CS480/680 – Spring 2019
Assignment 3 solutions

1. Question 1

$$
\begin{aligned}
k(x,z) &= exp\left(-\frac{||x-z||^2}{2\sigma^2}\right) \\
&= exp\left(-\frac{x^T x}{2\sigma^2}\right) exp\left(-\frac{z^T z}{2\sigma^2}\right) exp\left(\frac{x^T z}{\sigma^2}\right) \\
&= exp\left(-\frac{x^T x}{2\sigma^2}\right) exp\left(-\frac{z^T z}{2\sigma^2}\right) \sum_{j=0}^{\infty} \frac{1}{j!}\left(\frac{x^T z}{\sigma^2}\right)^j \\
&= exp\left(-\frac{x^T x}{2\sigma^2}\right) exp\left(-\frac{z^T z}{2\sigma^2}\right) \left[\sum_{j=0}^{\infty} \frac{1}{\sigma^{2j} j!} \sum_{i_1=0}^{j} ... \sum_{i_j=0}^{j} (x_{i_1}...x_{i_j})(z_{i_1}...z_{i_j})\right] \\
&= \sum_{j=0}^{\infty} \left[\frac{(x_{i_1}...x_{i_j})}{\sigma^j \sqrt{j!}} exp\left(-\frac{x^T x}{2\sigma^2}\right)\right] \left[\frac{(z_{i_1}...z_{i_j})}{\sigma^j \sqrt{j!}} exp\left(-\frac{z^T z}{2\sigma^2}\right)\right]
\end{aligned}
$$

where $\sum_{j=0}^{\infty} \frac{1}{\sigma^{2j} j!} \sum_{i_1=0}^{j} ... \sum_{i_j=0}^{j} (s_{i_1}...s_{i_j}) = 1$ if $j = 0$. Therefore, computing the Gaussian kernel is equivalent to taking the inner product after mapping the input to an infinite dimensional feature space, where each element of the mapping $\phi(x)$ has the form

$$
\frac{(x_{i_1}...x_{i_j})}{\sigma^j \sqrt{j!}} exp\left(-\frac{x^T x}{2\sigma^2}\right)
$$

2. Question 2

   (a) The graph for generalized regularized linear regression is given in Figure 1. The best degree is 3 and the test set mean squared error is 0.064593. The running time increases exponentially as the degree increases.

   (b) The graph for Bayesian generalized linear regression is given in Figure 2. The best degree is 3 and the test set mean squared error is 0.064593. The running time increases exponentially as the degree increases. Briefly, the two models are the same because we start with a prior having an identity co-variance matrix, output variance being 1 and the regularization term was 1. The major difference between the models is that while generalized regularized linear regression returns point estimates, Bayesian generalized linear regression returns a probability distribution over the estimates. The point estimate is taken to be the mean of this probability distribution.

   (c) The mean squared error for the identity kernel is 3.5906 (or 1.29) depending on the implementation. Both answers are correct. The graphs for the Gaussian and polynomial kernels are shown respectively in Figures 3 and 4. For the Gaussian kernel, the best σ is 4 and the test set mean squared error is 0.2587. For the polynomial kernel, the best degree is 4 and the test set mean squared error is 0.04577. The running time is independent of the number of basis functions. Instead it grows cubically with the amount of data.

(d) The graph for neural network regression is given in Figure 5. The best number of hidden units is 10 and the test set mean squared error is 0.10318. The running time grows linearly with the number of hidden units.

