UNIVERSITY OF OXFORD

Machine Learning
Nando de Freitas
February 19, 2015

Exercise Sheet 3

1 Gradient and Hessian of log-likelihood for logistic regression

1. Let $\sigma(a) = \frac{1}{1+e^{-a}}$ be the sigmoid function. Show that

$$\frac{d\sigma(a)}{da} = \sigma(a)(1-\sigma(a)) \tag{1}$$

2. Using the previous result and the chain rule of calculus, derive an expression for the gradient of the log likelihood of logistic regression discussed in the lectures (also appearing in Section 8.3.1 of Kevin's book).

3. The Hessian can be written as $\mathbf{H} = \mathbf{X}^T\mathbf{S}\mathbf{X}$, where $\mathbf{S} := \text{diag}(\mu_1(1-\mu_1), \ldots, \mu_n(1-\mu_n))$. Show that $\mathbf{H}$ is positive definite. (You may assume that $0 < \mu_i < 1$, so the elements of $\mathbf{S}$ will be strictly positive, and that $\mathbf{X}$ is full rank.)

2 Gradient and Hessian of log-likelihood for multinomial logistic regression

1. Please read the Kevin Murphy's textbook section on Multi-class logistic regression (Section 8.3.7 in my edition).

2. Let $\mu_{ik} = p(y_i = k|\mathbf{x}_i, \mathbf{W}) = \mathcal{S}(\boldsymbol{\eta}_i)_k$. Prove that the Jacobian of the softmax is

$$\frac{\partial \mu_{ik}}{\partial \eta_{ij}} = \mu_{ik}(\delta_{kj} - \mu_{ij}) \tag{2}$$

   where $\delta_{kj} = I(k = j)$.

3. Hence show that

$$\nabla_{\mathbf{w}_c}\ell \;\; = \;\; \sum_i (y_{ic} - \mu_{ic})\mathbf{x}_i \tag{3}$$

   Hint: use the chain rule and the fact that $\sum_c y_{ic} = 1$.

4. Show that the block submatrix of the Hessian for classes c and c' is given by

$$\mathbf{H}_{c,c'} \;\; = \;\; -\sum_i \mu_{ic}(\delta_{c,c'} - \mu_{i,c'})\mathbf{x}_i\mathbf{x}_i^T \tag{4}$$

3 Neural nets and XOR gates

Show that if the activation function of the hidden units is linear, a 3-layer (1 input layer $\mathbf{x}$, 1 hidden layer $\mathbf{h}$ and 1 output layer $\mathbf{y}$) network is equivalent to a 2-layer one. Use your result to explain why a three-layer network with linear hidden units cannot solve a non-linearly separable problem such as XOR.

4 Research question: Rectified linear units and dropout

- Why are ReLU or maxout units preferable to sigmoid neurons in convolutional neural networks?
- Explain dropout, a technique invented by Geoff Hinton and collaborators, in the context of deep learning. How does it interact with weight decay?