Regularization for Deep Learning

1 Reminding Backpropagation Algorithm

We re-explain backpropagation briefly in a mathematical form, and will present various methods for training and regularizing neural networks.

As introduced in previous week, the backpropagation is an approach for computing each neuron derivation. Consider Figure 1 as a neuron from a neural network. The symbol y refers to the activity (output) of a neuron. Similarly, the symbol z refers to the logit (input) of the neuron. We start by taking a look at the base case of the dynamic programming problem. Specifically, we calculate the error function derivatives at the output layer:

$$E = \frac{1}{2} \sum_{j \in \text{output}} (t_j - y_j)^2 \Rightarrow \frac{\partial E}{\partial y_j} = -(t_j - y_j)$$

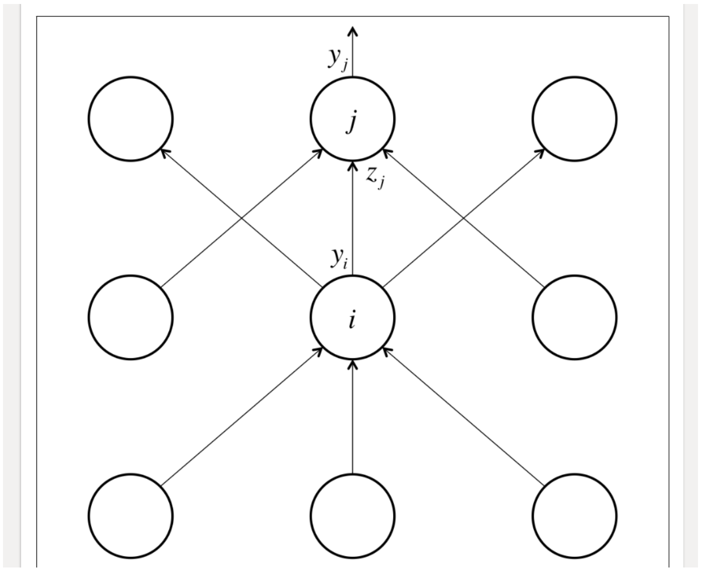

Figure 1: Diagram example for the derivation of the backpropagation algorithm.

Notice that t_j is the exact value and y_j is the predicted output of our i neuron. Now we tackle the inductive step. Let's presume we have the error derivatives for layer j . We now aim to calculate the error derivatives for the layer below it, layer i . To do so, we must accumulate information about how the output of a neuron in layer i affects the logits(inputs) of every neuron in layer j. This can be done as follows, using the fact that the partial derivative of the logit(input) with respect to the incoming output data from the layer beneath is merely the weight of the connection w_{ij} :

$$\frac{\partial E}{\partial y_j} = \sum_j \frac{\partial E}{\partial z_j}\frac{dz_j}{dy_i} = \sum_j w_{ij}\frac{\partial E}{\partial z_j}$$

Furthermore, we observe the following:

$$\frac{\partial E}{\partial z_j} = \frac{\partial E}{\partial y_j}\frac{dy_j}{dz_j} = y_j(1-y_j)\frac{\partial E}{\partial y_j}$$

Combining these two together, we can finally express the error derivatives of layer i in terms of the error derivatives of layer j:

$$\frac{\partial E}{\partial y_i} = \sum_j w_{ij} y_j (1-y_j)\frac{\partial E}{\partial y_j}$$

Then once we've gone through the whole dynamic programming routine, having filled up the table appropriately with all of our partial derivatives (of the error function with respect to the hidden unit activities), we can then determine how the error changes with respect to the weights. This gives us how to modify the weights after each training example:

$$\frac{\partial E}{\partial w_{ij}} = \frac{\partial z_j}{\partial w_{ij}}\frac{\partial E}{\partial z_j} = y_i y_j (1-y_j)\frac{\partial E}{\partial y_j}$$

Finally, to complete the algorithm, we merely sum up the partial derivatives over all the training examples in our dataset. This gives us the following modification formula:

$$\Delta w_{ij} = -\sum_{k \in \text{Dataset}} \epsilon y_i^{(k)} y_j^{(k)} (1-y_j^{(k)})\frac{\partial E^{(k)}}{\partial y_j^{(k)}}$$

This completes our description of the backpropagation algorithm!

