h2-written

kgz2437

February 2023

1 Question 1

Question 1b: This controls for confounding factors such as the quality of label collection + cleaning, underlying distributions. And it allows for fair comparison of different approaches.

Question 1c: The test set is the portion of the dataset that's not used to training. validation set is used for model selection, therefore it should be taken out of the training set.

2 Question 2

3 Question 3

$$\frac{\partial}{\partial \hat{y}_i} L(\hat{y}, y) = -1(i - 1 = y) + \frac{e^{\hat{y}_i}}{\hat{y}_1 + \hat{y}_2} = -1(i - 1 = y) + \textbf{softmax}(\hat{y})$$

.

The derivative contains the output of the softmax, therefore combining the two makes the derivative very easy to compute. However, if we were to separate them: then we would need to compute the derivative of the softmax, which involves computing an additional term.

$$\frac{\partial}{\partial \hat{y}_i} \textbf{softmax}(\hat{y}) = \frac{e^{\hat{y}_1 + \hat{y}_2}}{\left(e^{\hat{y}_1} + e^{\hat{y}_2}\right)^2}$$

.

4 Question 4

LSTM using the pretrained GLOVE embeddings achieved a test performance of 0.878, while without GLOVE it only achieved 0.677. It appears for both cases, lower learning rate than the default produces better out of sample performance. And secondly large batch sizes lead to better performance for non-pretrained.

pretrained	batch_size	lr	val_acc	test_acc
False	16	0.005	0.5896	0.59104
		0.010	0.5652	0.56944
		0.050	0.5126	0.50000
	32	0.005	0.6262	0.62216
		0.010	0.5784	0.58084
		0.050	0.5142	0.51952
	64	0.005	0.6822	0.66968
		0.010	0.6146	0.61444
		0.050	0.5124	0.50912
True	16	0.005	0.8750	0.87756
		0.010	0.8542	0.86020
		0.050	0.7288	0.72872
	32	0.005	0.8724	0.87804
		0.010	0.5350	0.53500
		0.050	0.6196	0.61348
	64	0.005	0.8592	0.86648
		0.010	0.8090	0.80968
		0.050	0.7896	0.79688

Table 1: Results from hyperparameter search.