BST 261: Data Science II
Lecture 3

Feedforward networks in Python with Keras, Regularization

Heather Mattie
Harvard T.H. Chan School of Public Health

Spring 2 2021

Recipe of the day

Mustikkapiirakka

(Finnish blueberry pie)

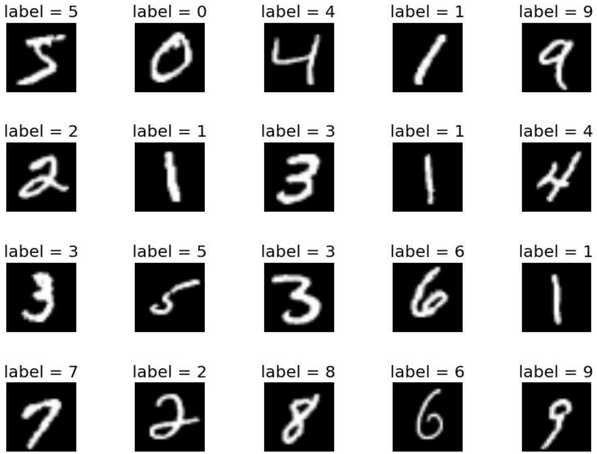
MLPs in Python/Keras

MNIST Data Example

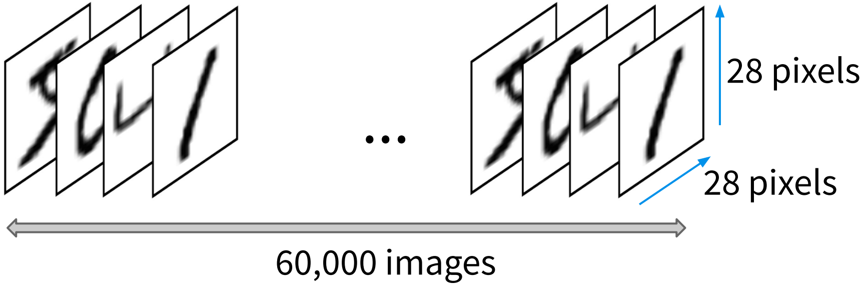
- The MNIST data set includes handwritten digits with corresponding labels

- Training set: 60,000 images of handwritten digits and corresponding labels
  - Each digit is represented as a 28 x 28 matrix of grayscale values 0 - 255
  - The entire training set is stored in a 3D tensor of shape (60000, 28, 28)
  - The corresponding image values are stored as a 1D tensor of values 0 - 9

- Testing set: 10,000 images with the same set up as the training set



Training Data



MNIST Data Example

Data wrangling

- We'll get into RGB images later, but for grayscale images, we need to first transform the matrix of values into a vector of values, and then normalize them to be between 0 and 1. It is not strictly necessary to normalize your inputs, but smaller numbers help speed up training and avoid getting stuck in local minima. This also ensures the gradients don't "explode" or "vanish"
  - Reshape each image from a 28 x 28 matrix of grayscale values 0 - 255 to a vector of length 28*28 = 784 of values 0 - 1 (divide each by 255)

- We now have 10 classes (categories; the digits 0-9)
  - We need to have multiclass labels that tell the network which digit the example is
  - Reshape each corresponding image label to a vector of length 10 of values 0 or 1
  - Example: the digit 3 would be represented as [0, 0, 0, 1, 0, 0, 0, 0, 0, 0]
  - You can think of this as "dummy coding" the labels

Activation and Loss Function Choices

Task	Last-layer activation	Loss function
Binary classification	sigmoid	Binary cross-entropy
Multiclass, single-label classification	softmax	Categorical cross-entropy
Multiclass, multilabel classification	sigmoid	Binary cross-entropy
Regression to arbitrary values	None	Mean square error (MSE)
Regression to values between 0 and 1	sigmoid	MSE or binary cross-entropy

Softmax function

$$\mathrm{softmax}(\boldsymbol{z})_i = \frac{\exp(z_i)}{\sum_j \exp(z_j)}$$

- Softmax units are used as outputs when predicting a discrete variable *y* with *j* possible values
- In this setting, which can be seen as a generalization of the Bernoulli distribution, we need to produce a vector $\hat{\mathbf{y}}$ with $\hat{y}_i = P(y = i|x)$
- We require that each $\hat{y}_i$ lie in the [0, 1] interval and that the entire vector sums to 1
- We first compute $z = w^T x + b$ as usual
- Here, $z_i = log[\tilde{P}(y = i|x)]$ represents an unnormalized log probability for class *i*
- The softmax function then exponentiates and normalizes *z* to obtain $\hat{\mathbf{y}}$