2 Stochastic and Minibatch gradient Descent

In the "Backpropagation Algorithm", we've been using a version of gradient descent known as **batch gradient descent**. The idea behind batch gradient descent is that we use our entire dataset to compute the error surface and then follow the gradient to take the path of steepest descent. For a simple quadratic error surface, this works quite well. But in most cases, our error surface may be a lot more complicated. Consider the scenario in Figure 2 for illustration.

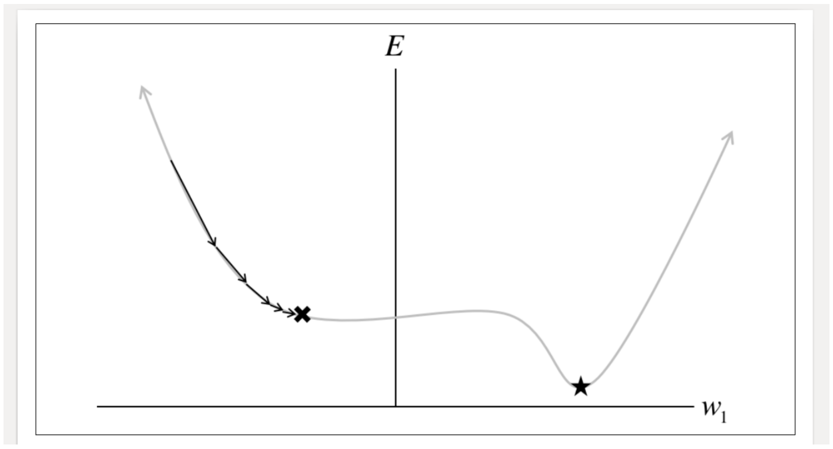

Figure 2: Batch gradient descent is sensitive to saddle points, which can lead to premature convergence.

We only have a single weight, and we use random initialization and batch gradient descent to find its optimal setting. The error surface, however, has a flat region (also known as saddle point in high-dimensional spaces), and if we get unlucky, we might find ourselves getting stuck while performing gradient descent.

Another potential approach is stochastic gradient descent (SGD), where at each iteration, our error surface is estimated only with respect to a single example. This approach is illustrated by Figure 3, where instead of a single static error surface, our error surface is dynamic. As a result, descending on this stochastic surface significantly improves our ability to navigate flat regions.

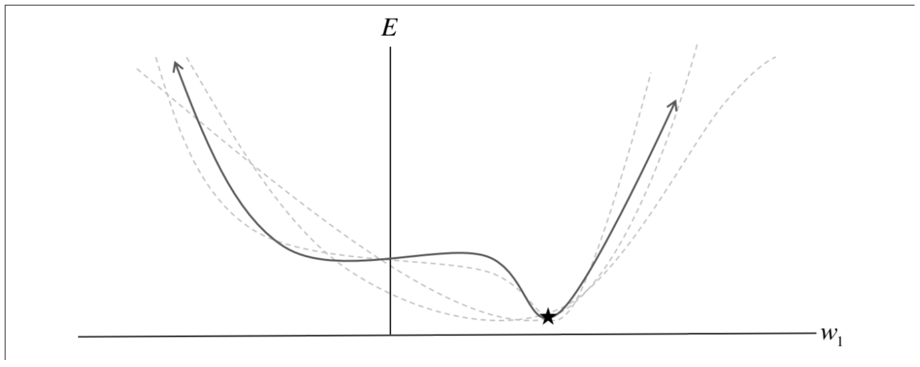

Figure 3: The stochastic error surface fluctuates with respect to the batch error surface, enabling saddle point avoidance.

The major pitfall of stochastic gradient descent, however, is that looking at the error incurred one example at a time may not be a good enough approximation of the error surface. This, in turn, could potentially make gradient descent take a significant amount of time. One way to combat this problem is using mini-batch gradient descent. In **mini-batch gradient descent**, at every iteration, we compute the error surface with respect to some subset of the total dataset (instead of just a single example). This subset is called a minibatch, and in addition to the learning rate, minibatch size is another hyperparameter. Minibatches strike a balance between the efficiency of batch gradient descent and the local-minima avoidance afforded by stochastic gradient descent. In the context of backpropagation, our weight update step becomes:

$$\Delta w_{ij} = - \sum_{k \in minibatch} \epsilon y_i^{(k)} y_j^{(k)} \left(1 - y_j(k) \frac{\partial E^{(k)}}{\partial y_j^{(k)}}\right)$$

This is identical to what we derived in the previous section, but instead of summing over all the examples in the dataset, we sum over the examples in the current minibatch.

