Exercises
Machine Learning
HS 2016

Series 6, Dec 12, 2016 (Neural Networks)

Machine Learning Laboratory
Dept. of Computer Science, ETH Zürich
Prof. Joachim M. Buhmann
Web https://ml2.inf.ethz.ch/courses/ml/

For questions 1,2,3 : **Emiliano Diaz**
piazza
For questions :

Solution 1 (Backpropagation for classification):

Using notation for a multi-level neural network used in the tutorial we have:

$$
\begin{aligned}
\frac{\partial L_n}{\partial w_{ik}^{L+1}} &= \frac{\partial}{\partial w_{ik}^{L+1}} - \sum_{r=1}^{I} y_r \ln \hat{y}_r + (1 - y_r)\ln(1 - \hat{y}_r) && (1)\\
&= -\Big(\frac{y_i}{\hat{y}_i} - \frac{1 - y_i}{1 - \hat{y}_i}\Big)\frac{\partial \hat{y}_i}{\partial w_{ik}^{L+1}} && (2)\\
&= \Big(\frac{1 - y_i}{1 - \hat{y}_i} - \frac{y_i}{\hat{y}_i}\Big)\frac{\partial \hat{y}_i}{\partial w_{ik}^{L+1}} && (3)\\
&= \Big(\frac{\hat{y}_i(1 - y_i) - y_i(1 - \hat{y}_i)}{\hat{y}_i(1 - \hat{y}_i)}\Big)\frac{\partial}{\partial w_{ik}^{L+1}}\sigma\Big(\sum_{m=1}^{K(L)} w_{im}^{L+1} z_m^L\Big) && (4)\\
&= \Big(\frac{\hat{y}_i - y_i}{\hat{y}_i(1 - \hat{y}_i)}\Big)\sigma'\Big(\sum_{m=1}^{K(L)} w_{im}^{L+1} z_m^L\Big) z_k^L && (5)\\
&= \delta_i^{L+1} z_k^L && (6)
\end{aligned}
$$

where

$$
\delta_i^{L+1} = \Big(\frac{\hat{y}_i - y_i}{\hat{y}_i(1 - \hat{y}_i)}\Big)\sigma'\Big(\sum_{m=1}^{K(L)} w_{im}^{L+1} z_m^L\Big)
$$

Now

$$
\begin{aligned}
\frac{\partial L_n}{\partial w^L_{mk}} &= \frac{\partial L_n}{\partial z^L_m}\frac{\partial z^L_m}{\partial w^L_{mk}} && (7)\\
&= \frac{\partial}{\partial z^L_m} - \sum_{i=1}^{I} y_i \ln \hat{y}_i + (1-y_i)\ln(1-\hat{y}_i)\frac{\partial}{\partial w^L_{mk}} h\Big(\sum_{r=1}^{K(L-1)} w^L_{mr} z^{L-1}_r\Big) && (8)\\
&= \sum_{i=1}^{I}\{\Big(\frac{\hat{y}_i - y_i}{\hat{y}_i(1-\hat{y}_i)}\Big)\frac{\partial}{\partial z^L_m}\sigma\Big(\sum_{r=1}^{K(L)} w^{L+1}_{ir} z^L_r\Big)\} h'\Big(\sum_{r=1}^{K(L-1)} w^L_{mr} z^{L-1}_r\Big) z^{L-1}_k && (9)\\
&= \sum_{i=1}^{I}\{\Big(\frac{\hat{y}_i - y_i}{\hat{y}_i(1-\hat{y}_i)}\Big)\sigma'\Big(\sum_{r=1}^{K(L)} w^{L+1}_{ir} z^L_r\Big) w^{L+1}_{im}\} h'\Big(\sum_{r=1}^{K(L-1)} w^L_{mr} z^{L-1}_r\Big) z^{L-1}_k && (10)\\
&= \sum_{i=1}^{I}\{\delta^{L+1}_i w^{L+1}_{im}\} h'\Big(\sum_{r=1}^{K(L-1)} w^L_{mr} z^{L-1}_r\Big) z^{L-1}_k && (11)\\
&= \delta^L_m z^{L-1}_k && (12)
\end{aligned}
$$

where

$$
\delta^L_m = \sum_{i=1}^{I}\{\delta^{L+1}_i w^{L+1}_{im} h'\Big(\sum_{r=1}^{K(L-1)} w^L_{mr} z^{L-1}_r\Big)
$$

We perform one more iteration:

