COMP4670/6467
Introduction to Statistical Machine Learning
Tutorial 2

Christfried Webers

12/14 March 2013

1 Regression

1.1 The Data Set

We will use a dataset of established house price indices for Canberra in the years 2008-2011 relative to the house price in 2003-4 (= 100%) which can be found in the text file *Canberra-Houses.txt*. The data are sourced from

`http://www.ausstats.abs.gov.au/ausstats/meisubs.nsf/`
`0/79561B63B460F917CA257996000F9D38/$File/64160_dec%202011.pdf}`

Write a python routine to read in the data. What are the input data $\mathbf{x}$? What are the targets t ? What else do you need to do in order to *preprocess* the data in order to apply regression?

1.2 Plotting

In order to plot results, the following code can be used.

```
import pylab

...

pylab.plot(X, Y, "bo")
pylab.plot(X1, Y1, "r.")
pylab.show()
```

where X, Y and $X1, Y1$ are numpy arrays, and the string "bo" designates a plot with a blue circle for each data point, and "r." a red dot for each data point. (These modifier strings are similar to Matlab.)

1.3 Regression without Regulariser

Implement the regression procedure to find the maximum likelihood solution w_{ML} for a sum-of-squares error function. Use a number $M = 10$ of basis functions uniformly distributed over the input space. And use subroutines which allow you to easily switch between polynomial basis functions, Gaussian basis functions, and sigmoidal basis functions.

The parameters μ_j for $j = 0 \ldots M - 1$, and s can be stored in a global variable to simplify the code.

(Hint: Use subroutines to calculate $\phi_i(\mathbf{x})$, $\boldsymbol{\phi}(\mathbf{x})$ and $\boldsymbol{\Phi}$. Then a change of basis functions will be reflected in a change of $\phi_i(\mathbf{x})$ only. Or, as predictions works with one row of $\boldsymbol{\phi}$ evaluated at the test x, you can also modularise per row in $\boldsymbol{\phi}$. Nevertheless, try to hide dealing with subscripts in subroutines so you don't get confused!)

1.4 Training

Choose half of the available data to train the model by calculating the maximum likelihood parameter $\mathbf{w}_{ML}$ for the model and some choice of basis functions.

Assume you choose every second data pair for training in one experiment, and the first half of the data in a second experiment. For which experiment do you expect a better prediction on the the remaining test data?

1.5 Testing

Use the remaining data to evaluate the sum-of-squares error function between the targets of the test set and the predicted values $y(x)$ of the regressor.

1.6 Exploring

Does another set of basis function give better results (smaller error) in the test phase after training the model to find $\mathbf{w}_{ML}$?

How does the error change if the basis functions are not uniformly distributed over the input space?

(Note: Use the same training and test data as before.)

1.7 Regression with Regulariser

Use the squared length of the weight vector as regulariser multiplying it with a regularisation coefficient $\lambda > 0$ to reduce the complexity of the model. Incorporate λ into your code to calculate $\mathbf{w}_{ML}$ for regression with the given regulariser.

1.8 Exploring the Regression with Regulariser

Can you setup an experiment which tries to find the optimal model complexity by varying λ for each train/test cycle?

(Note: Use the same training and test data as before.)