3 Test Sets, Validation Sets, and Overfitting

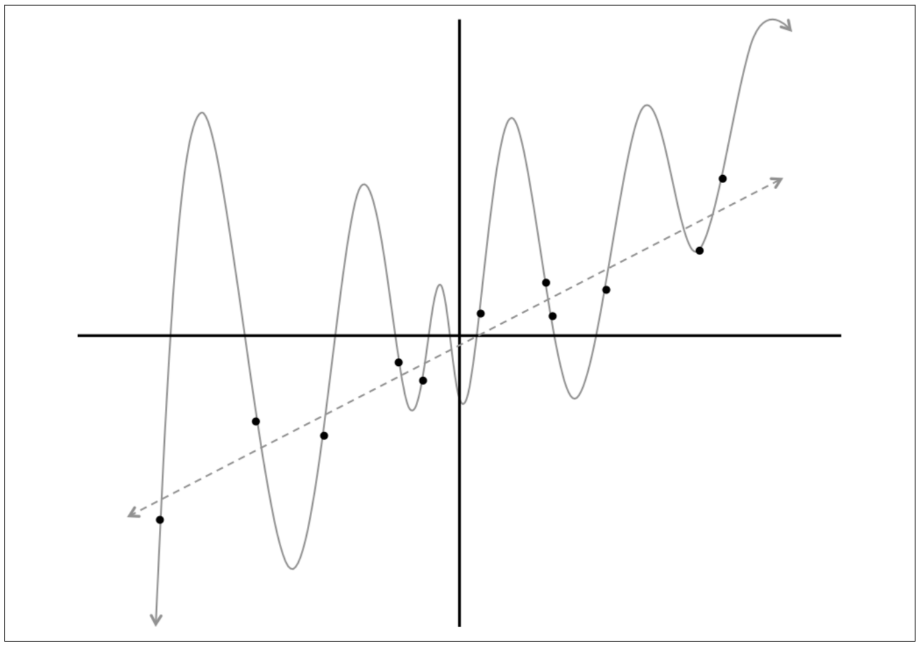

Figure 4: Two potential models that might describe our dataset: a linear model versus a degree 12 polynomial.

One of the major issues with artificial neural networks is that the models are quite complicated. For example, let's consider a neural network that's pulling data from an image from the MNIST database (28×28 pixels), feeds into two hidden layers with 30 neurons, and finally reaches a softmax layer of 10 neurons. The total number of parameters in the network is nearly 25000. This can be quite problematic, and to understand why, let's consider a new toy example, illustrated in Figure 5.

We are given a bunch of data points on a flat plane, and our goal is to find a curve that best describes this dataset (i.e., will allow us to predict the y-coordinate of a new point given its x-coordinate). Using the data, we train two different models: a linear model and a degree 12 polynomial. Which curve should we trust? The line which gets almost no training example correctly? Or the complicated curve that hits every single point in the dataset? At this point we might trust the linear fit because it seems much less contrived. But just to be sure, let's add more data to our dataset. The result is shown in Figure **??**.