Categorical cross-entropy

- In this case we want to maximize

$$log[P(y = i; z)] = log[\text{softmax}(z)_i] = z_i - log \sum_j exp(z_j)$$

- The first term shows that the input always has a direct contribution to the loss function

- Because $log \sum_j exp(z_j) \approx max_j z_j$, the negative log-likelihood loss function always strongly penalizes the most active incorrect prediction

MNIST Data Example

Network Architecture

- Let’s start with 2 layers:
  - Hidden layer will have 512 hidden units and the **relu activation function**

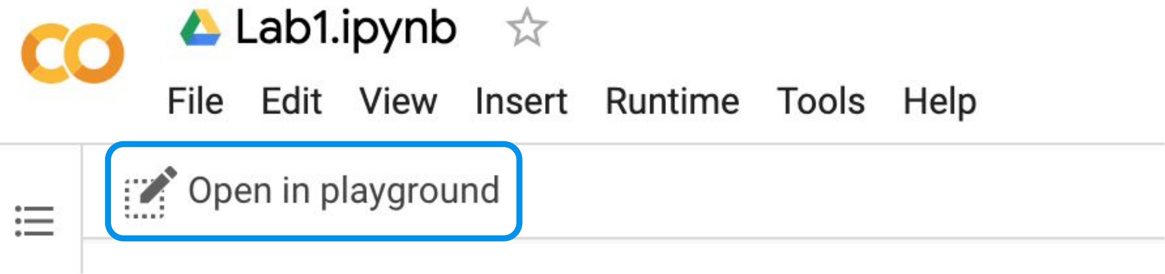
  - Output layer with 10 units (one for each possible digit) and the **softmax activation function** (this produces a vector of length 10, where each element is a probability between 0 and 1 of the image being classified as that digit)
  - Example: [0, 0.3, 0, 0, 0, 0, 0, 0.7, 0, 0] - the highest probability corresponds to a label of 7, so the network would classify this image as a 7

  - **rmsprop optimization algorithm**
  - **categorical_crossentropy loss function**
  - **accuracy performance measure** (the proportion of times the correct class is chosen)

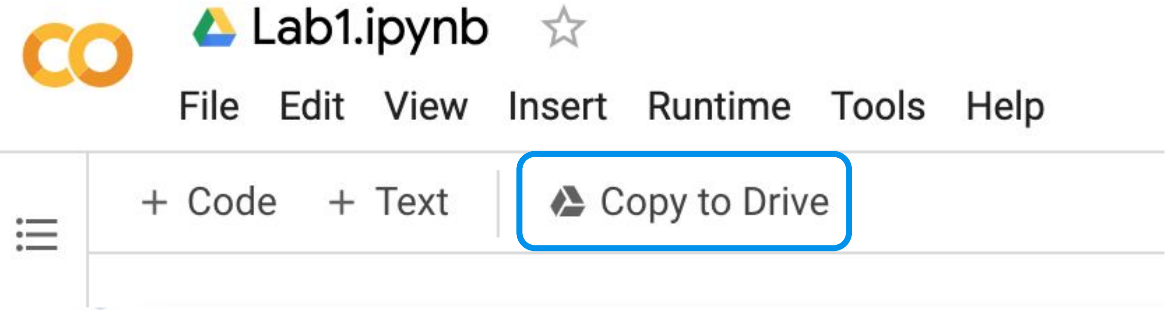
MNIST Data Example

Colab link

Step 1



Step 2



IMDb Data Example

The IMDb data set is a set of movie reviews that have been labeled as either positive or negative, based on the text content of the reviews

- Training set: 25,000 either positive or negative movie reviews that have each been turned into a vector of integers
  - We'll see how to actually do this later in the course
  - Each review can be of any length
  - Only the top 10,000 most frequently occurring words are kept i.e. rare words are discarded
  - Each review includes a label: 0 = negative review and 1 = positive review

- Testing set: 25,000 either positive or negative movie reviews, similar to the training set

IMDb Data Example

Data Wrangling

- Each review is of a varying length and is a list of integers - we need to turn this into a tensor with a common length for each review
- Create a 2D tensor of shape 25,000 x 10,000
  - 25,000 reviews and 10,000 possible words
- Use the **vectorize_sequences** function to turn a movie review list of integers into a vector of length 10,000 with 1s for each word that appears in the review and 0s for words that do not
- The labels are already 0s and 1s, so the only thing we need to do is make them float numbers

Activation and Loss Function Choices

Task	Last-layer activation	Loss function
Binary classification	sigmoid	Binary cross-entropy
Multiclass, single-label classification	softmax	Categorical cross-entropy
Multiclass, multilabel classification	sigmoid	Binary cross-entropy
Regression to arbitrary values	None	Mean square error (MSE)
Regression to values between 0 and 1	sigmoid	MSE or binary cross-entropy

