Intro to Machine Learning Homework Assignment 3

Goktug Saatcioglu

NetId: gs2417

1. Using cross-validation for early-stopping will mean that we will have to pick the fold with the lowest validation cost instead of averaging out the validation costs over all folds. In other words, we would have to pick the "best fold" and use the model trained on the data as defined by that fold. However, doing this goes against the idea of cross-validation since the purpose of cross-validation is to estimate the empirical cost of a hypothesis set. By selecting the "best fold" we end up picking a model that performs exceptionally well on a particular subset of our data set. This may lead to higher generalization error as a model that performs well for a specific case is unlikely to perform well for all cases (by cases we mean subsets of datasets and future datasets we may encounter) and we could run into issues of overfitting. Thus, we cannot use cross-validation for early-stopping.

2. We derive the distance function below where $y^* = M^*(x)$.

$$
\begin{aligned}
D(y^*, M, x) &= -\log p_{y^*} && \text{by definition} \\
&= -\log \frac{a^+_{y^*}}{\sum_{k=1}^K a^+_k} && \text{since } p = \frac{1}{\sum_{k=1}^k a^+_k} a^+ \\
&= -\log \frac{\exp(a_{y^*})}{\sum_{k=1}^K \exp(a_k)} && \text{since } a^+ = \exp(a) \\
&= -(\log(\exp(a_{y^*})) - \log \sum_{k=1}^K \exp(a_k)) && \text{since } a^+ = \exp(a) > 0 \\
&= -(a_{y^*} - \log \sum_{k=1}^K \exp(a_k)) && \text{simpify terms} \\
&= -a_{y^*} + \log \sum_{k=1}^K \exp(a_k) && \text{distribute the negative} \\
\therefore D(y^*, M, x) &= -a_{y^*} + \log \sum_{k=1}^K \exp(a_k)
\end{aligned}
$$

Note: Notice that we never had to use the definition of a, which is given by $a = W\tilde{x}$, for our deriviation of the distance function.

3. First consider the following observation.

$$\begin{aligned}
\frac{\partial}{\partial w_y}(\sum_{k=1}^{K}\exp(w_k^T\tilde{x})) &= \sum_{k=1}^{K}\frac{\partial}{\partial w_y}\exp(w_k^T\tilde{x}) && \text{by linearity}\\
&= \sum_{k=1}^{K}\begin{cases}\tilde{x}\exp(w_y^T\tilde{x}), & \text{if } w_k = w_y\\ 0, & \text{otherwise}\end{cases} && \text{cases for the derivative}\\
&= \tilde{x}\exp(w_y^T\tilde{x})\\
\therefore \frac{\partial}{\partial w_y}(\sum_{k=1}^{K}\exp(w_k^T\tilde{x})) &= \tilde{x}\exp(w_y^T\tilde{x}) && (3.1)
\end{aligned}$$

Next we take the gradient of the distance function with respect to the weight vector that corresponds to the correct class outputted by the reference machine, i.e. the y^*-th row vector.

$$\begin{aligned}
\frac{\partial D(y^*, M, x)}{\partial w_{y^*}} &= -\frac{\partial}{\partial w_{y^*}}(a_{y^*} - \log\sum_{k=1}^{K}\exp(a_k))\\
&= -\frac{\partial}{\partial w_{y^*}}(w_{y^*}^T\tilde{x} - \log\sum_{k=1}^{K}\exp(w_k^T\tilde{x})) && \text{since } a = W\tilde{x}\\
&= -(\frac{\partial}{\partial w_{y^*}}w_{y^*}^T\tilde{x} - \frac{\partial}{\partial w_{y^*}}\log\sum_{k=1}^{K}\exp(w_k^T\tilde{x})) && \text{by linearity}\\
&= -(\tilde{x} - \frac{\partial}{\partial w_{y^*}}\log\sum_{k=1}^{K}\exp(w_k^T\tilde{x})) && \text{simplify first term}\\
&= -(\tilde{x} - \frac{\partial}{\partial w_{y^*}}(\sum_{k=1}^{K}\exp(w_k^T\tilde{x}))\frac{1}{\sum_{k=1}^{K}\exp(w_k^T\tilde{x})}) && \text{by chain rule}\\
&= -(\tilde{x} - \tilde{x}\exp(w_{y^*}^T\tilde{x})\frac{1}{\sum_{k=1}^{K}\exp(w_k^T\tilde{x})}) && \text{using 3.1}\\
&= -(1 - \frac{\exp(w_{y^*}^T\tilde{x})}{\sum_{k=1}^{K}\exp(w_k^T\tilde{x})})\tilde{x} && \text{factor out } \tilde{x}\\
&= -(1 - p(C = y^* \mid x))\tilde{x} && \text{using the definition of } p(C = y^* \mid x)\\
\therefore \frac{\partial D(y^*, M, x)}{\partial w_{y^*}} &= -(1 - p(C = y^* \mid x))\tilde{x} && (3.2)
\end{aligned}$$

Similarly, let's consider the gradient of the distance function with respect to the weight vector that

corresponds to any other incorrect class.

$$
\begin{aligned}
\frac{\partial D(y^*, M, x)}{\partial w_y} &= -\frac{\partial}{\partial w_y}(a_{y^*} - \log \sum_{k=1}^{K} \exp(a_k)) \\
&= -\frac{\partial}{\partial w_y}(w_{y^*}^T \tilde{x} - \log \sum_{k=1}^{K} \exp(w_k^T \tilde{x})) && \text{since } a = W\tilde{x} \\
&= -(\frac{\partial}{\partial w_y} w_{y^*}^T \tilde{x} - \frac{\partial}{\partial w_y} \log \sum_{k=1}^{K} \exp(w_k^T \tilde{x})) && \text{by linearity} \\
&= -(0 - \frac{\partial}{\partial w_y} \log \sum_{k=1}^{K} \exp(w_k^T \tilde{x})) && \text{simplify first term} \\
&= -(0 - \frac{\partial}{\partial w_y} (\sum_{k=1}^{K} \exp(w_k^T \tilde{x})) \frac{1}{\sum_{k=1}^{K} \exp(w_k^T \tilde{x})}) && \text{by chain rule} \\
&= -(0 - \tilde{x} \exp(w_y^T \tilde{x}) \frac{1}{\sum_{k=1}^{K} \exp(w_k^T \tilde{x})}) && \text{using 3.1} \\
&= -(0 - \frac{\exp(w_y^T \tilde{x})}{\sum_{k=1}^{K} \exp(w_k^T \tilde{x})}) \tilde{x} && \text{factor out } \tilde{x} \\
&= -(0 - p(C = y \mid x)) \tilde{x} && \text{using the definition of } p(C = y \mid x) \\
\therefore \frac{\partial D(y^*, M, x)}{\partial w_y} &= -(0 - p(C = y \mid x)) \tilde{x} && (3.3)
\end{aligned}
$$

We can then combine Eq. (3.2) and Eq. (3.3) into a single vector equation to get the gradient of the distance function with respect to the weight matrix W.

$$\nabla_W D(y^*, M, x) = -(y^* - p)\tilde{x}^T, \tag{3.4}$$

where

$$y^* = \begin{bmatrix} 0, \\ \vdots, \\ 1, \\ \vdots, \\ 0 \end{bmatrix} \leftarrow y^*\text{-th row},$$

and p is given as in the lecture notes (Eq. (1.26)). Thus, Eq (3.4) is the same equation as Eq (1.28) of the lecture notes and we have derived the learning rule.

4. See hw3.ipnyb.