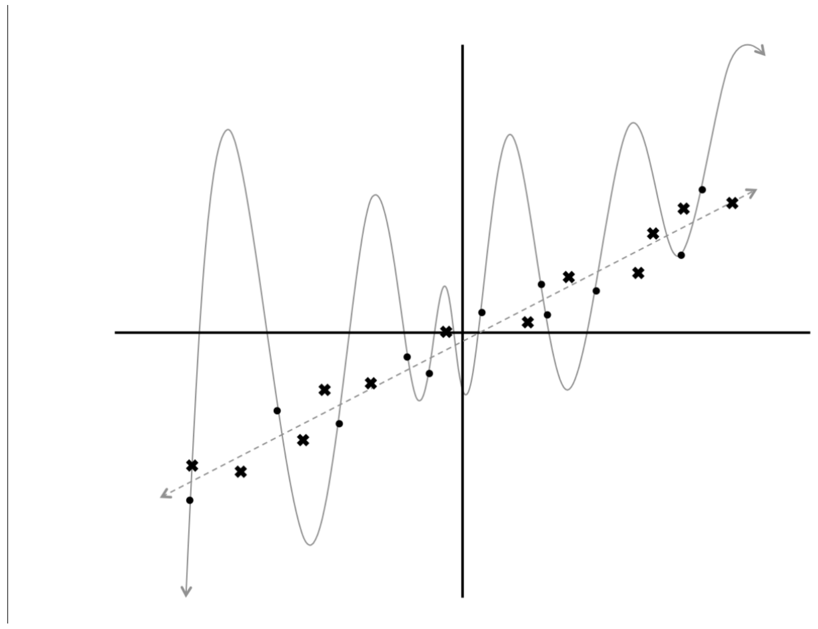

Figure 5: Evaluating our model on new data indicates that the linear fit is a much better model than the degree 12 polynomial.

Now the verdict is clear: the linear model is not only better subjectively but also quantitatively (measured using the squared error metric). But this leads to a very interesting point about training and evaluating machine learning models. By building a very complex model, it's quite easy to perfectly fit our training dataset because we give our model enough degrees of freedom to contort itself to fit the observations in the training set. But when we evaluate such a complex model on new data, it performs very poorly. In other words, the model does not **generalize** well. This is a phenomenon called **overfitting**, and it is one of the biggest challenges that a machine learning engineer must combat. This becomes an even more significant issue in deep learning, where our neural networks have large numbers of layers containing many neurons. The number of connections in these models is astronomical, reaching the millions. As a result, overfitting is commonplace.

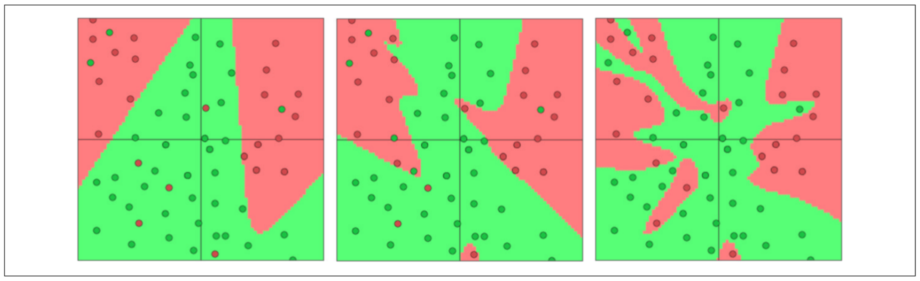

Figure 6: A visualization of neural networks with 3, 6, and 20 neurons (in that order) in their hidden layer.

Let's see how this looks in the context of a neural network. Let's say we have a neural network with two inputs, a softmax output of size two, and a hidden layer with 3, 6,or 20 neurons. We train these networks using mini-batch gradient descent (batch size 10), and the results, visualized using ConvNetJS[1], are shown in Figure 6.

It's already quite apparent from these images that as the number of connections in our network increases, so does our propensity to overfit to the data. We can similarly see the phenomenon of overfitting as we make our neural networks deep. These results are shown in Figure 7, where we use networks that have one, two, or four hidden layers of three neurons each.

This leads to three major observations. First, the machine learning engineer is always working with a direct trade-off between overfitting and model complexity. If the model isn't complex enough, it may not be powerful enough to capture all of the useful information necessary to solve a problem. However, if our model is very complex (especially if we have a limited amount of data at our disposal), we run the risk of overfitting. Deep learning takes the approach of solving very complex problems with complex models and taking additional countermeasures to prevent overfitting. We'll see a lot of these measures in this section.

