This techniques of ‘completing squares’ is a general technique.
- We can use it with multidimensional Gaussians also.
- Here is a general result of which what we showed is a special case.

Calculations with Gaussian densities

- Suppose we have

$$\begin{aligned} f(z) &= \mathcal{N}(\mu, \Lambda^{-1}) \\ f(y|z) &= \mathcal{N}(Az + b, L^{-1}) \end{aligned}$$

- Then we get

$$\begin{aligned} f(y) &= \mathcal{N}(A\mu + b, L^{-1} + A\Lambda^{-1}A^T) \\ f(z|y) &= \mathcal{N}(\Sigma[A^T L(y - b) + \Lambda\mu], \Sigma) \end{aligned}$$

where $\Sigma = (\Lambda + A^T LA)^{-1}$.

Density of Binary Random variables

- 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

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

- 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

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

- We have $\Gamma(a+1) = a\Gamma(a)$ and $\Gamma(1) = 1$.
- For $n > 0$ integer, $\Gamma(n) = (n-1)!$.

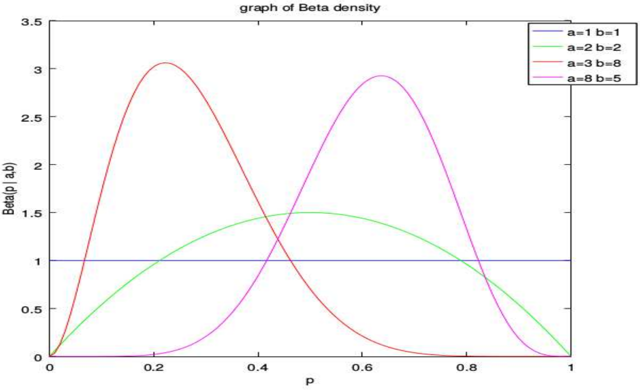
The Beta density

- The Beta(a, b) density is

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

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

Plot of Beta density



Mean and Mode of Beta density

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

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

- By differentiating we can easily show that its mode is at $\frac{a-1}{a+b-2}$.
- We can show that its mean is $\frac{a}{a+b}$ and its variance is $\frac{ab}{(a+b)^2(a+b+1)}$

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

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

- To show that this is a density we need to show

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

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

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

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

Now we interchange the order of integration.
We have x going from 0 to ∞ and for each x, t going from x to ∞.
To get same region in x–t space but with a changed order, we have t going from 0 to ∞ and x going from 0 to t.

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

In the inner integral change the variable from x to u as:
$x = tu$.
(When x goes from 0 to t, u goes from 0 to 1; $dx = tdu$).

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

Thus what we have is

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

This implies

$$\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\int_0^1 u^{a-1}\,(1-u)^{b-1}\, du = 1$$

This completes the proof that this is a density

We can find expected value of Beta density as follows

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

- Similarly one can show that

$$\text{Variance} = \frac{ab}{(a+b)^2(a+b+1)}$$

- 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$(\sum x_i + a,\ n + b - \sum x_i)$ density

- Suppose we want the MAP estimate.
- Recall posterior is Beta$(\sum x_i + a,\ n + b - \sum x_i)$ density
- 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.
- Since $x \in \{0,\ 1\}$, we need only $P(x = 1 \mid \mathcal{D})$.

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

- This is simply the mean of the posterior.

- Recall that the posterior density is Beta$(\sum x_i + a,\ n + b - \sum x_i)$.
- Hence we have

$$
\begin{aligned}
P[x = 1|\mathcal{D}] &= \frac{\sum_{i=1}^{n} x_i \ + \ a}{\sum_{i=1}^{n} x_i \ + \ a + n + b - \ \sum_{i=1}^{n} x_i} \\
&= \frac{\sum_{i=1}^{n} x_i \ + \ a}{n + a + b}
\end{aligned}
$$

- We can take this posterior mean as the Bayesian estimate

- 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.

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

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

- 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.

- Once again it turns out that the Bayesian estimate is a convex combination of MLE and prior estimate.
- Let $n_1 = \sum_{i=1}^{n} x_i$. (Number of 'heads' in data)
- Let $\alpha_0 = (a + b)$. (Sample size in prior)
- Let $m_1 = a/\alpha_0$. (Estimate from prior)

- We have

$$
\begin{aligned}
\hat{p}_B &= \frac{\sum_{i=1}^{n} x_i + a}{n + a + b} \\
&= \frac{n_1 + m_1\alpha_0}{n + \alpha_0} \\
&= \left(\frac{n}{n+\alpha_0}\right)\frac{n_1}{n} + \left(\frac{\alpha_0}{n+\alpha_0}\right) m_1 \\
&= \left(\frac{n}{n+\alpha_0}\right)\hat{p}_{\mathrm{ML}} + \left(\frac{\alpha_0}{n+\alpha_0}\right)\hat{p}_{\mathrm{prior}}
\end{aligned}
$$

- In document classification, 'bag of words' representation involves estimating a Bernoulli density.
- The parameters could be: θ_{jc} – probability that j^{th} word is present given document is from class c.
- If we want a Bayesian approach, we would use this 'Beta-Bernoulli model'.
- This will ensure that none of the Bernoulli parameters become 1 or 0 (unlike the case with ML).
- We can use same hyperparameters for the prior for all parameters. (For example, $a = b = 1$ and posterior mean as the estimate).