LECTURE 7: STATISTICAL CONCEPTS (CONTINUED)

COS 513, FALL 2010
LECTURER: DAVID BLEI
SCRIBE: ANURADHA

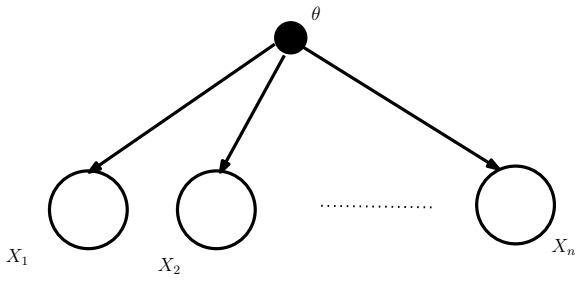
Given n data points, $X_1, X_2, \ldots, X_n$ that are known to be drawn from a Gaussian distribution, how do we estimate the distribution? The Gaussian distribution is completely parametrized by the mean and the variance of the distribution. If X is a random variable with a Gaussian distribution then,

$$p(X|\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left\{-\frac{1}{2\sigma^2}(x-\mu^2)\right\}$$

where μ and σ^2 are the mean and the variance respectively and we call these parameters $\theta \triangleq \{\mu, \sigma^2\}$. Our problem is to estimate θ from the data $X \triangleq X_{1:N}$.

Maximum Likelihood Estimation

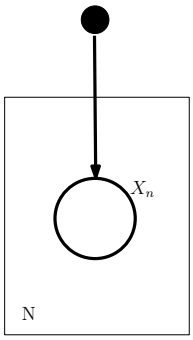
The graphical model for this problem is -



where the small black circle represents the parameter. This can be, more succinctly, represented by the plate model as -

From the Bayes Ball algorithm we know that given θ, the X_is are independent. Also, the X_is are identically distributed. Hence, $X_1, X_2, \ldots, X_n$ are IID (independently and

Date: Oct. 11, 2010.



identically distributed).

$$\begin{aligned} p(X|\theta) &= \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} exp\left\{-\frac{1}{2\sigma^2}(x_n - \mu^2)\right\} \\ &= \frac{1}{\left(\sqrt{2\pi\sigma^2}\right)^N} exp\left\{-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n - \mu^2)\right\} \end{aligned}$$

To estimate θ, we interpret the likelihood function as a function of θ and maximise the log likelihood (call it $l(\theta; x)$) -

$$\log p(X|\theta) \triangleq \; l(\theta; x) = \; -\frac{N}{2}\log 2\pi - \frac{N}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n - \mu)^2$$

The Maximum likelihood estimate (MLE) of θ is the θ that optimizes $l(\theta; x)$

$$\begin{aligned} \frac{\partial l}{\partial \mu} &= \; -\frac{1}{2\sigma^2}\sum_{n=1}^{N} 2(x_n - \mu)(-1) \\ &= \frac{1}{\sigma^2}\sum_{n=1}^{N}(x_n - \mu) \end{aligned}$$

Setting the derivative to 0 and solving for μ,

$$\begin{aligned} \frac{1}{\sigma^2}\sum_{n=1}^{N}(x_n - \mu) \; &= \; 0 \\ \hat{\mu}_{ML} \; &= \; \frac{1}{N}\sum_{n=1}^{N} x_n \end{aligned}$$

The MLE of the mean of the Gaussian is just the mean of the data! Solving for σ^2,

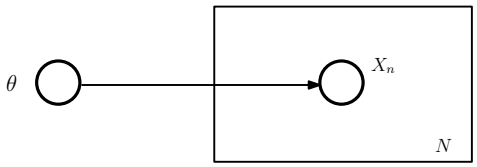
$$\frac{\partial l}{\partial \sigma^2} = -\frac{N}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{n=1}^{N}(x_n - \mu)^2 = 0$$

$$\hat{\sigma^2}_{MLE} = \frac{1}{N}\sum_{n=1}^{N}(x_n - \hat{\mu}_{ML})^2$$

which is just average squared distance of the data points from the sample mean.

Bayesian Computation

In the Bayesian case, the parameter is thought of as a random variable and we have the following model,

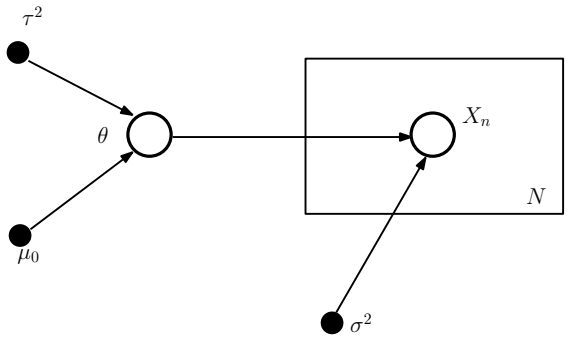


Suppose the variance σ^2 is fixed, the posterior of the mean is

$$\begin{aligned} p(\mu|x_{1:N}) &\propto p(\mu, X) \\ &\propto p(\mu)p(X_{1:N}|\mu) \end{aligned}$$