Second, it is very misleading to evaluate a model using the data we used to train it. Using the example in Figure 5, this would falsely suggest that the degree 12 polynomial model is preferable to a linear fit. As a result, we almost

[1]`http://stanford.io/2pOdNhy`

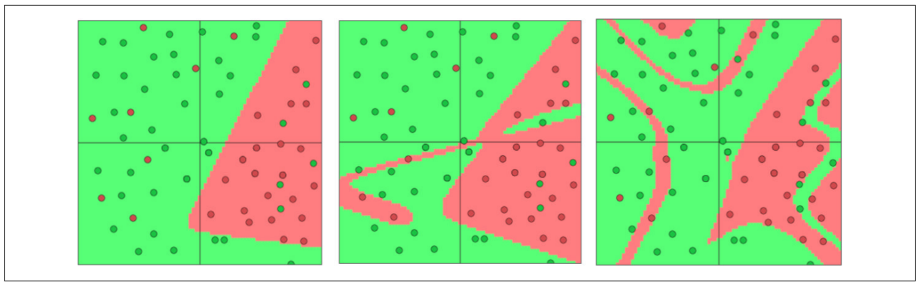

Figure 7: A visualization of neural networks with one, two, and four hidden layers (in that order) of three neurons each

never train our model on the entire dataset. Instead, as shown in Figure 8, we split up our data into a training set and a test set.

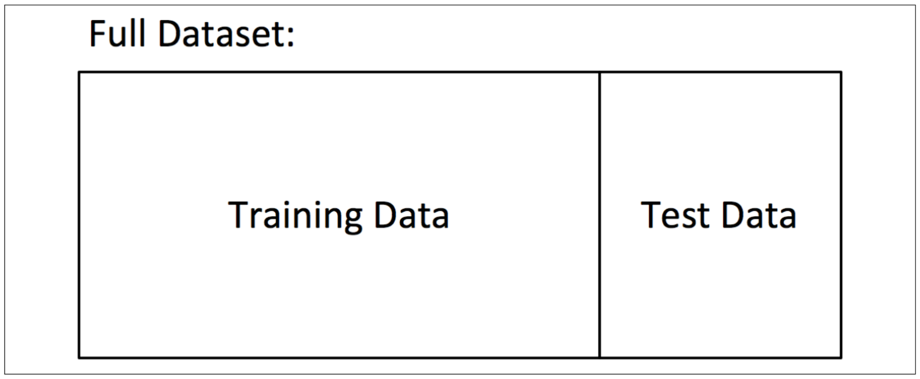

Figure 8: We often split our data into nonoverlapping training and test sets in order to fairly evaluate our model.

This enables us to make a fair evaluation of our model by directly measuring how well it generalizes on new data it has not yet seen. In the real world, large datasets are hard to come by, so it might seem like a waste to not use all of the data at our disposal during the training process. Consequently, it may be very tempting to reuse training data for testing or cut corners while compiling test data. Be forewarned: if the test set isn't well constructed, we won't be able draw any meaningful conclusions about our model.

Third, it's quite likely that while we're training our data, there's a point in time where instead of learning useful features, we start overfitting to the training set. To avoid that, we want to be able to stop the training process as soon as we start overfitting, to prevent poor generalization. To do this, we divide our training process into **epochs**. An epoch is a single iteration over the entire training set. In other words, if we have a training set of size d and we are doing mini-batch gradient descent with batch size b, then an epoch would be equivalent to $\frac{d}{b}$ model updates. At the end of each epoch, we want to measure how well our model is generalizing. To do this, we use an additional validation set, which is shown in Figure 9. At the end of an epoch, the validation set will tell us how the model does on data it has yet to see. If the accuracy on the training set continues to increase while the accuracy on the **validation set** stays the same (or decreases), it's a good sign that it's time to stop training because we're overfitting.

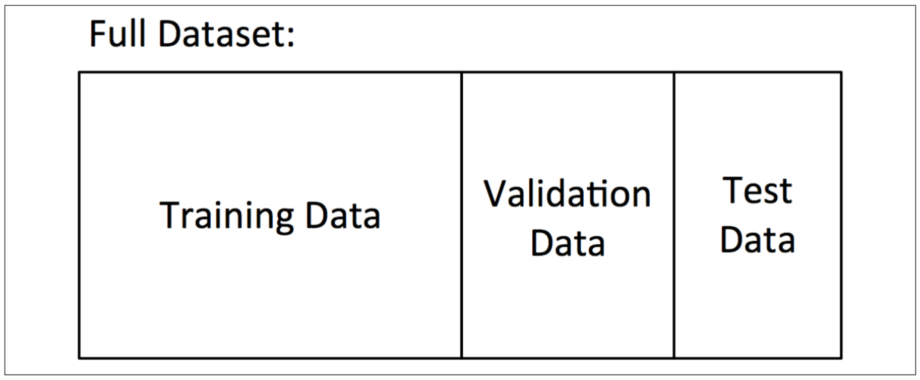

Figure 9: In deep learning we often include a validation set to prevent overfitting dur- ing the training process.

