Machine Learning
Exercise 6

Marc Toussaint
Machine Learning & Robotics lab, U Stuttgart
Universitätsstraße 38, 70569 Stuttgart, Germany

June 11, 2015

Note: We were again rather late in posting this exercise. Sorry. Please try to do at least the first part of the coding exercise, showing that training works (error reduces). The points a) - c) are optional.

1 Coding Neural Networks for classification

Consider the classification data set `../data/data2ClassHastie.txt`, which is a non-linear classification problem (the same as the one used in the slides on classification). It has a three-dimensional input, where the first is a constant 1.0 (allowing for a bias in the first hidden layer).

Train a NN for this classification problem. As this is a binary classification problem we only need one output neuron y. If $y > 0$ we classify as '+1' otherwise as '-1'.

- First code a routine that computes $y = f(x, w)$ for a 1-layer NN with hidden layer size $h_1 = 100$. Note that $w = (\beta, W_0)$ is a $10 + 300$ dimensional parameter.
- Initialize $g = (g_\beta, g_{W_0})$ (of same dimensions as w) to zero. Then loop through the data, for each x_i
  - Compute $f_i = f(x_i, w)$ and the error $e = \max\{0, 1 - y_i f_i\}$
  - If $e = 0$ do nothing. Otherwise define $\delta_f = -y_i$ and backpropagate the gradient to all weights. As we need to sum these gradients over all data points, add them to (g_β, g_{W_0})

  After this, make a small gradient descent step $w \leftarrow w - \alpha g$ on all weights.
- Iterate this.

a) Visualize the prediction by plotting $\sigma(f(x, w))$ over a 2-dimensional grid. The σ "squashes" the discriminative value f as if we'd interpret it as a logistic model. That gives a nicer illustration.

b) Test regularization. (No full-fledged CV necessary.)

c) **Stochastic gradient** means that, instead of computing proper gradients by summing over the full data, we only compute a "stochastic gradient" by picking a single (or a few) random data points, compute the gradient for this alone and make a (very small) gradient step in this direction. How does this perform?