IMDb Data Example

Network Architecture

- 3 layers
  - 2 hidden layers and 1 output layer
  - Hidden layers have 16 hidden units each and a **relu activation function**
  - Output layer has 1 unit (the probability a review is positive)
- **Sigmoid activation function**
- **rmsprop optimization algorithm**
- **binary_crossentropy loss function**
- **accuracy performance measure** (proportion of times the correct class is chosen)

IMDb Data Example

Colab link

Regularization

Regularization

- One of the biggest problems with neural networks is overfitting.
- Regularization schemes combat overfitting in a variety of different ways

A perceptron represents the following optimization problem:

$$\mathrm{argmin}_W l(y, f(X)) \quad \text{where} \quad f(X) = \frac{1}{1+\exp(-\phi(XW))} \text{ (for example)}$$

Regularization

One way to regularize is to introduce penalties and change

$$\text{argmin}_W l(y, f(X))$$

to

$$\text{argmin}_W l(y, f(X)) + \lambda R(W)$$

where R(W) is often the L1 or L2 norm of W. These are the well-known ridge and LASSO penalties, and referred to as **weight decay** by the neural net community.

L2 Regularization

We can limit the size of the L2 norm of the weight vector:

$$\text{argmin}_W l(y, f(X)) + \lambda \|W\|_2$$

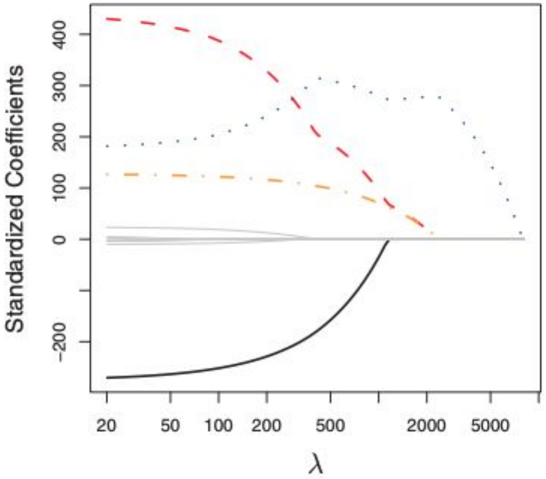
where

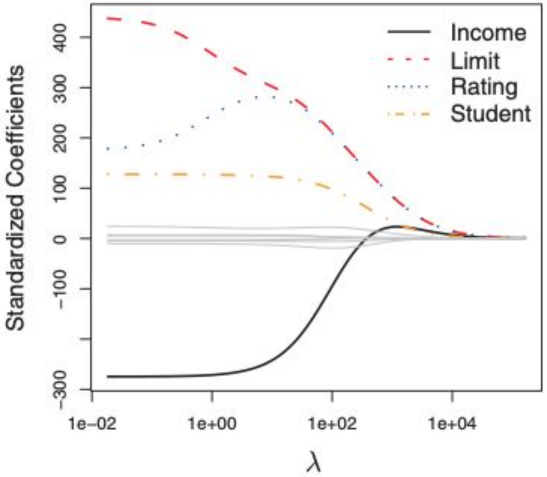
$$\|W\|_2 = \sum_{j=1}^{p} w_j^2$$

We can do the same for the L1 norm. What do the penalties do?

Shrinkage

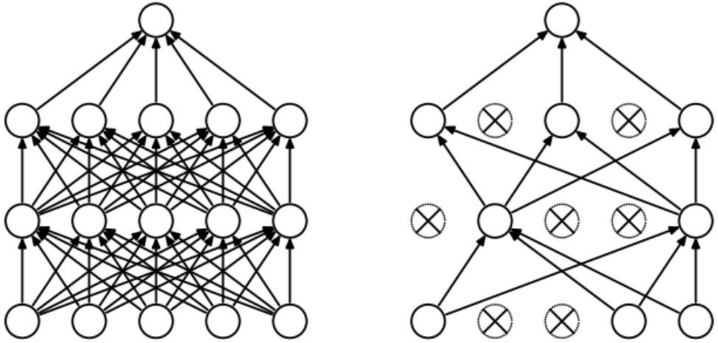
The L1 and L2 penalties shrink the weights to or towards 0.





Stochastic Regularization

- Why is this a good idea?
- One of the most popular ways to do this is **dropout**
- Given a hidden layer, we are going to set each element of the hidden layer to 0 with probability *p* each SGD update.

