Learning from Data, Tutorial Sheet for week 8

Division of Informatics, University of Edinburgh

Instructor: Amos Storkey

1. Given training data $D = \{(x^\mu, c^\mu), \mu = 1, \ldots, P\}$, $c^\mu \in \{0, 1\}$, and $\mathbf{x}$ are two dimensional, you decide to make a classifier using a neural network with a single hidden layer, with two hidden units, each with transfer function $g(x) = \exp(-0.5x^2)$. The output transfer function is the logistic sigmoid, $\sigma(x) = e^x/(1 + e^x)$, so that the network is,

$$p(c = 1|\mathbf{x}) = \sigma(b_0 + v_1 g(\mathbf{w}_1^T \mathbf{x} + b_1) + v_2 g(\mathbf{w}_2^T \mathbf{x} + b_2))$$

- Write down the log likelihood, based on the usual i.i.d assumption for this model.
- Calculate the derivatives of the log likelihood as a function of the network parameters, $\mathbf{w}_1, \mathbf{w}_2, b_1, b_2, \mathbf{v}, b0$
- Comment on the relationship between this model and logistic regression.
- Comment on the decision boundary of this model.

2 An example classification application is to classify Reuters news stories into categories (e.g. sport, politics, finance etc). This is an example of a multi-class problem. One approach to solving this is to use the so-called 'softmax' output which, for C classes takes the general form:

$$p(class = i|x) = \frac{e^{f(\boldsymbol{x},i)}}{\sum_{j=1}^{C} e^{f(\boldsymbol{x},j)}}$$

Comment on the relation of this general approach to logistic regression.

To use the above 'softmax' approach we need to convert text into a vector representation. This is usually done using the "bag of words" representation of the document where the order of the words is thrown away, so that simple counts of each word present in the document are used. There are some standard preprocessing techniques that are also used:

- Word stems are used rather than words;
- Words that occur infrequently in the training data are discarded;
- "Stop words" (like "and", "or" etc) are not counted;
- To deal with documents of different length, each document feature vector is normalized to have unit length.

Consider how long the feature vector may be and discuss why these preprocessing techniques may be beneficial. Are there situations where you think the "bag of words" representation would cause problems?

Outline how you would proceed to formulate and train an appropriate model.