Machine Learning using Matlab, Exercise sheet 3

Florian Wolf

Question 1

Overfitting: The model fits the data and the natural noise in the data pretty good, but fails to perform on new/not seen test data sets. The good trainings data fit is often based on a very complex model, containing way too many parameters, compared to the data's parameters. E.g. increasing the dimension of a polynomial fit until the trainings data perfectly matches, although the ("real") degree of the polynomial representing the structure of the data is way lower.

Identify:
- low loss in training, but relatively high in test data
- huge amount of parameters, e.g. of the order of magnitude of the number of training data

Address: . Model reduction → less parameters/less complex hypothesis for more simplicity

Underfitting: The model cannot represent or fit the data in an adequate way, because the number of parameters used is way to low, to be able to represent the higher dimensional correlation in the data. E.g. if you try to use a linear fit to match data, which correlates in a quadratic way → no matter how good the fit (locally is) it cannot represent the data due to lower dimension.

Identify: Bad/poor results in testing and training (a bad model can lead to this too)

Address: Increase the amount of parameters or/and make the models hypothesis more complex

Question 2 Let us define $h^{u}(x^{(i)}) := w^{(u)} x^{(i)} + b^{(u)}$, therefore we get

$\frac{\partial}{\partial w^{(m)}} h^{u}(x^{(i)}) = x^{(i)} \delta_{um}$ and $\frac{\partial}{\partial b^{(e)}} h^{u}(x^{(i)}) = \delta_{ue}$, where $\delta_{ue} = \chi_{u=e}$

for the indicator function χ. Let us write

$$\mathcal{L}(W,b) = -\frac{1}{m} \sum_{i=1}^{m} \sum_{u=1}^{K} \chi_{y^{(i)}=u} \cdot \log \underbrace{\frac{\exp(h^{(u)}(x^{(i)}))}{\sum_{j=1}^{K} \exp(h^{(j)}(x^{(i)}))}}_{=: F}$$

Now we get

$$\frac{\partial F}{\partial h^{(e)}} = \sum_{u=1}^{K} \chi_{y^{(i)}=u} \left(\frac{\sum_{j=1}^{K} \exp(h^{(j)}(x^{(i)}))}{\exp(h^{(u)}(x^{(i)}))} \cdot \left(\frac{\exp(h^{(u)}(x^{(i)})) \cdot \delta_{ue}}{\sum_{j=1}^{K} \exp(h^{(j)}(x^{(i)}))} - \right. \right.$$

$$\left. \left. \frac{\exp(h^{(u)}(x^{(i)}))}{\left(\sum_{j=1}^{K} \exp(h^{(j)}(x^{(i)}))\right)^{2}} \cdot \left(\sum_{j=1}^{K} \exp(h^{(j)}(x^{(i)})) \delta_{je} \right) \right) \right)$$

$$= \exp(h^{(e)}(x^{(i)}))$$

$$= \sum_{u=1}^{K} \chi_{y^{(i)}=u} \cdot \left(\delta_{ue} - \frac{\exp(h^{(e)}(x^{(i)}))}{\sum_{j=1}^{K} \exp(h^{(j)}(x^{(i)}))} \right)$$

$$= \chi_{y^{(i)}=e} - \underbrace{\sum_{u=1}^{K} \chi_{y^{(i)}=u}}_{=1} \left(\frac{\exp(h^{(e)}(x^{(i)}))}{\sum_{j=1}^{K} \exp(h^{(j)}(x^{(i)}))} \right) = \chi_{y^{(i)}=e} - \frac{\exp(h^{(e)}(x^{(i)}))}{\sum_{j=1}^{K} \exp(h^{(j)}(x^{(i)}))}$$

$\Rightarrow$ Using the chain rule, we obtain

$$\frac{\partial \mathcal{L}}{\partial w^{(e)}} = -\frac{1}{m} \sum_{i=1}^{m} \frac{\partial F}{\partial w^{(e)}} \overset{C.R.}{=} -\frac{1}{m} \sum_{i=1}^{m} \frac{\partial F}{\partial h^{(e)}} \underbrace{\frac{\partial h^{(e)}}{\partial w^{(e)}}}_{= x^{(i)}} = -\frac{1}{m} \sum_{i=1}^{m} \chi_{y^{(i)}=e} - \left(\frac{\exp(h^{(e)}(x^{(i)}))}{\sum_{j=1}^{K} \exp(h^{(j)}(x^{(i)}))} \right) x^{(i)}$$

$$\frac{\partial \mathcal{L}}{\partial b^{(e)}} = -\frac{1}{m} \sum_{i=1}^{m} \frac{\partial F}{\partial b^{(e)}} \overset{C.R.}{=} -\frac{1}{m} \sum_{i=1}^{m} \frac{\partial F}{\partial h^{(e)}} \underbrace{\frac{\partial h^{(e)}}{\partial w^{(e)}}}_{=1} = -\frac{1}{m} \sum_{i=1}^{m} \chi_{y^{(i)}=e} - \frac{\exp(h^{(e)}(x^{(i)}))}{\sum_{j=1}^{K} \exp(h^{(j)}(x^{(i)}))}.$$

This was our claim.