## Congratulations! You passed!

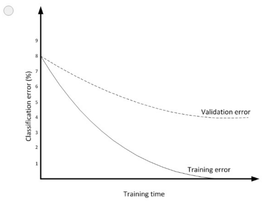
Next Item

1. 1 / 1 points

You are experimenting with two different models for a classification task. The figures below show the classification error you get as training progresses on the training data and the validation data for each of the two models. Which model do you think would perform better on previously unseen test data?.

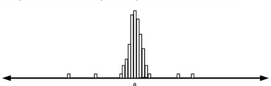

Correct

One model has got very low error on the training data but it has seriously overfitted, so its error on the validation data is much higher than on the test data and also higher than the validation error of the other model. The validation error is a much better predictor of test error than the training error, so we would expect the model with lower validation error to do better on test data.




2. 1 / 1 points

The figure below shows the histogram of weights for a learned Neural Network.




Which regularization technique has been used during learning?

- no regularization has been used
- adding weight noise
- L2 regularization

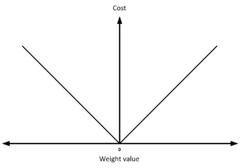
Correct

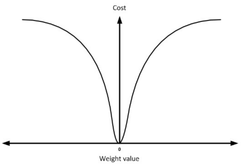
L1 regularization would have forced a lot of the weights to exactly zero. Their distribution looks more like a Gaussian around zero which is what you would expect for L2 regularization. The few weights that are many standard deviations from zero could be explained by the demands of the task. It might be necessary to make some weights quite unlikely under their Gaussian prior in order to make the outputs close to the target values.

- L1 regularization

3. 1 / 1 points

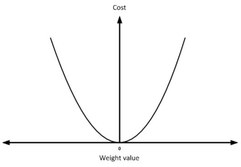
Suppose you want to regularize the weights of a neural network during training so that lots of its weights are quite close to zero, but a few are a very long way from zero. Which cost function you would add to your objective function?




Correct

L1 regularization would usually force many of the weights to be exactly zero. Strong L2 regularization could force a lot of weights to be close to zero, but it would then be extremely costly to have any very big weights.




4. 1 / 1 points

In a linear regression task, a $d$ dimensional input vector $x$ is used to predict the output value $y$ using the weight vector $w$ where $y = w^T x$. The error function $E = \frac{1}{2}(t - w^T x)^2$ where $t$ is the target output value. We want to use a student-t cost for the weights: $C = \frac{1}{2}\sum_{i=1}^{d} \log(1 + w_i^2)$.

The total error to be optimized $E_{tot} = E + C$. What is the expression for $\frac{\partial E_{tot}}{\partial w_i}$?

- $\frac{\partial E_{tot}}{\partial w_i} = -(t - w_i x_i) - \lambda w_i$
- $\frac{\partial E_{tot}}{\partial w_i} = -(t - w_i x_i) - 2\lambda \frac{w_i}{(1+w_i^2)}$
- $\frac{\partial E_{tot}}{\partial w_i} = -(t - w^T x) x_i + \frac{\lambda w_i}{(1+w_i^2)}$

Correct

First you compute $dE/dy$ then you multiply it by $dy/dw_i$ then you add on $dC/dw_i$

- $\frac{\partial E_{tot}}{\partial w_i} = -(t - w^T x) x_i + \frac{\lambda}{(1+w_i^2)}$

5. 1 / 1 points

Different regularization methods have different effects on the learning process. For example $L2$ regularization penalizes high weight values. $L1$ regularization penalizes weight values that do not equal zero. Adding noise to the weights during learning ensures that the learned hidden representations take extreme values. Sampling the hidden representations regularizes the network by pushing the hidden representation to be binary during the forward pass which limits the modeling capacity of the network.

Given the shown histogram of activations (just before the nonlinear logistic nonlinearity) for a Neural Network, what is the regularization method that has been used (check all that apply)?


- [x] Sampling the hidden representation

Correct

When you sample the hidden states, the sampling creates noise if the logistic is in its sensitive region. The learning tends to find solutions that minimize this noise by keeping units firmly on or firmly off.

- [ ] $L2$ regularization

Un-selected is correct

- [ ] $L1$ regularization

Un-selected is correct

- [x] Adding weight noise

Correct

Noise in the weights will make the outputs of the units noisy unless they are firmly on or firmly off. The learning will therefore tend to stop once the units behave like this.

6. 1 / 1 points

Suppose we have trained a neural network with one hidden layer and a single logistic output unit to predict whether or not an image contains a bird. If we retrain the network in the same way on the same data but using half as many hidden units, which of the following statements is true:

- [x] It will almost certainly do worse on the training data.

Correct

- [ ] It will almost certainly do better on the test data.

Un-selected is correct

- [ ] It will almost certainly do better on the training data.

Un-selected is correct

- [ ] It will almost certainly do worse on the test data.

Un-selected is correct