HW4 Written Part

Due 5pm Friday Feb 28, 2020

What is your name? Solutions

Problem 1

Suppose that you have a classification model that provides about 80% classification accuracy on the training set and on out-of-sample test data. If you ask a human to do this task, they will generally be able to achieve classification accuracy of close to 100%.

(a) For each of the following possible next steps in your analysis, choose one option that you might consider next, and write 1 sentence explaining your reasoning.

- Change L_2 penalty by:

Increasing λ (Decreasing λ)

Why? Since we have similar accuracy on the training and test data, we are not overfitting the training data. We can't tell for sure based on the information provided, but since the training set accuracy is much lower than what a human could do, it's possible we are underfitting the training data. If that were the case, we could improve our fit to the training data by using a more flexible model, with a smaller L_2 penalty.

- Change the dropout rate by:

Increasing dropout rate (Decreasing dropout rate)

Why?
A smaller dropout rate means that more units are included in each step of gradient descent, reducing the regularizing effect of dropout.

- Change training time by:

(Increasing the number of epochs) Decreasing the number of epochs

Why?
The longer we train the model, the more we optimize training set loss. Running the estimation process for more epochs may result in better training set classification accuracy.

(unless we reach a bad minimum[1] of cost)

- Change model structure by:

Adding more hidden layers Reducing the number of hidden layers

Why?

A model with more hidden layers can represent more complex functions.

- Change model structure by:

Adding more units to hidden layers Reducing the number of units in hidden layers

Why?

A model with more units in the hidden layers can represent more complex functions.

(b) Collecting more data is always good. In this case, would you say that collecting more data is a higher priority than trying some of the ideas in part (a), or a lower priority? Why?

In this case, collecting more data is not as high of a priority as modifying the model using one or more of the ideas in part (a). It could be that by improving our model, we could reach 100% accuracy on the training and test data. We should do as well as possible with our existing data before spending time and money on additional data collection.

Problem 2

(a) I divided my labeled classification data into a training set used for model construction and another portion for validation. I then evaluated 1000 neural architectures by estimating parameters with gradient descent on the training set and computing classification accuracy on the validation set. Discuss why the resulting model is likely to yield poorer accuracy on out-of-sample test data as compared to the validation data, even though the validation data was not directly used for estimating the bias and weight parameters. This could be answered in 1-3 sentences.

If we evaluate the performance of 1000 different models on the validation data and pick the model with best validation set accuracy, we may overfit the validation data. This would happen if we selected a model that captured something in the validation data that is not going to generalize to the test set, which is more likely if we examine performance of many models on the validation set

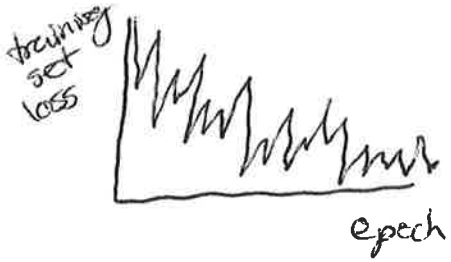
(b) Suppose I now look at performance on my test set and find that indeed, classification accuracy is lower on the test set than it was on the training and validation sets. I now continue to refine my model based on combined validation and test set performance, fitting another 200 variations on neural network models. Is my test set performance a reliable measure of the quality of my final model? This could be answered in 1 sentence.

No for a similar reason as above. If we try many models and choose the one with best test set performance, we may select a model that overfits the test set and does not generalize to new data well.

Problem 3

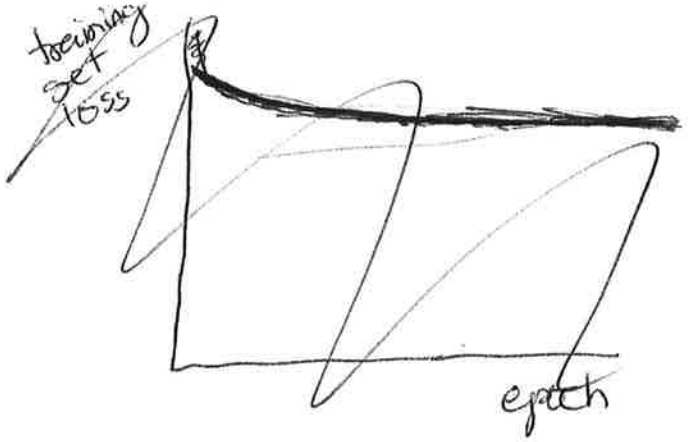
(a) What problem(s) result from having a learning rate that is too high? How would you detect such problems in a plot of training set loss as a function of training epoch? You can answer in 1-2 sentences.

