Multiclass Classification

Neal Moorthy

3/4/18

1. If you used early stopping then each of the K models will have differently sized test and validation sets and we would most likely fall into a local minima without being given a chance to get out.

2. C = M(x) and a = Wx

$$\begin{aligned} D(M^*, M, x) &= -logp(C = M^*(x)|x) \\ &= -logP_{M^*(x)} \\ \overrightarrow{a^+} &= exp(a) \\ P &= \frac{a^+}{\sum_{i=1}^{K} a_i^+} \\ \overrightarrow{P} &= \frac{exp(Wx)}{\sum exp(W_i X_i)} \\ \log(P) &= W\tilde{x} - \log(\sum exp(W_i X_i) \\ &= -a_y{}^* + log \sum_{k=1}^{K} exp(a_k) \end{aligned}$$

3. When you substitute in W$\tilde{x}$ for a and perform the partial derrivative with respect to W then we have $-$(1$-$ the probablity that our machine is correct) and $-$(0$-$ the probablity that our machine is correct). 1 and 0 are simply the possible values of the reference machine so really we have $-$(y*$-$ the probability that our machine is correct)

$$\begin{aligned} D(y^*, M, x) &= -a_{y^*} + log \sum_{k=1}^{K} exp(a_K) \\ \frac{\partial D(y^*, M, x)}{\partial W_y} &= \frac{\partial}{\partial W_y}(-exp(W^*x) + log \sum_{k=1}^{K} exp(W_k x_k)) \\ &= -(y * -p)\tilde{x}^T \end{aligned}$$