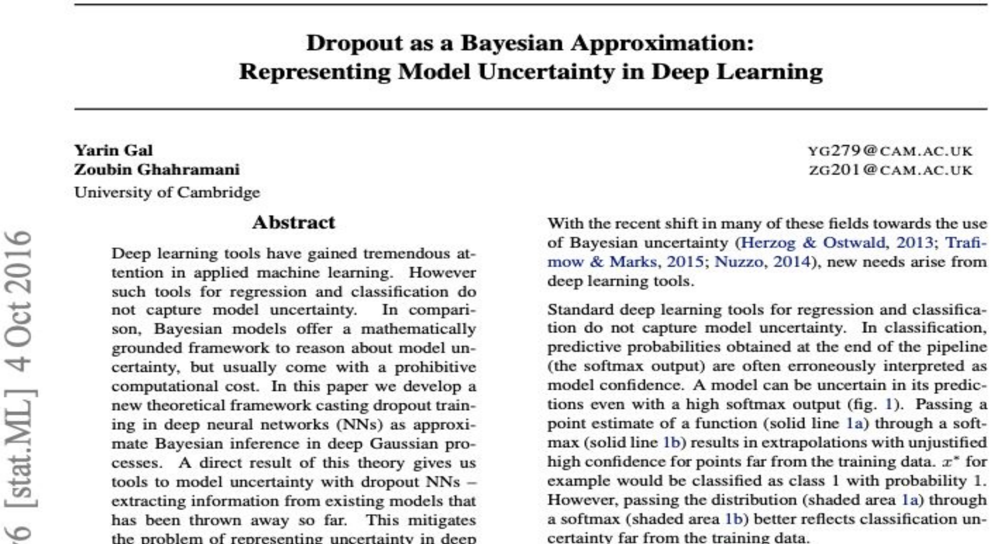


(a) Standard Neural Net (b) After applying dropout.

https://medium.com/@amarbudhiraja/https-medium-com-amarbudhiraja-learning-less-to-learn-better-dropout-in-deep-machine-learning-74334da4bfc5

Stochastic Regularization

- One way to think of this is the network is trained by bagged versions of the network.
- **Bagging** reduces variance.
- Others have argued this is an approximate Bayesian model



Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning

Yarin Gal YG279@CAM.AC.UK
Zoubin Ghahramani ZG201@CAM.AC.UK
University of Cambridge

6 [stat.ML] 4 Oct 2016

Abstract

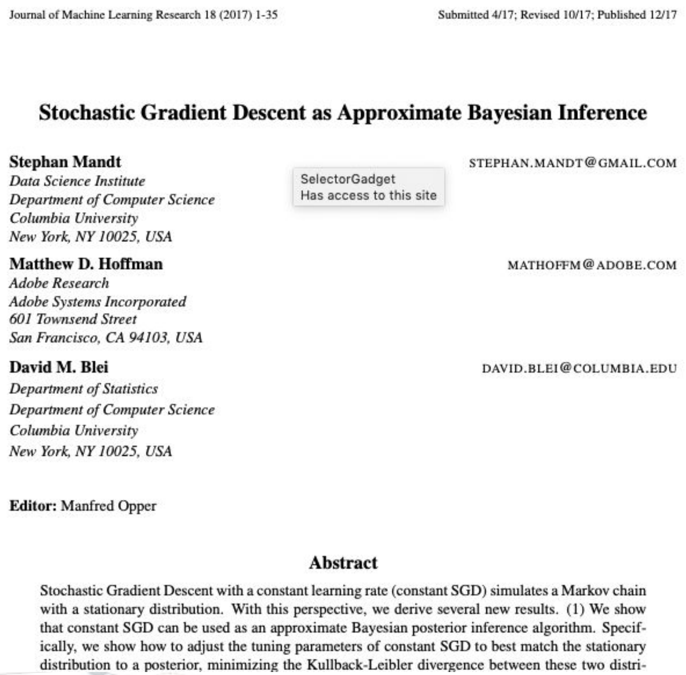
Deep learning tools have gained tremendous attention in applied machine learning. However such tools for regression and classification do not capture model uncertainty. In comparison, Bayesian models offer a mathematically grounded framework to reason about model uncertainty, but usually come with a prohibitive computational cost. In this paper we develop a new theoretical framework casting dropout training in deep neural networks (NNs) as approximate Bayesian inference in deep Gaussian processes. A direct result of this theory gives us tools to model uncertainty with dropout NNs – extracting information from existing models that has been thrown away so far. This mitigates the problem of representing uncertainty in deep

With the recent shift in many of these fields towards the use of Bayesian uncertainty (Herzog & Ostwald, 2013; Trafimow & Marks, 2015; Nuzzo, 2014), new needs arise from deep learning tools.