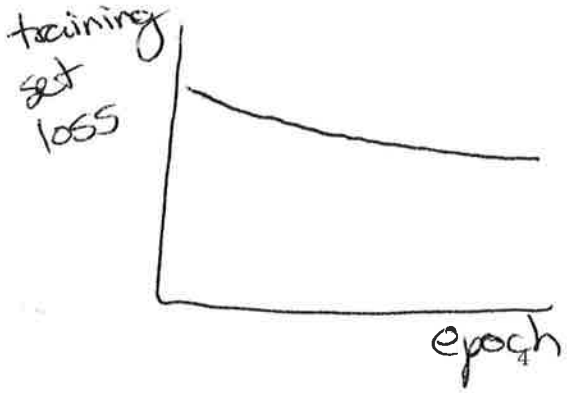
If the learning rate is too large we may take a very large step during gradient descent and move away from a region of low cost in the parameter space. A learning rate that is too high will be visible in a plot of training set loss vs. epoch in the form of a line that oscillates up and down:



(a) What problem(s) result from having a learning rate that is too small? How would you detect such problems in a plot of training set loss as a function of training epoch? You can answer in 1-2 sentences.

If the learning rate is too small the training set loss will not decrease very quickly.





Problem 4

I'm fitting a neural network using L_1 regularization. Suppose that in the current iteration of gradient descent, my weight parameter $w_{11}^{[1]}$ is positive. (The point of this is that although the derivative of the absolute value is not defined everywhere, it is defined if the argument to the absolute value is not 0.)

(a) Find an expression for $\frac{\partial J(b,w)}{\partial w_{11}^{[1]}}$. Your answer can involve a term like $-\frac{1}{m}\frac{\partial \ell(b,w)}{\partial w_{11}^{[1]}}$, where $\ell(b,w)$ is the log-likelihood for your model. (You should not be doing any involved chain rule calculations.)

$$\frac{\partial J(b,w)}{\partial w_{11}^{[1]}} = \frac{\partial}{\partial w_{11}^{[1]}}\left[-\frac{1}{m}\ell(b,w) + \sum_{l=1}^{L}\lambda^{[l]}\sum_{ij}|w_{ij}^{[l]}|\right]$$

$$= -\frac{1}{m}\frac{\partial \ell(b,w)}{\partial w_{11}^{[1]}} + \lambda^{[1]}$$

(b) Suppose the current value of $w_{11}^{[1]}$ is small (positive, but close to 0). Is the effect of the penalty on a gradient descent update step for $w_{11}^{[1]}$ larger for L_1 regularization or L_2 regularization?

The gradient descent update step for $w_{11}^{[1]}$ is $w_{11}^{[1]new} = w_{11}^{[1]old} - \alpha \cdot \frac{\partial}{\partial w_{11}^{[1]}} J(b,w)$

For L_1 regularization, this works out to

$$w_{11}^{[1]new} = w_{11}^{[1]old} - \alpha\left[\frac{-1}{m}\ell(b,w) + \lambda^{[1]}\right]$$

For L_2 regularization we obtain

$$w_{11}^{[1]new} = w_{11}^{[1]old} - \alpha\left[\frac{-1}{m}\ell(b,w) + 2\lambda^{[1]}w_{11}^{[1]}\right]$$

If $w_{11}^{[1]}$ is close to 0 the modification to the gradient descent step is smaller for L_2 regularization than for L_1 regularization, since in that case $2\lambda^{[1]}w_{11}^{[1]} < \lambda^{[1]}$

(c) In general, would you expect the final estimate of $w_{11}^{[1]}$ to be closer to 0 if you use L_1 regularization or L_2 regularization? You can answer in 1 sentence, but you should justify your answer based on your result from part (b).

This question was not precisely enough stated. If the estimate of $w_{11}^{[1]}$ we could obtain without regularization is close to 0 the penalized estimate we'd get from L_1 regularization would be closer to 0 than the penalized estimate from L_2 regularization, in general. This can be seen from the gradient descent updates above since if $w_{11}^{[1]}$ is near 0, L_1 regularization shrinks towards 0 more aggressively.