まず、

$$\hat{y}_{ik} = \frac{1}{1 + \exp\left(-b_k - \sum_j x_{ij} w_{jk}\right)}$$

コスト関数は、

$$J = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k}[y_{ik}\log\hat{y}_{ik} + (1-y_{ik})\log(1-\hat{y}_{ik})] + \lambda\sum_{j}\sum_{k} w_{jk}^2$$

微分は、

$$\frac{\partial J}{\partial w_{jk}} = -\frac{1}{N}\sum_{i=1}^{N}\left[(y_{ik} - \hat{y}_{ik})x_{ij}\right] + 2\lambda w_{jk}$$

$$\frac{\partial J}{\partial b_k} = -\frac{1}{N}\sum_{i=1}^{N}(y_{ik} - \hat{y}_{ik})$$

なんで、普通に squared error をコスト関数として使わないかというと、squared error だと non convex になるかららしい。まず、上の微分の二階微分は、

$$\begin{cases} \dfrac{\partial^2 J}{\partial w_{jk}^2} = 2\lambda \geq 0 \\ \dfrac{\partial^2 J}{\partial b_k^2} = 0 \end{cases}$$

よって、convex なので、gradient descent で最適化できる。Squared error の方はまだチェックしてない（実際に計算してチェックしてみよう）。

しかし、連続値の場合でも、このコスト関数で良いのか、怪しいので要チェックだ。つまり、$y_{ik} = \hat{y}_{ik}$の時に、上記のコストが最小値となることを証明できれば良いはず。

追記：
証明できた。

$$\begin{cases} \dfrac{\partial J}{\partial \hat{y}_{ik}} = -\dfrac{1}{N}\dfrac{y_{ik}-\hat{y}_{ik}}{\hat{y}_{ik}(1-\hat{y}_{ik})} \\ \dfrac{\partial^2 J}{\partial \hat{y}_{ik}^2} = \dfrac{1}{N}\left[\dfrac{y_{ik}-\hat{y}_{ik}}{\hat{y}_{ik}(1-\hat{y}_{ik})}\right]^2 \geq 0 \end{cases}$$

二階微分が非負より、傾きは常に増加するということ。つまり、下に凸なカーブだ。そして、一階微分より、$\hat{y}_{ik} = y_{ik}$の時に最小値となることが分かる。よって、このコスト関数は、ちゃんと$\hat{y}_{ik} = y_{ik}$に近づけるという目的を果たすことができる。