The validation set is also helpful as a proxy measure of accuracy during the process of **hyperparameter optimization**. We've covered several hyperparameters so far in our discussion (learning rate, minibatch size, etc.), but we have yet to develop a framework for how to find the optimal values for these hyperparameters. One potential way to find the optimal setting of hyperparameters is by applying a **grid search**, where we pick a value for each hyperparameter from a finite set of options (e.g., $\epsilon \in \{0.001, 0.01, 0.1\}$, batch size $\in \{16, 64, 128\}$ and etc), and train the model with every possible permutation of hyperparameter choices. We elect the combination of hyperparameters with the best performance on the validation set, and report the accuracy of the model trained with best combination on the test set.

With this in mind, before we jump into describing the various ways to directly combat overfitting, let's outline the workflow we use when building and training deep learning models. The workflow is described in detail in Figure 10. It is a tad intricate, but it's critical to understand the pipeline in order to ensure that we're properly training our neural networks.

First we define our problem rigorously. This involves determining our inputs, the potential outputs, and the vectorized representations of both. For instance, assume our goal was to train a deep learning model to identify cancer. Our input

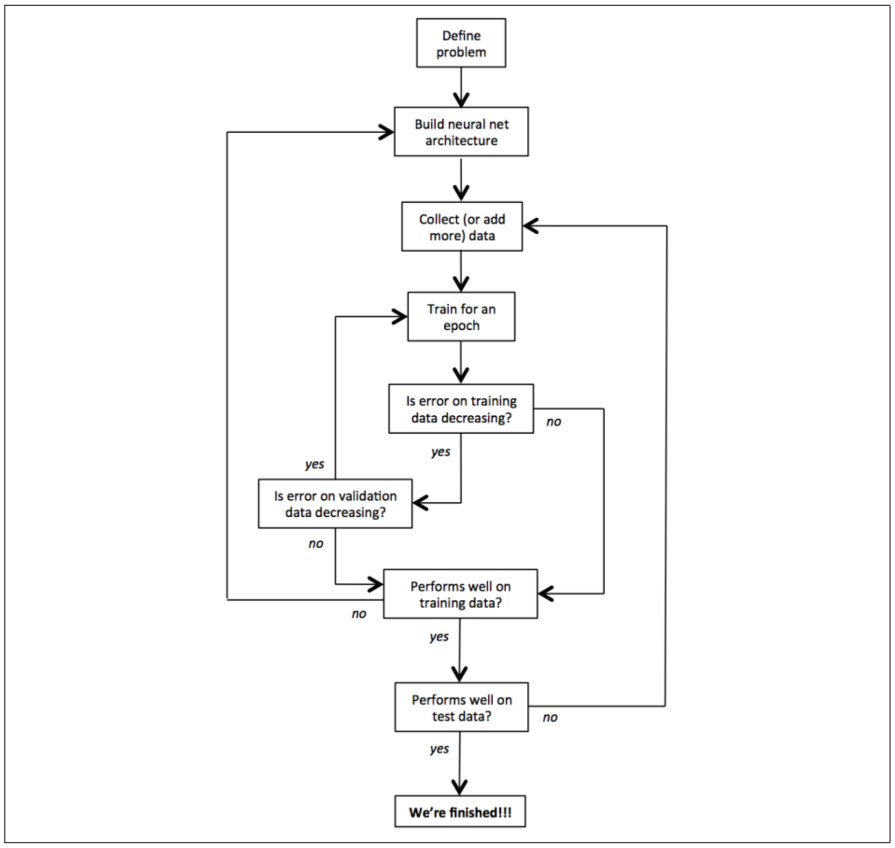
Define problem
Build neural net architecture
Collect (or add more) data
Train for an epoch
Is error on training data decreasing?
yes
no
Is error on validation data decreasing?
yes
no
Performs well on training data?
no
yes
Performs well on test data?
no
yes
We're finished!!!