Modulo normalization, the posterior is the product of the prior distribution of μ and the probability of the data under the Gaussian assumption. But what is the prior distribution of μ? We can come up with several candidate distributions and even use the data for this purpose. Here, we shall assume the prior to have a Gaussian distribution as well. When the posterior and the prior distributions are from the same family of distributions, the prior is called a conjugate prior. Let $\mu \sim N(\mu_0, \tau^2)$. Ideally we should treat the parameters μ_0 and τ^2 as random variables and obtain their distribution. These parameters are called hyperparameters. We could go up another level and put a prior on these hyperparameters and continue for more levels. However, there is no good way to set these hyperparameters and we simply assume μ_0 and τ^2 to be arbitrary constants. As we go higher up the levels of hyperparameters of prior distributions, the influence of any assumption about the hyperparameters on the inference decreases. We shall also see that as we have large amounts of data, the influence of the hyperparameters diminishes.

We now have the following model : $\mu \sim N(\mu_0, \tau^2)$, $X_i \sim N(\mu, \sigma^2)$



$$p(\mu) = \frac{1}{\sqrt{2\pi\tau^2}} exp\left\{-\frac{1}{2\tau^2}(\mu-\mu_0)^2\right\}$$
$$p(X_{1:N}|\mu) = \frac{1}{\left(\sqrt{2\pi\sigma^2}\right)^N} exp\left\{-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu^2)\right\}$$

From the Bayes Ball rules, we know the X_is need not be independent. However, conditioned on μ, they are independent and we can compute $p(X_{1:N}|\mu)$. This gives the joint probability

$$p(X,\mu) = \frac{1}{(2\pi\sigma^2)^{N/2}} exp\left\{-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu^2)\right\} \frac{1}{(2\pi\tau^2)^{1/2}} exp\left\{-\frac{1}{2\tau^2}(\mu-\mu_0)^2\right\}$$

which on normalizing gives $p(\mu|x_{1:N})$. Doing the algebra gives -

$$p(\mu|x_{1:N}) = \frac{1}{\sqrt{2\pi\tilde{\sigma}^2}} exp\left\{-\frac{1}{2\tilde{\sigma}^2}(\mu-\tilde{\mu})^2\right\}$$

where,

$$\tilde{\mu} = \frac{N/\sigma^2}{N/\sigma^2+1/\tau^2}\bar{x} + \frac{1/\tau^2}{N/\sigma^2+1/\tau^2}\mu_0$$
$$\tilde{\sigma}^2 = \left(\frac{N}{\sigma^2}+\frac{1}{\tau^2}\right)^{-1}$$

and $\bar{x}$ is the sample mean.

We see that the posterior mean $\tilde{\mu}$ is the weighted average of the sample mean and the prior mean μ_0 and $\tilde{\mu}$ approaches $\hat{\mu}_{ML}$ as N approaches infinity. As we see more data, we rely more on the data than the hyperparameters. Our estimate of the variance $\tilde{\sigma}^2$ goes to 0 for large N. This matches the intuition that the uncertainty in the estimate of the mean diminishes when we have more evidence.

An aside : De Finetti's theorem. Notice that $p(X_{1:N}|\mu)$ was also computed in the maximum likelihood estimate. There we thought of the X_is as IID while here we only assume that they are independent conditioned on μ. It can be seen from the following theorem that this is a much weaker assumption

Fact 1 (De Finetti's Theorem)**.** *If $X_1, X_2, \ldots, X_n$ are exchangeable, then*

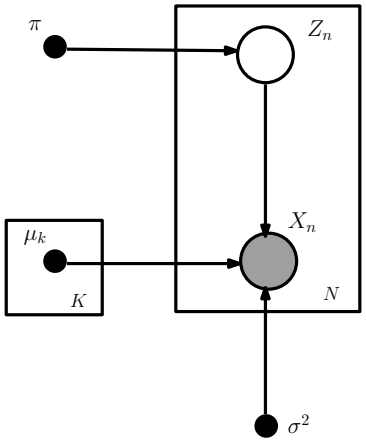
$$p(X_1, X_2, \ldots, X_n) = \int_\theta p(\theta) \prod_n p(X_n|\theta) d\theta$$

Thus, if the random variables are exchangeable i.e if their distribution is invariant under permutations then they are independent conditioned on a parameter random variable. So, in the Bayesian approach to estimating μ, we only need the assumption that the X_is are exchangeable meaning the order in which they come doesn't matter.

Posterior inference may be hard

The posterior inference may be difficult to compute in general and we may have to settle for an approximate inference. We now look at an example where inference is hard.

Consider a mixture model where the data points $x_{1:N}$ come from one of 2 Gaussians and it is not known which Gaussian each data point comes from. We model the problem as -



Here, the data points come from one of K Gaussian distributions having means $\mu_{1:K}$ and variance σ^2. Z_n is a hidden variable which tells us which distribution, X_n is drawn from. Z_n has a multinomial distribution with parameter π where π is a fixed distribution over $\{1, 2, \ldots, K\}$. Thus, $Z_n \sim Mult(\pi)$ and $X_n \sim N(\mu_{Z_n}, \sigma^2)$. If we put a prior on the μ_is then inferring their posterior distribution is hard i.e computing $p(\mu_i|x_{1:N})$ for any $i \in [k]$ is hard. Why? We shall see this in the next class.