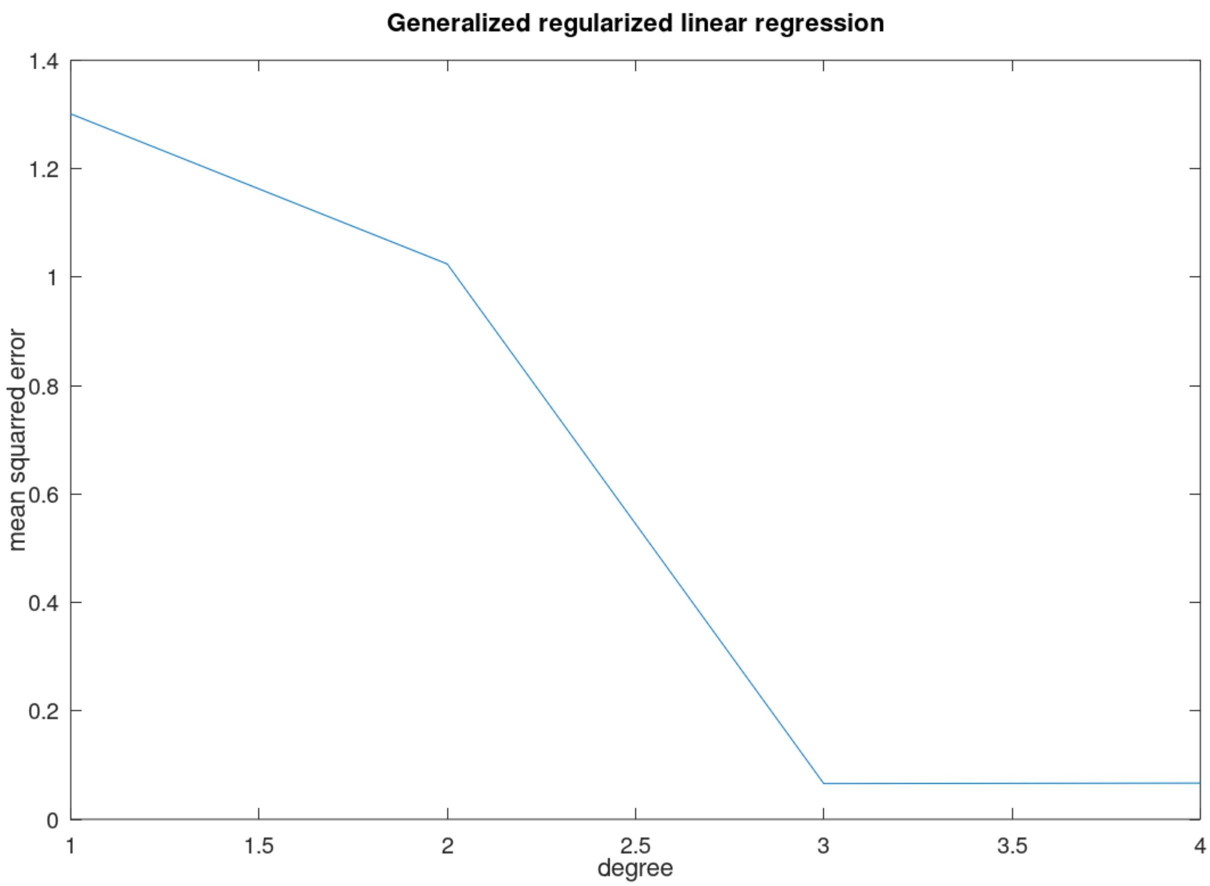

Figure 1: Generalized regularized linear regression

