TDA 231 Machine Learning: Homework 4

Instructor: Chiranjib Bhattacharyya and Devdatt Dubhashi

Due Date: March 9, 2012

Goal: Regression, EM

1. (2 points) Consider dataset *q1.mat*. It has 100 examples of 2 dimensional data (X), with corresponding output (Y). Use matlab command *mvregress* to perform linear regression of X on Y. Implement SVM regression using matlab *quadprog* with linear kernel, choosing $C = 1.0$ and $\epsilon = 0.1$
   (a) Submit code for both regressions.
   (b) Submit in a table the regression coefficient w as well the error in the residuals ($norm(y - \hat{y}, 2)$) obtained for cases.

2. (2 points) Consider dataset *q2.mat*. It has 100 examples of 2 dimensional data (X), with corresponding output (Y). Repeat the above question, with original X and alternate feature set

$$\phi(x) = [1, x_1, x_2, x_1^2, x_2^2, x_1x_2]$$

   where x_1 and x_2 are the first and second feature for original data point. Submit in a table the regression coefficient w as well the error in the residuals ($norm(y - \hat{y}, 2)$) obtained for cases.

3. (2 points) Consider dataset *data_henk.mat*. It has the variables *X_test*, *X_train* *Y_test*, *Y_train* where each X has data consisting of 7 features, and Y is the corresponding output. Use gaussian process regression to get predictions *Y_pred* for the test data *X_test* using (*X_train*, *Y_train*) for training. You can use the implementation available at http://www.gaussianprocess.org/gpml/code/matlab/doc/. Using the code at this website, you can run gp regression using the commands:

```
hyp = struct;
hyp.mean = [];
hyp.cov = [];
sn = 0.1;
hyp.lik = log(sn);
negloglik = gp(hyp, [] , [], @covLIN, [], X_train, Y_train);
[Y_pred m2] = gp(hyp, [] , [], @covLIN, [], X_train, Y_train, X_test);
```

   Report the residual error (*norm(Y_pred - Y_test, 2)*), and submit code. Compare with result obtained using *mvregress*. Can you run your implementation of svm regression for this problem? What are the results?

4. (4 points) Consider dataset *q3.mat* containing two-dimensional data generated from mixture of two Gaussian distributions with unknown means and covariances. Implement the EM algorithm discussed in class to identify the unknown means and covariances.
   (a) Report μ_1, Σ_1 and μ_2, Σ_2
   (b) Plot the loglikelihood with increasing EM iterations.
   (c) Submit your implementation.