Standard deep learning tools for regression and classification do not capture model uncertainty. In classification, predictive probabilities obtained at the end of the pipeline (the softmax output) are often erroneously interpreted as model confidence. A model can be uncertain in its predictions even with a high softmax output (fig. 1). Passing a point estimate of a function (solid line 1a) through a softmax (solid line 1b) results in extrapolations with unjustified high confidence for points far from the training data. x^* for example would be classified as class 1 with probability 1. However, passing the distribution (shaded area 1a) through a softmax (shaded area 1b) better reflects classification uncertainty far from the training data.

Stochastic Regularization

- Many have argued that SGD itself provides regularization



Journal of Machine Learning Research 18 (2017) 1-35 Submitted 4/17; Revised 10/17; Published 12/17

Stochastic Gradient Descent as Approximate Bayesian Inference

Stephan Mandt STEPHAN.MANDT@GMAIL.COM
Data Science Institute
Department of Computer Science
Columbia University
New York, NY 10025, USA

SelectorGadget
Has access to this site

Matthew D. Hoffman MATHOFFM@ADOBE.COM
Adobe Research
Adobe Systems Incorporated
601 Townsend Street
San Francisco, CA 94103, USA

David M. Blei DAVID.BLEI@COLUMBIA.EDU
Department of Statistics
Department of Computer Science
Columbia University
New York, NY 10025, USA

Editor: Manfred Opper

Abstract

Stochastic Gradient Descent with a constant learning rate (constant SGD) simulates a Markov chain with a stationary distribution. With this perspective, we derive several new results. (1) We show that constant SGD can be used as an approximate Bayesian posterior inference algorithm. Specifically, we show how to adjust the tuning parameters of constant SGD to best match the stationary distribution to a posterior, minimizing the Kullback-Leibler divergence between these two distri-

Initialization Regularization

- The weights in a neural network are given random values initially.
- There is an entire literature on the best way to do this initialization
  - Normal
  - Truncated Normal
  - Uniform
  - Orthogonal
  - Scaled by number of connections
  - Etc.
- Try to “bias” the model into initial configurations that are easier to train

Initialization Regularization

- A popular way is to do **transfer learning**

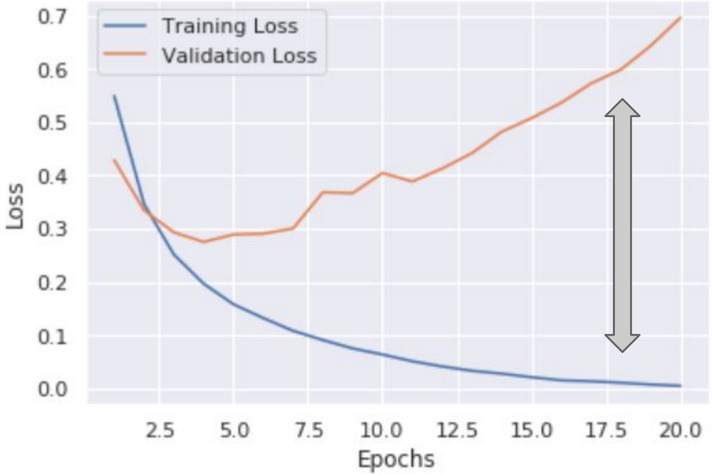
Train the model on auxiliary task where lots of data is available

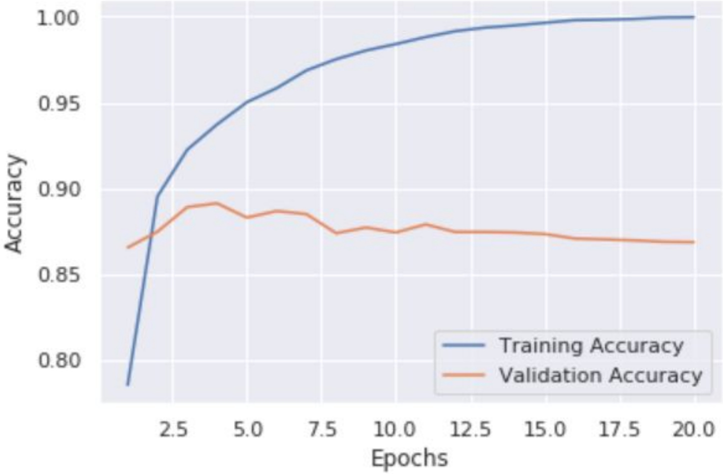
→

Use final weight values from previous task as initial values and “fine tune” on primary task

IMDb Example

We saw overfitting in the IMDb example:





How do we make this model better?