Figure 10: Detailed workflow for training and evaluating a deep learning model.

would be an RBG image, which can be represented as a vector of pixel values. Our output would be a probability distribution over three mutually exclusive possibilities: 1) normal, 2) benign tumor (a cancer that has yet to metastasize), or 3) malignant tumor (a cancer that has already metastasized to other organs).

After we define our problem, we need to build a neural network architecture to solve it. Our input layer would have to be of appropriate size to accept the raw data from the image, and our output layer would have to be a softmax of size 3. We will also have to define the internal architecture of the network (number of hidden layers, the connectivities, etc.). We'll further discuss the architecture of image recognition models when we talk about convolutional neural networks in future sections. At this point, we also want to collect a significant amount of data for training or modeling. This data would probably be in the form of uniformly sized pathological images that have been labeled by a medical expert. We shuffle and divide this data up into separate training, validation, and test sets.

Finally, we are ready to begin gradient descent. We train the model on our training set for an epoch at a time. At the end of each epoch, we ensure that our error on the training set and validation set is decreasing. When one of these stops to improve, we terminate and make sure we are happy with the model's performance on the test data. If we are unsatisfied, we need to rethink our architecture or reconsider whether the data we collect has the information required to make the prediction we're interested in making. If our training set error stopped improving, we probably need to do a better job of capturing the important features in our data. If our validation set error stopped improving, we probably need to take measures to prevent overfitting.

If, however, we are happy with the performance of our model on the training data, then we can measure its performance on the test data, which the model has never seen before this point. If it is unsatisfactory, we need more data in our dataset because the test set seems to consist of example types that weren't well represented in the training set. Otherwise, we are finished.

4 Preventing Overfitting in Deep Neural Networks

There are several techniques that have been proposed to prevent overfitting during the training process. One method of combatting overfitting is called **regularization**. Regularization modifies the objective function that we minimize by adding additional terms that penalize large weights. In other words, we change the objective function so that it becomes $Error + \lambda f(\theta)$, where $f(\theta)$ grows larger as the components of θ grow larger, and λ is the regularization strength (another hyperparameter). The value we choose for λ determines how much we want to protect against overfitting. A $\lambda = 0$ implies that we do not take any measures against the possibility of overfitting. If λ is too large, then our model will prioritize keeping θ as small as possible over trying to find the parameter values that perform well on our training set. As a result, choosing $\lambda$is a very important task and can require some trial and error.

The most common type of regularization in machine learning is **L_2 regularization**. It can be implemented by augmenting the error function with the squared magnitude of all weights in the neural network. In other words, for every weight w in the neural network, we add $\frac{1}{2}\lambda w^2$ to the error function. The L_2 regularization has the intuitive interpretation of heavily penalizing peaky weight vectors and preferring diffuse weight vectors. This has the appealing property of encouraging the network to use all of its inputs a little rather than using only some of its inputs a lot. Of particular note is that during the gradient descent update, using the L_2 regularization ultimately means that every weight is decayed linearly to zero. Because of this phenomenon, L_2 regularization is also commonly referred to as **weight decay**.

We can visualize the effects of L_2 regularization using ConvNetJS. Similar to Figures 6 and 7, we use a neural network with 2 inputs, a softmax output of size

2, and a hidden layer with 20 neurons. We train the networks using mini-batch gradient descent (batch size 10) and regularization strengths of $0.01, 0.1$, and 1. The results can be seen in Figure 11.

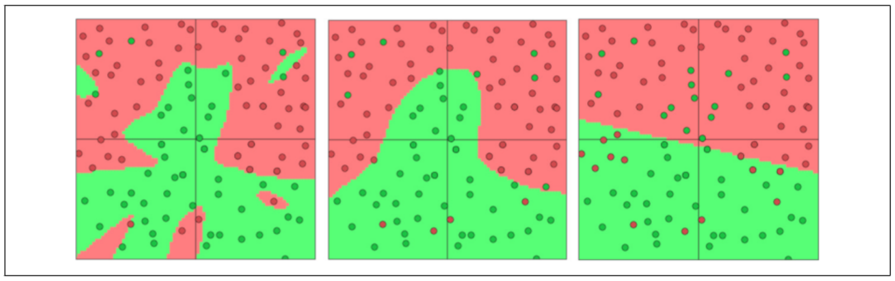

Figure 11: A visualization of neural networks trained with regularization strengths of $0.01, 0.1$, and 1 (in that order).

