CMPS 242: HW3

December 31, 2018

1. Linear Regression
   Generate a 100-example training set as follows. For each example, draw feature x uniformly at random from $[100, 100]$, and label t from $N(f(x), 1)$ where $f(x) = 0.2 + 2x + x^2 + 3x^3$ (remember to seed your random number generator). To fit this curve we will use the probabilistic model:

   $$p(t|x, w, \beta) = N(t|y(x, w), \beta^{-1})$$

   where $y(x, w) = w_0 + w_1 x + w_2 x^2 + w_3 x^3$.

   (a) Perform MLE estimation of w and β. You may use the optimize module from scipy for this task. Give the estimates and describe how well they match the true parameters used to generate the data. Use 1000 and 10,000 training points (instead of the original 100). How do the parameter estimates change?

   (b) Repeat the first part, but consider additional parameters in the model. In particular, create the data as before, but estimate the parameters for a function $y(x, w) = w_0 + w_1 x + w_2 x^2 + w_3 x^3 + w_4 x^4 + w_5 x^5$. Comment on what changes.

   (c) Now consider estimating a $w = (w_0, w_1, w_2, w_3)$ using Bayesian linear regression (see section 3.3.1 of the text). Set the hyperparameter $\alpha = 1$ (as used in equation 3.52) and prior mean $m_0 = 0$, so the prior on w is (refer to (3.48):

   $$p(w)N(w|0, I).$$

   The w vector maximizing the posterior is the maximizer of (3.55), which is a least-squares problem with a 2-norm regularizer. With hyperparameter $\beta = 1$, find the w vector maximizing this poster for 100, 1000, and 10,000 training examples and plot these three curves as well as the true function $f(x) = 0.2 + 2x + x^2 + 3x^3$. Repeat for 5th degree polynomial estimation using the larger parameter vector $w = (w_0, w_1, w_2, w_3, w_4, w_5)$. How do the curves for the 5th degree model compare with those of the third degree model?

2. Naive Bayes Text Classification Download the Enron-spam dataset from http://www.aueb.gr/users/ion/data/enron-spam/. Use the pre-processed datasets Enron1, . . . , Enron5 as training data and Enron6 as test data. Implement the Naive Bayes algorithm that we discussed in the class (one reference is http://nlp.stanford.edu/IR-book/pdf/13bayes.pdf). Remember to perform all calculations on the log scale to prevent underflow.

(a) Report any decisions made, such as how you decided to handle punctuation.

(b) Report the accuracy on the test set.

(c) How do you account for different prior probabilities for spam and ham?

(d) Implement versions with and without Laplace smoothing (adding a fictitious observation of each word to each class)? How does the performance of the classifier change when Laplace smoothing is added?

(e) What are the most discriminative words based on the learned probabilities?