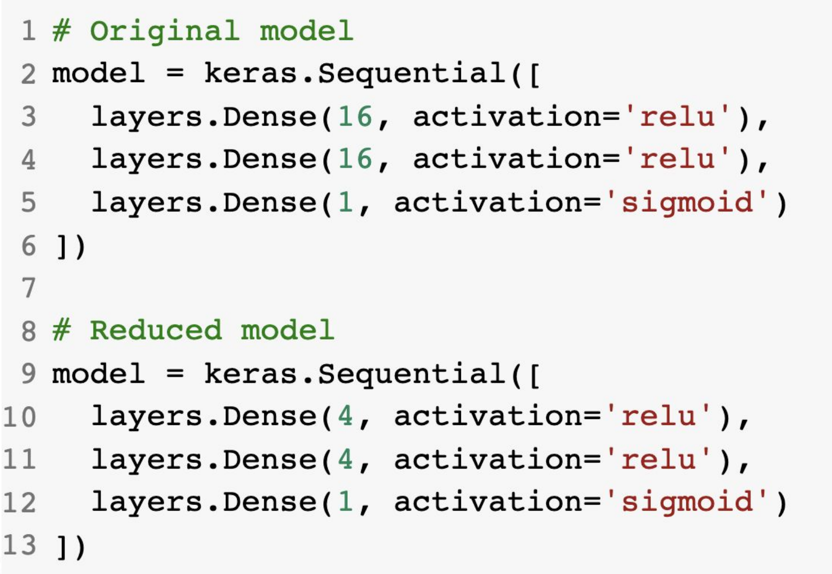
Regularization

1. Reduce network size
2. Weight regularization
3. Dropout

Back to the IMDb colab notebook

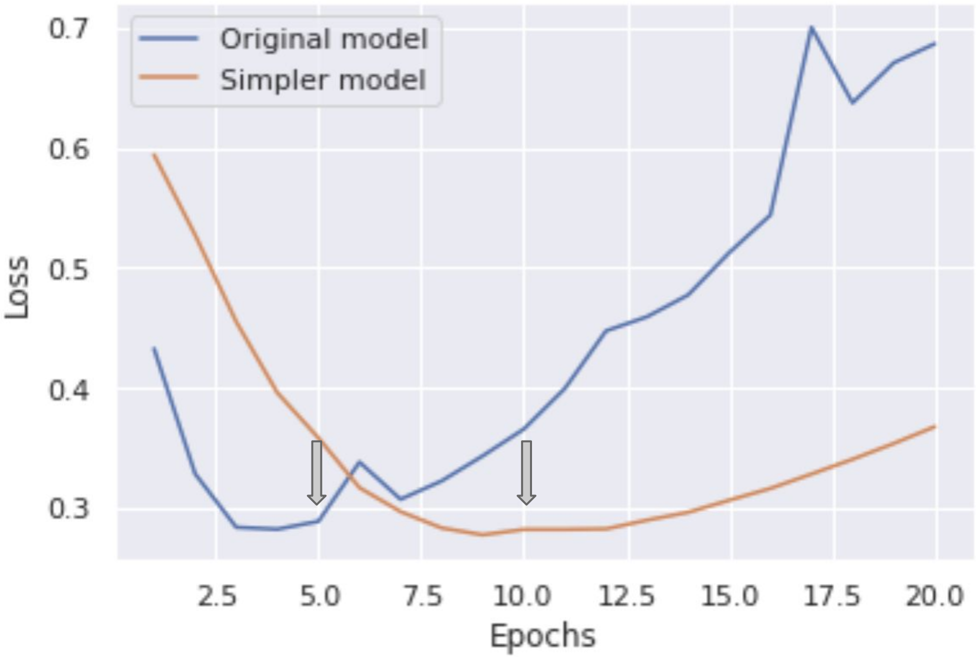
Regularization: reducing network size

When we are battling overfitting, one option is to simplify the model. Let's compare the performance we get from a simpler model. Here we have simplified the model by reducing the number of hidden units in each hidden layer.