Another common type of regularization is **$\mathbf{L_1}$ regularization**. Here, we add the term $\lambda|w|$ for every weight w in the neural network. The L_1 regularization has the intriguing property that it leads the weight vectors to become sparse during optimization (i.e., very close to exactly zero). In other words, neurons with L_1 regularization end up using only a small subset of their most important inputs and become quite **resistant to noise** in the inputs. In comparison, weight vectors from L_2 regularization are usually diffuse, small numbers. L_1 regularization is very useful when you want to understand exactly which features are contributing to a decision. If this level of feature analysis isn't necessary, we prefer to use L_2 regularization because it empirically performs better.

Max norm constraints have a similar goal of attempting to restrict θ from becoming too large, but they do this more directly. Max norm constraints enforce an absolute upper bound on the magnitude of the incoming weight vector for every neuron and use projected gradient descent to enforce the constraint. In other words, any time a gradient descent step moves the incoming weight vector such that $||w|||_2 > c$, we project the vector back onto the ball (centered at the origin) with radius c . Typical values of c are 3 and 4. One of the nice properties is that the parameter vector cannot grow out of control (even if the learning rates are too high) because the updates to the weights are always bounded.

Dropout is a very different kind of method for preventing overfitting that has become one of the most favored methods of preventing overfitting in deep neural networks. While training, dropout is implemented by only keeping a neuron active with some probability p (a hyperparameter), or setting it to zero otherwise. Intuitively, this forces the network to be accurate even in the absence of certain information. It prevents the network from becoming too dependent on any one (or any small combination) of neurons. Expressed more mathematically, it prevents overfitting by providing a way of approximately combining exponentially many different neural network architectures efficiently. The process of dropout is expressed pictorially in Figure 12.

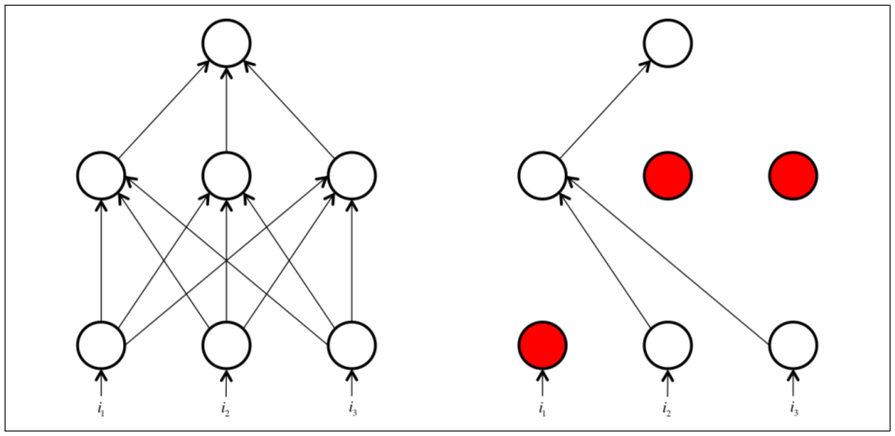

Figure 12: Dropout sets each neuron in the network as inactive with some random probability during each minibatch of training.

Dropout is pretty intuitive to understand, but there are some important intricacies to consider. First, we'd like the outputs of neurons during test time to be equivalent to their expected outputs at training time. We could fix this naively by scaling the output at test time. For example, if $p = 0.5$, neurons must halve their outputs at test time in order to have the same (expected) output they would have during training. This is easy to see because a neuron's output is set to 0 with probability $1 - p$. This means that if a neuron's output prior to dropout was x, then after dropout, the expected output would be $E[\text{output}] = px + (1 - p) \cdot 0 = px$. This naive implementation of dropout is undesirable, however, because it requires scaling of neuron outputs at test time. Test-time performance is extremely critical to model evaluation, so it's always preferable to use inverted dropout, where the scaling occurs at training time instead of at test time. In inverted dropout, any neuron whose activation hasn't been silenced has its output divided by p before the value is propagated to the next layer. With this fix, $E[\text{output}] = p\frac{x}{p} + (1-p) \cdot 0 = x$ and we can avoid arbitrarily scaling neuronal output at test time.