CS195f Homework 4

Mark Johnson and Erik Sudderth

Homework due at 2pm, 20th October 2009

In this problem set, we study different approaches to linear regression using a one-dimensional dataset collected from a simulated motorcycle accident. The input variable, x, is the time in milliseconds since impact. The output variable, y, is the recorded head acceleration. The dataset is available here:

`/course/cs195f/asgn/regression/motor.mat`

We have divided the full dataset into 40 training examples (variables `Xtrain` and `Ytrain`), and 53 test examples (variables `Xtest` and `Ytest`).

When fitting polynomial functions, as explored below, numerical problems can arise when the input variables take even moderate values. To minimize these, all training features should be scaled to lie in the interval $[-1, +1]$ before fitting. Note that an equivalent scaling must then be applied to all test data. Here is an example script to get you started:

`/course/cs195f/asgn/regression/motorDemo.m`

Question 1:

a) *Consider a polynomial basis, with functions $\phi_j(x) = x^j$. Write a function which evaluates these polynomial functions at a vector of points $x_i \in \mathbb{R}$, for any j. In a single figure, plot $\phi_j(x)$ for $-1 \leq x \leq 1$, and $j = 0, 1, 2, \ldots, 19$.*

   Hint: *To create a dense regular grid of points at which to evaluate and plot these functions, use the* `linspace` *command.*

b) *Consider the standard linear regression model, in which observations y_i follow a Gaussian distribution centered around a linear function w of a fixed set of basis functions:*

$$p(y_i \mid x_i, w, \beta) = \mathcal{N}(y_i \mid w^T \phi(x_i), \beta^{-1})$$

   Here, β is the inverse variance or precision. Define a family of regression models, each of which contains all polynomials $\phi_j(x)$ of order $j \leq M$, where M is a parameter controlling model complexity. Compute maximum likelihood (ML) estimates $\hat{w}$ of the regression parameters for models of order $M = 0, 1, 2, \ldots, 19$. Plot, as a function of x, the mean prediction $\hat{w}^T \phi(x)$ for models of order $M = 0, 1, 3, 5, 19$.

   Hint: *To compute $x = A^{-1}b$ in Matlab, rather than explicitly calling the* `inv` *command, use the following command to improve numerical stability:*

```
>> x = A \ b;
```

c) *Consider the following error metric, which is defined for any set of N points (x_i, y_i):*

$$L(x, y \mid \hat{w}) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{w}^T\phi(x_i))^2}$$

What is the relationship between this quantity and the ML estimate of the inverse error variance, $\hat{\beta}$? Evaluate and plot $L(x, y \mid \hat{w})$ as a function of the model order, M, for the 40 training examples. Also do this for the 53 test examples. Which model has the smallest training error, and which has the smallest test error? Together with the test data, plot the mean prediction $\hat{w}^T\phi(x)$ for both of these models.

d) *In constructing the training and test sets, we excluded one point from the original dataset: $x = 57.6$, $y = 10.7$. What is the error in the prediction of this point for the two models selected in part (c)? Discuss any qualitative differences between this datapoint and the other test data.*

e) *We now consider an alternative, radial basis function basis of the form*

$$\phi_j(x) = \exp\left(-\frac{(x-\mu_j)^2}{2\sigma^2}\right) \qquad j = 1, \ldots, M$$

For any model order M, we space the basis function centers μ_j evenly between -1 and 1. In Matlab, this can be done with the following command:

```
>> mu = linspace(-1,1,M);
```

We then set the bandwidth to triple the distance between basis centers, $\sigma = 3(\mu_2 - \mu_1)$. Finally, for any M, we also include a constant bias term $\phi_0(x) = 1$. In three figures, plot these basis functions for $-1 \leq x \leq 1$, and for models of order $M = 5, 10, 15$.

f) *Repeat parts (b-d) for the radial basis function basis of part (e), and models of order $M = 5, 10, 15, 20, 25, 30$. Which basis performs better for this data?*

In the previous question, you may have noticed that unregularized ML estimates can become unstable for large model orders, M. We now consider an alternative, Bayesian approach in which the regression coefficients are assigned a Gaussian prior

$$p(w) = \mathcal{N}(w \mid 0, \alpha^{-1}I_M)$$

where α is a hyperparameter discussed further below.