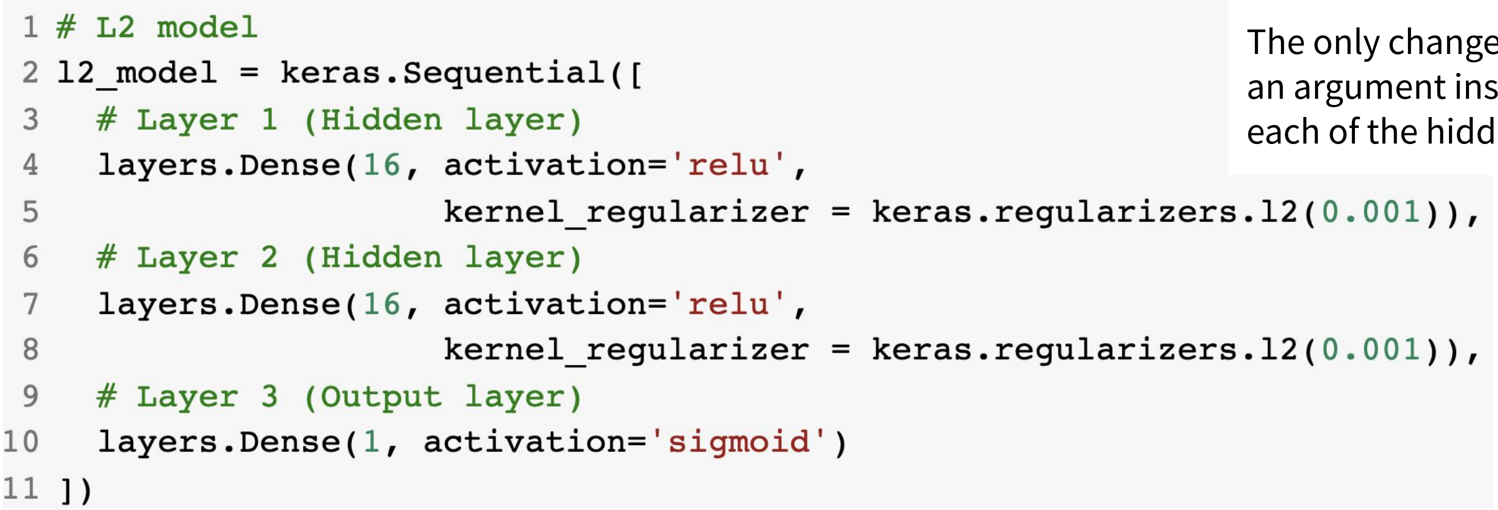
```
# Original model
model = keras.Sequential([
  layers.Dense(16, activation='relu'),
  layers.Dense(16, activation='relu'),
  layers.Dense(1, activation='sigmoid')
])

# Reduced model
model = keras.Sequential([
  layers.Dense(4, activation='relu'),
  layers.Dense(4, activation='relu'),
  layers.Dense(1, activation='sigmoid')
])
```

Regularization: reducing network size



The smaller network performs better than the original model - it starts to overfit at epoch 10 rather than epoch 6. These values are when the validation loss starts to increase.

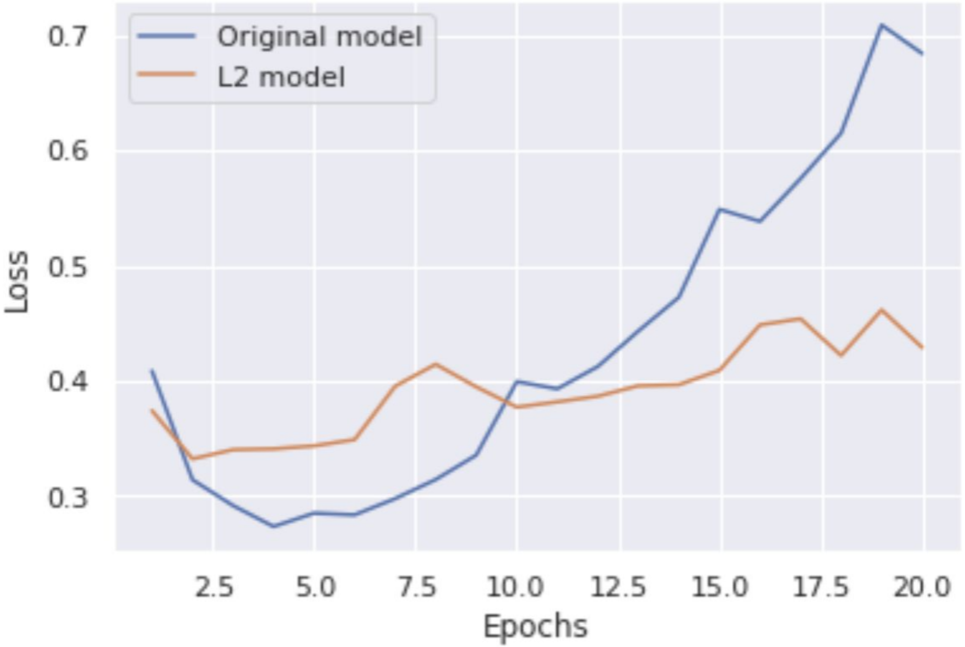
Regularization: weight regularization



```
# L2 model
l2_model = keras.Sequential([
  # Layer 1 (Hidden layer)
  layers.Dense(16, activation='relu',
                     kernel_regularizer = keras.regularizers.l2(0.001)),
  # Layer 2 (Hidden layer)
  layers.Dense(16, activation='relu',
                     kernel_regularizer = keras.regularizers.l2(0.001)),
  # Layer 3 (Output layer)
  layers.Dense(1, activation='sigmoid')
])
```