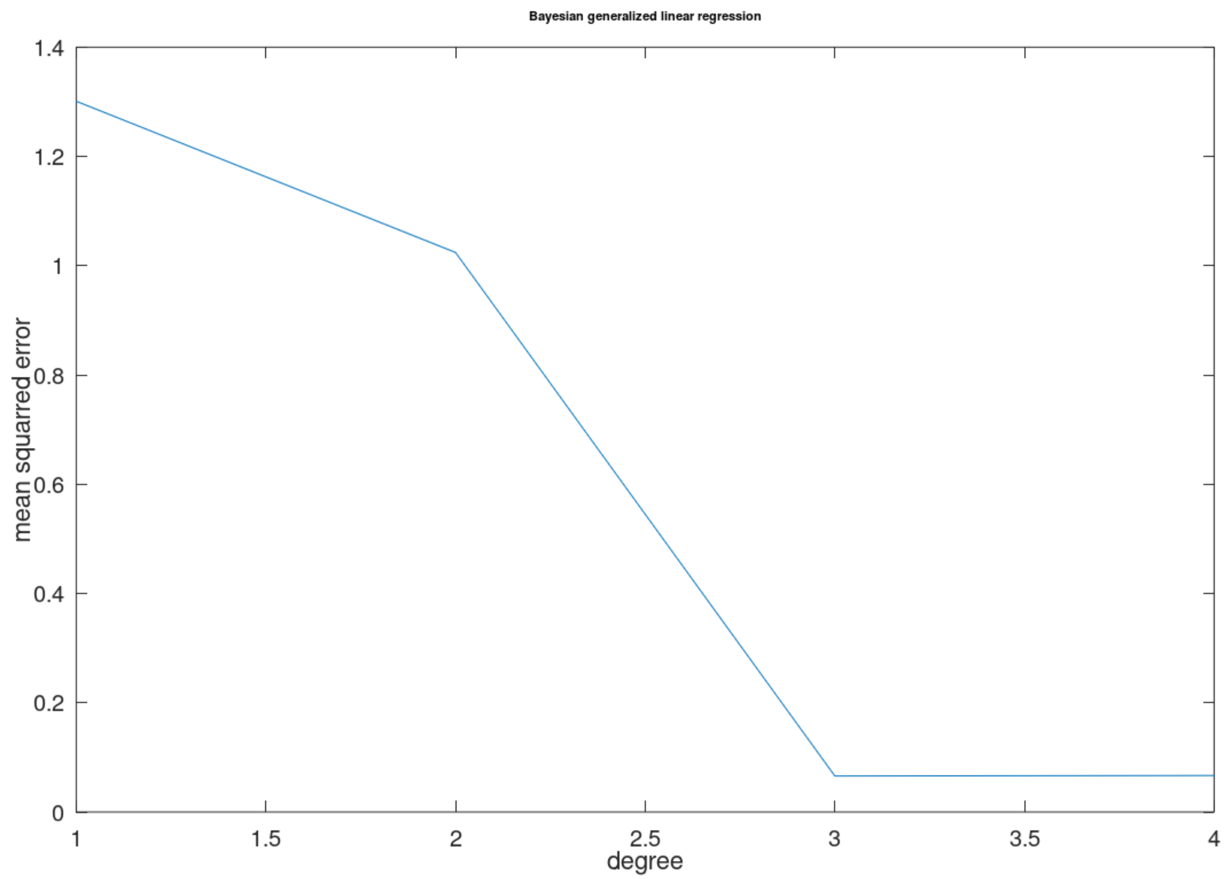

Figure 2: Bayesian generalized linear regression

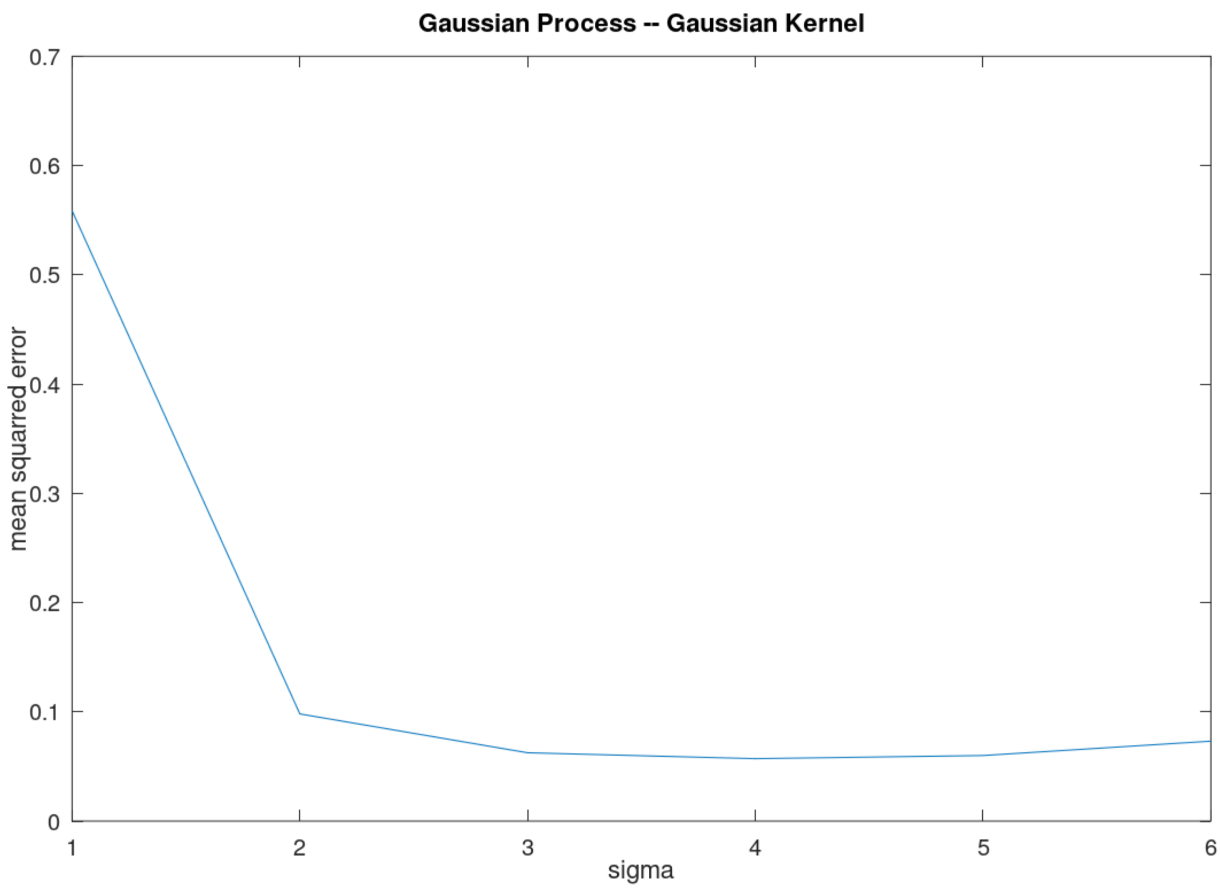

