1. The MLE of $\theta$ is given by $\frac{1}{n} \sum_{i=1}^{n} x_i$.

2. The MLE of $\theta$ is just the maximum element in the given set of data. This result is interesting because the MLE may not need to be revised in light of new data (which is not the case in the previous question).

3. In the notation given in the question, the prior is normal with a mean of $\theta_0$ and variance $\sigma_0$, and the data comes from another normal distribution with mean $\theta$ and variance $\sigma$. $\theta$ is unknown here. We can write the posterior for $\theta$ as:

$$P(\theta|x) \propto P(x|\theta) \times P(\theta)$$

$$= \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\theta)^2}{\sigma^2}} \frac{1}{\sqrt{2\pi\sigma_0}} e^{-\frac{(\theta-\theta_0)^2}{2\sigma_0^2}}$$

$$\propto e^{-\frac{\theta^2}{2\left(\frac{1}{\sigma} + \frac{1}{\sigma_0}\right)} + \theta\left(\frac{\theta_0}{\sigma_0} + \frac{x}{\sigma}\right)}$$

$$\propto e^{-\frac{(\theta-\mu)^2}{2\alpha}}$$

Notice that we do not really care about the proportionality constants at this point; this is because we can always fix them later by noticing that the pdf must integrate to 1 (cute, isn’t it?). Now, the two parameters introduced in the last step are given by:

$$\alpha = \frac{1}{\sigma_0 + \frac{1}{\sigma}}$$

$$\mu = \alpha\left(\frac{\theta_0}{\sigma_0} + \frac{x}{\sigma}\right)$$

The final distribution is thus normal with mean $\mu$ and variance $\alpha$. The above Bayesian update can be repeated again, over new data, to yield new posteriors, all of which will be normal. Do not get confused by the form of the normal distribution in the above equations; in class we used $\sigma^2$ to refer to the variance, whereas this question uses $\sigma$.

4. Given that measurement errors are normally distributed (around zero) and assuming independence of observations, maximum likelihood estimation reduces to finding a least squares model fit. You will find that the cubic has the best likelihood, followed by the quadratic, then the linear, and constant. You should keep in mind that this ranking does not capture the true spirit of the Bayesian decision making paradigm where we would have marginalized over all possible ranges of the unknown parameters, rather than pick one best value. The AIC and BIC criteria explicitly account for model complexity and actually flip the quadratic and the cubic in their rankings. Thus the quadratic will be chosen as the best model.

It is now time to reveal how the data was generated! We used the following MATLAB code, which has constant, linear, and cubic terms (sometimes overshadowed in magnitude by the error):

```matlab
a = []; for i=0:0.01:1
```
5. You are expected to estimate 12 bimodal distributions (4 attributes \( \times \) 3 classes) for the Iris dataset. Not all of the corresponding datasets have a bimodal nature, however, and you might encounter difficulties in estimation. This problem is one of identifiability and basically means that there are not enough dimensions in the data to identify all the degrees of freedom posited by the model. Recall that EM estimation only conducts a local search.

The purpose of fitting models to such messy data is so that we can use the model parameters for further analysis. Given the means and variances, for instance, we can then attempt to understand the relationship between the features and the class attribute. It is well known that Iris-Setosa is linearly separable from Iris-Versicolor and Iris-Virginica, whereas the latter two are not linearly separable. The estimated Gaussian components should demonstrate when the projections of the classes overlap. In the ideal case, we would have estimated a truly four-dimensional Gaussian mixture model, with all features taken together, and including covariance terms for the interactions between the dimensions. As such a study will reveal, two of the classes are indeed not linearly separable.