Question 2:

a) *What is the MAP estimate of w under the Gaussian prior above, and linear observation model of part 1(b)? Consider radial basis function and polynomial bases of orders $M = 100$. For each of these two models, determine the MAP estimate $\hat{w}$ assuming hyperparameter values of $\alpha = 0.01$, $\beta = 0.0025$. Plot the mean prediction $\hat{w}^T\phi(x)$ for both of these models. Would it be possible to compute ML estimates for models of order $M = 100$?*

b) *Fix $\beta = 0.0025$, and consider 100 candidate values for the regularization parameter α, logarithmically spaced between 10^{-8} and $10^0 = 1.0$:*

```
>> alpha = logspace(-8,0,100);
```

Using the `semilogx` *command, plot the error metric of problem 1(c) versus α for both the training and test datasets, and both basis families. Then plot the mean prediction $\hat{w}^T\phi(x)$ for the models which minimize the training and test error, for each basis family.*

c) *Consider the two models from part (b) which minimize the test error for their corresponding basis families. Given these hyperparameters and the training data, what are the corresponding posterior distributions over the prediction function $f(x) = w^T\phi(x)$? Draw and plot 10 samples from each of these two distributions.*

Hint: *The most efficient way to sample the specified functions is to determine the posterior distribution of w, draw samples $\tilde{w}$ from it, and then determine corresponding prediction functions $\tilde{f}(x) = \tilde{w}^T\phi(x)$ via the appropriate linear transformation. In Matlab, to draw a sample from a multivariate normal distribution with mean* `mu` *and covariance* `Sigma`*, use the following command:*

```
>> SigmaRoot = chol(Sigma, 'lower');
>> wSamp = mu + SigmaRoot * randn(size(mu));
```

d) *Again consider the pair of models from part (c). What is the error in their corresponding predictions of the held-out test point from problem 1(d)? Are the relative magnitudes of these errors predictable from the posterior distributions plotted in part (c)?*

e) *Does the "soft" regularization approach explored in this question seem more or less effective than the model selection approach of question 1? Why?*

In the previous questions, we compared the accuracy of various models on test data, but did not provide a mechanism for choosing among models given solely training data. Cross-validation methods provide one popular, but computationally intensive, solution to this problem. The following question instead explores a Bayesian approach to model selection.

Question 3: (200-level credit)

a) *In Sec. 3.5.1 of Bishop's textbook,* Pattern Recognition and Machine Learning*, the marginal likelihood of the training data is shown to take the following form:*

$$p(y \mid \alpha, \beta) = \frac{M}{2}\log\alpha + \frac{N}{2}\log\frac{\beta}{2\pi} + \frac{1}{2}\log|S_N| - \frac{\alpha}{2}||m_N||^2 - \frac{\beta}{2}\sum_{i=1}^{N}(y_i - m_N^T\phi(x_i))^2$$

Here, m_N and S_N are the posterior mean and variance of the weight vector, w, given N observations. Plot this quantity as a function of α, for the pair of basis families and range of hyperparameter values considered in problem 2(b). For each model family, what is the test accuracy for the hyperparameters which maximize the marginal likelihood of the training data? How do these compare to the models which actually performed best in problem 2(b)?

Hint: *To avoid numerical underflow when computing the marginal likelihood above, you can exploit the following identity:*

$$\log |S_N| = \log \prod_{j=1}^{M} \lambda_j = \sum_{j=1}^{M} \log \lambda_j \qquad\qquad S_N u_j = \lambda_j u_j$$

Here, λ_j are the eigenvalues of the covariance matrix S_N.

b) Suggest an alternative family of basis functions for this regression problem. As in problem 1(a,e), plot examples of this family.

c) Repeat part (a) for the basis family proposed in part (b). How does this new family perform on test data, relative to the radial basis function and polynomial bases? Is this relative performance accurately predicted by the marginal likelihood?