Figure 3: Gaussian Process – Gaussian Kernel

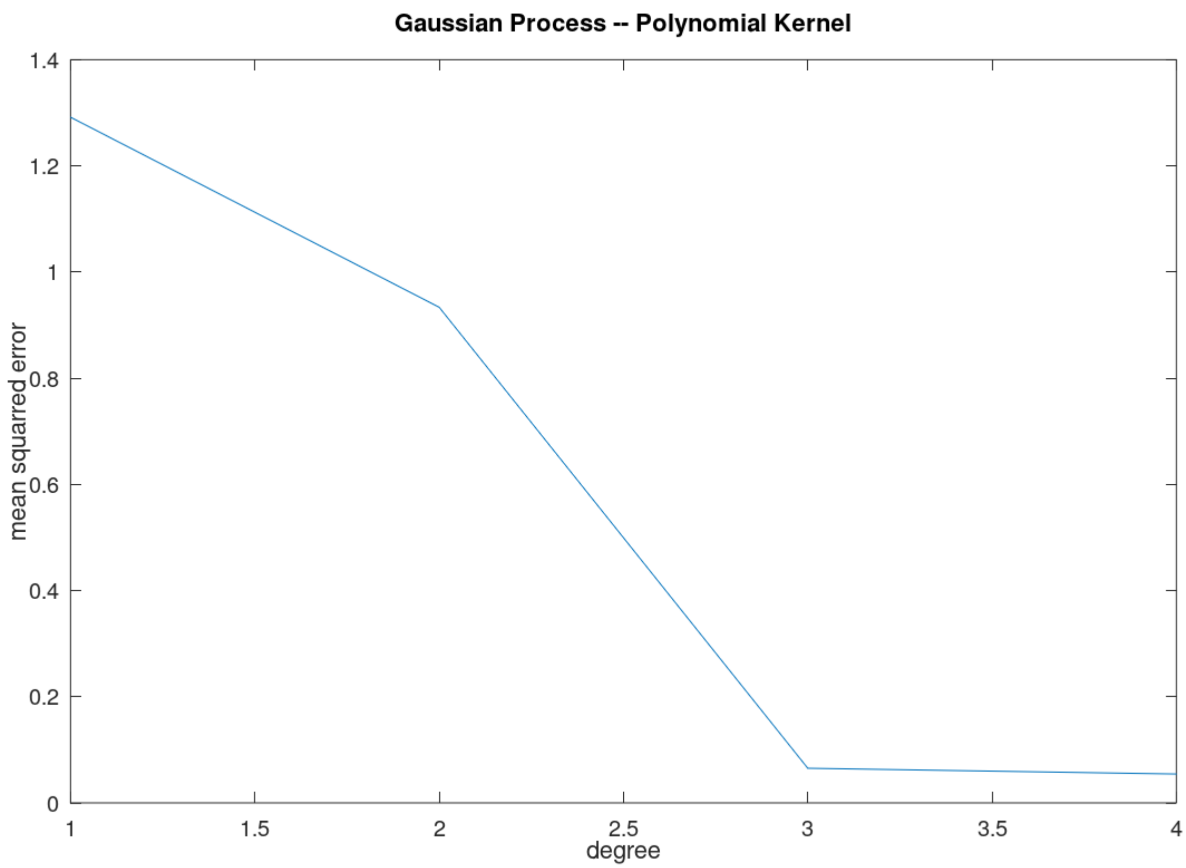