$$
\begin{aligned}
\frac{\partial L_n}{\partial w^{L-1}_{mk}} &= \frac{\partial L_n}{\partial z^{L-1}_m}\frac{\partial z^{L-1}_m}{\partial w^{L-1}_{mk}} && (13)\\
&= \frac{\partial}{\partial z^{L-1}_m} - \sum_{i=1}^{I} y_i \ln \hat{y}_i + (1-y_i)\ln(1-\hat{y}_i)\frac{\partial}{\partial w^{L-1}_{mk}} h\Big(\sum_{r=1}^{K(L-2)} w^{L-1}_{mr} z^{L-2}_r\Big) && (14)\\
&= \sum_{i=1}^{I}\{\Big(\frac{\hat{y}_i - y_i}{\hat{y}_i(1-\hat{y}_i)}\Big)\frac{\partial}{\partial z^{L-1}_m}\sigma\Big(\sum_{r=1}^{K(L)} w^{L+1}_{ir} z^L_r\Big)\} h'\Big(\sum_{r=1}^{K(L-2)} w^{L-1}_{mr}) z^{L-2}_r\Big) z^{L-2}_k && (15)\\
&= \sum_{i=1}^{I}\{\Big(\frac{\hat{y}_i - y_i}{\hat{y}_i(1-\hat{y}_i)}\Big)\sigma'\Big(\sum_{r=1}^{K(L)} w^{L+1}_{ir} z^L_r\Big)\frac{\partial}{\partial z^{L-1}_m}\sum_{r=1}^{K(L)} w^{L+1}_{ir} h\Big(\sum_{s=1}^{K(L-1)} w^L_{rs} z^{L-1}_s\Big)\} h'\Big(\sum_{r=1}^{K(L-2)} w^{L-1}_{mr}) z^{L-2}_r\Big) z^{L-2}_k && (16)\\
&= \sum_{i=1}^{I}\{\delta^{L+1}_i \sum_{r=1}^{K(L)} w^{L+1}_{ir}\frac{\partial}{\partial z^{L-1}_m} h\Big(\sum_{s=1}^{K(L-1)} w^L_{rs} z^{L-1}_s\Big)\} h'\Big(\sum_{r=1}^{K(L-2)} w^{L-1}_{mr}) z^{L-2}_r\Big) z^{L-2}_k && (17)\\
&= \sum_{r=1}^{K(L)}\{(\sum_{i=1}^{I}\delta^{L+1}_i w^{L+1}_{ir}) h'\Big(\sum_{s=1}^{K(L-1)} w^L_{rs} z^{L-1}_s\Big) w^L_{rm}\} h'\Big(\sum_{r=1}^{K(L-2)} w^{L-1}_{mr}) z^{L-2}_r\Big) z^{L-2}_k && (18)\\
&= \sum_{r=1}^{K(L)}\{\delta^L_r w^L_{rm}\} h'\Big(\sum_{r=1}^{K(L-2)} w^{L-1}_{mr}) z^{L-2}_r\Big) z^{L-2}_k && (19)\\
&= \delta^{L-1}_m z^{L-2}_k && (20)\\
& && (21)
\end{aligned}
$$

where

$$\delta_m^{L-1} = \sum_{r=1}^{K(L)} \{\delta_r^L w_{rm}^L\} h'\Big(\sum_{r=1}^{K(L-2)} w_{mr}^{L-1}) z_r^{L-2}\Big)$$

We see that in general we the backpropagation equations are:

1. $\frac{\partial L_n}{\partial w_{ik}^{L+1}} = \delta_i^{L+1} z_k^L$ and $\delta_i^{L+1} = \Big(\frac{\hat{y}_i - y_i}{\hat{y}_i(1-\hat{y}_i)}\Big)\sigma'\Big(\sum_{m=1}^{K(L)} w_{im}^{L+1} z_m^L\Big)$
2. $\frac{\partial L_n}{\partial w_{mk}^l} = \delta_m^l z_k^{l-1}$ and $\delta_m^l = \sum_{r=1}^{K(l+1)} \{\delta_r^{l+1} w_{rm}^{l+1}\} h'\Big(\sum_{r=1}^{K(l-1)} w_{mr}^l) z_r^{l-1}\Big)$ for $l \in \{2, ..., L\}$
3. $\frac{\partial L_n}{\partial w_{kj}^l} = \delta_k^1 x_j$ and $\delta_k^1 = \sum_{r=1}^{K(2)} \{\delta_r^2 w_{rm}^2\} h'\Big(\sum_{j=1}^{J} w_{kj}^1) x_j\Big)$

Solution 2 (Maximum likelihood estimator for regression):

The likelihood function for an i.i.d. data set, $\{(x_1, t_1), ..., (x_N, t_N)\}$, under the conditional distribution:

$$p(t|x, w) = N(t|y(x, w), \beta^{-1} I)$$

is given by

$$\prod_{n=1}^{N} N(t_n|y(x_n, w), \beta^{-1} I)$$

If we take the logarithm of this, using the definition of a multivariate normal distribution, we get

$$\sum_{n=1}^{N} \ln N(t_n|y(x_n, w), \beta^{-1} I) \tag{22}$$

$$= -\frac{1}{2}\sum_{n=1}^{N} (t_n - y(x_n, w))^T (\beta I)(t_n - y(x_n, w)) + K \tag{23}$$

$$= -\frac{\beta}{2}\sum_{n=1}^{N} ||t_n - y(x_n, w)||^2 + K \tag{24}$$

where K comprises terms which are independent of w. The first term on the right hand side is proportional to the negative of:

$$E(w) = \frac{1}{2}\sum_{n=1}^{N} ||y(x_n, w) - t_n||^2$$

and hence maximizing the log-likelihood is equivalent to minimizing the sum-of-squares error.

Solution 3 (Maximum likelihood estimator for classification):

For the given interpretation of $y_k(x, w)$, the conditional distribution of the target vector for a multiclass neural network is

$$p(t|w_1, ..., w_K) = \prod_{k=1}^{K} y_k^{t_k}$$

Thus, for a data set of N points, the likelihood function will be

$$p(T|w_1, ..., w_K) = \prod_{n=1}^{N} \prod_{k=1}^{K} y_{nk}^{t_{nk}}$$

Taking the negative logarithm in order to derive an error function we obtain

$$E(w) = -\sum_{n=1}^{N} \sum_{k=1}^{K} t_{kn} \ln y_k(x_n, w)$$

as required.