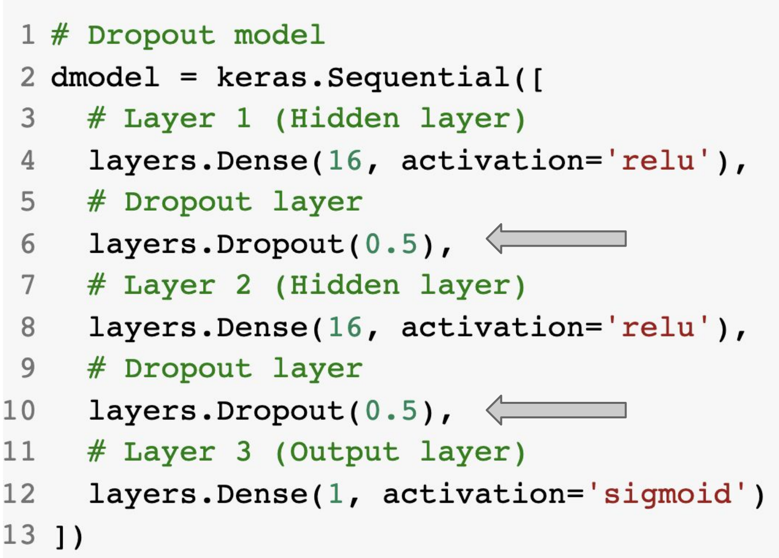
The only change is adding an argument inside of each of the hidden layers

Regularization: weight regularization



The L2-regularized model is much more resistant to overfitting - the validation loss starts to increase at a much slower rate

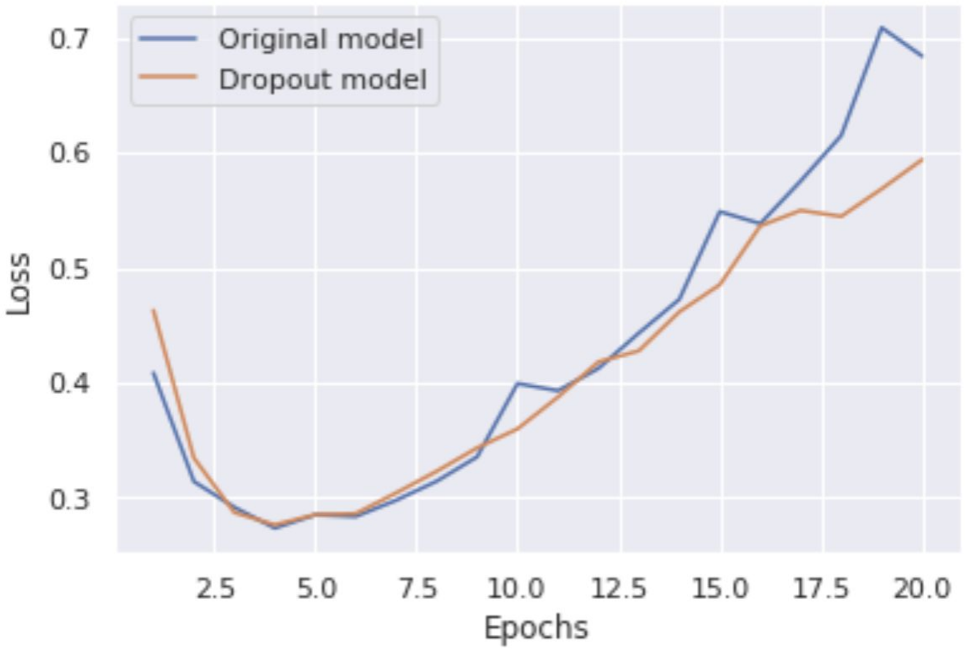
Regularization: adding dropout



```
# Dropout model
dmodel = keras.Sequential([
  # Layer 1 (Hidden layer)
  layers.Dense(16, activation='relu'),
  # Dropout layer
  layers.Dropout(0.5),
  # Layer 2 (Hidden layer)
  layers.Dense(16, activation='relu'),
  # Dropout layer
  layers.Dropout(0.5),
  # Layer 3 (Output layer)
  layers.Dense(1, activation='sigmoid')
])
```

The 0.5 indicates a 50% probability of dropping out a unit. Typically, 20% is used in practice but you can try different values and see what performs best.

Regularization: adding dropout



The dropout model is slightly better than the original model (in terms of overfitting) but does not control for overfitting as well as the L2 network