Assignment 3

Machine Learning, Summer term 2013, Ulrike von Luxburg

Solutions due by April 29

Exercise 1 (Least square regression, 8 point) In this exercise you will implement the linear least square method for regression. First generate a simple synthetic dataset

```
sigma = 0.1;
[X_train,Y_train] = genLinData(50,sigma);
[X_test,Y_test] = genLinData(30,sigma);
```

1. Plot the input data and describe it by reading `genLinData.m`
2. Preprocess the training data by concatenating 1 (for the bias term) to each training point.
3. Write a function `w = LLS(X,Y)` that given the training data X and the training labels Y, returns a linear classifier (vector of weights) w. X is a $n \times d$ matrix consisting n points in the d dimensional space. Y is a vector of size n and w is a vector of size d.
4. In the same figure from Step 1, plot the fitted line.
5. Predict labels for `X_test` (preprocess `X_test` as you did it for the training data). Plot it in the figure from Step 1 with a different color.
6. Write a function `err = lossL2(Y,Y_pred)` which returns the empirical squared loss of predicting `Y_pred` instead of `Y`.
7. Find the average error for 10 different runs of your code. Plot the average error for

   ```
   sigma = [0.01 0.1 0.4 0.9 1];
   ```

8. Add an extreme outlier to your training data

   ```
   X_train = [X_train;10];
   Y_train = [Y_train;10];
   ```

   and run your code to see its effect in linear least square regression.

Exercise 2 (Ridge regression, 6 point) In this exercise you will implement the ridge regression algorithm. Load the synthetic train and test data from `dataRidge.mat`.

1. For preprocessing, do the same as Step 2 from Exercise 1.
2. Run `LLS` from exercise 1 and plot training points, the predicted line and predicted values for the test data.
3. Run `LLS` with polynomial basis functions

   $$\Phi_i(x) = x^i; i = 1, ..., 15 \tag{1}$$

   Illustrate the learned regression function by applying it on `xx=-1.5:0.01:2.5`.
4. Describe the class of functions that you can learn with these basis functions.

Web page: `http://www.informatik.uni-hamburg.de/ML/contents/people/luxburg/teaching/2013-ss-vorlesung-ml/`
Login: "machine", Password: "learning"

5. Write a function `RidgeLLS(X,Y,lambda)` which implements the ridge regression. Here, X is the design matrix.

6. Apply the ridge regression on the test data using the set of basis functions in Equation 1. Plot the prediction function (on the previous figure) for $\lambda \in \{0.0001, 0.1, 10\}$ by applying it on `xx=-1.5:0.01:2.5`.

7. Report the prediction error with respect to λ for $\lambda \in \{2^i; i = -15, -14, ..., 1\}$.

Exercise 3 (1 point) Compare the prediction running time for linear least square method and kNN regression (computational complexity). How much information do you need to keep for predicting with each method (space complexity)?

Exercise 4 (Convexity, 2 point) Prove that the least squares loss function $\|Y - Xw\|^2$ is a convex function of w.

Exercise 5 (optional, 0 points)
In the lecture I mentioned the law case of Lucia de Berk. She is a nurse who was suspected to have murdered a number of patients and was sentenced to life imprisonment. The main arguments that led to her conviction were based on wrong statistical arguments. One of the main things that went wrong were that the same data that was used to suggest the hypothesis that she might be a murderer also was used to test this hypothesis. In machine learning terms, the test error was computed on the training set.
We tried to make an exercise out of this case, but it is too complex and requires too much background knowledge in statistics. But we would like to encourage you to read up on the case, here are a couple of links:
The Wikipedia entry that gives the general background:
`http://en.wikipedia.org/wiki/Lucia_de_Berk`

A talk given by Peter Grünwald that summarizes the statistical flaws:
`http://homepages.cwi.nl/~pdg/presentations/evidencehandout.pdf`

A short comment in Nature about the case (that was in 2007, before the appeal)
`http://www.nature.com/nature/journal/v445/n7125/full/445254a.html`

And finally, a paper that explains many of the statistical arguments that went wrong.
Meester, Collins, Gill, van Lambalgen: On the (ab)use of statistics in the legal case against the nurse Lucia de B.
`http://lpr.oxfordjournals.org/content/5/3-4/233.full.pdf`

If some of you are interested, we can offer an extra discussion session on this case later in the semester (please tell you TA if you'd be interested in such a meeting).