Machine Learning
Homework 3

Yao Song
301266041

November 6, 2014

Question 1

$$\frac{\mathrm{d}}{\mathrm{d}x}\left((1+\exp(-a)^{-1})\right) \tag{1}$$
$$=-\left(1+\exp(-a)\right)^{-2}\frac{\mathrm{d}}{\mathrm{d}x}\left(\exp(-a)\right) \tag{2}$$
$$=\frac{exp(-a)}{(1+\exp(-a))^2} \tag{3}$$
$$=\frac{1}{1+\exp(-a)}-\frac{1}{(1+\exp(-a))^2} \tag{4}$$
$$=g(a)-g(a)^2 \tag{5}$$
$$=g(a)(1-g(a)) \tag{6}$$

Question 2

1

No. They are not independent.

2

The assumption is reasonable. Because the driver's age is not necessarily reflected on the mileage of the vehicle. It may be a second hand car owned by a new driver or an experienced driver may just get a new car.

3

The ExtraCar node tells us something about the SocioEcon, which will affect the MedicalCost. So ExtraCar information is useful in determining MedicalCost in an indirect way. For example, if we know a customer has an extra car, his or her SocioEcon is more likely to be better than the customer's SocioEcon without an extra car. So we know this customer is able to afford more insurance cost. That is important information to be incorporated into our determination.

4

Denote MeicalCost = x, LiabilityCost = y, PropertyCost =z, age= a, VehicleYear = v. Then the expectation of the total insurance cost of the prospective client can be written as follows:

$$E[x+y+z|a,v]=\int_x\int_y\int_z(x+y+z)Pr(x,y,z|a,v)\,dxdydz,$$

where $Pr(x,y,z|a,v)$ is the joint conditional probability of x,y,z given a and v.

Question 3

The likelihood function can be written as

$$L(\boldsymbol{\theta}) = P(\mathbf{x}^1, \mathbf{x}^1, \ldots, \mathbf{x}^N|\boldsymbol{\theta}) \tag{7}$$

$$= \prod_{i=1}^{N} P(\mathbf{x}^i|\boldsymbol{\theta}) = \prod_{i=1}^{N} P(x_1^i, x_2^i, \ldots, x_D^i|\boldsymbol{\theta}) \tag{8}$$

$$= \prod_{i=1}^{N}\prod_{k=1}^{D} P(x_k^i|pa_k;\boldsymbol{\theta}_k) \tag{9}$$

$$= \prod_{k=1}^{D} P(x_k^1, x_k^2, \ldots, x_k^N|pa_k;\boldsymbol{\theta}_k) \tag{10}$$

$$= \prod_{k=1}^{D} P(\boldsymbol{x_k}|pa_k;\boldsymbol{\theta}_k) \tag{11}$$

where we have defined $\boldsymbol{x_k} = [x_k^1, x_k^2, \ldots, x_k^N]$. So we can see that learning the parameters for the network according to the maximum likelihood criterion can be done by separately learning the parameters $\boldsymbol{\theta}_k$ of each conditional distribution using the relevant components of the training data.

Question 4

1

Please see the plot in the last page.

2

By commenting out line 39 in the code, I changed the initialization of the first layer weights.

3

The first thing we can observe is the accuracy does not change any more. Going one step further, we will find dW1 and dW2 do not change. So the weights will not change. Obviously that is due to the improper initialization of the weights. When we go into the process, we can observe that the returned values A and Z from the function "feedforward" do not change at all. So we need to analyze the feedfoward function.

By analysis of feedfoward function, we can find that A_1 has large values. That makes $Z_1 = 1./(1 + exp(-A_1))$ equal to vector of ones and remain unchanged.

If we add line39 in the code. We can observe A_1 with small values and as a result Z_1 will have different values and keep changing.

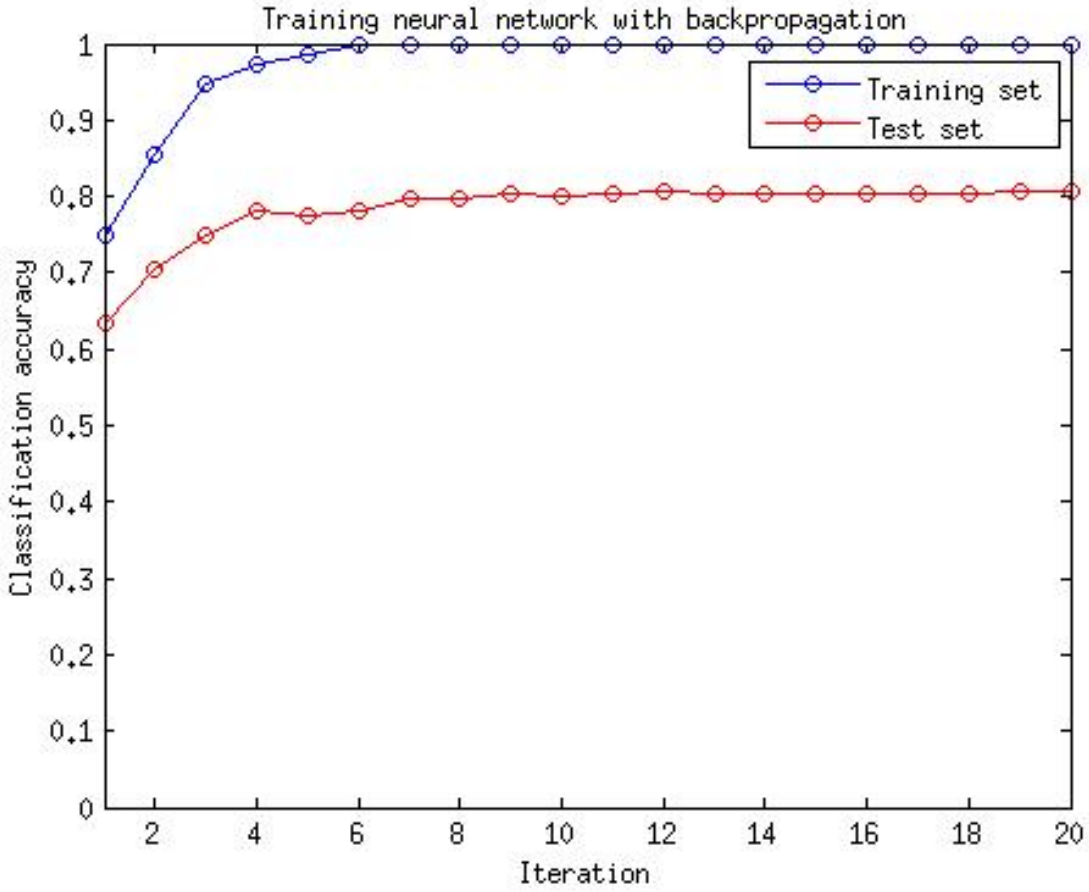

Figure 1: training/testing error plot