Figure 4: Gaussian Process - Polynomial Kernel

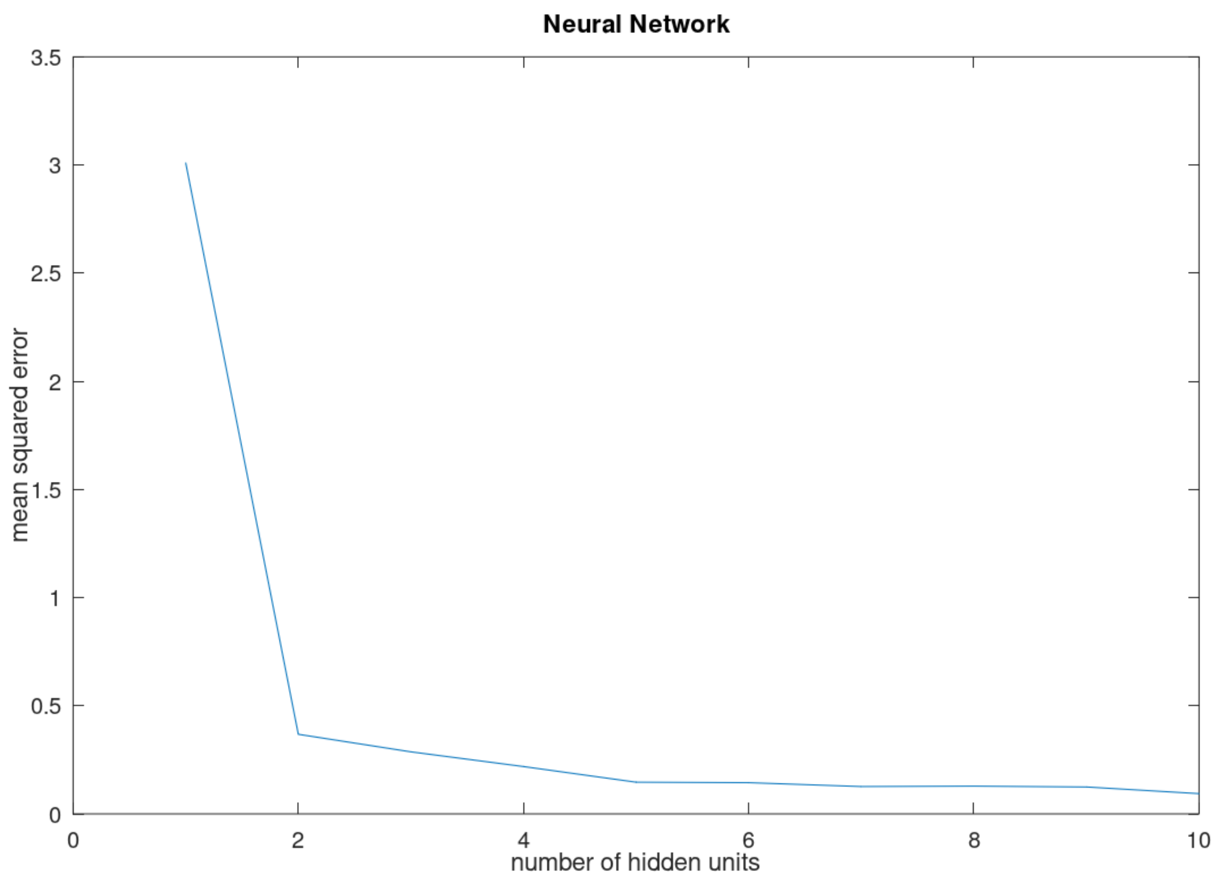

Figure 5: Neural Network