Pointers for Final Assignment:

1. Choose at least two datasets – Can go with MNIST and CIFAR10 as Keras models are easily available for them online.
2. Choose at least two models – I’d recommend going with both keras models as optimizers are implemented for it. Adam was used in week 8 assignment and for constant step size we can use SGD optimizer without Nesterov. One can be CNN, the code for which we already have. Another can be Logistic Regression, as used in the papers. Resources are easily available to build Logistic Regression using keras as well. (Not adding them so that not everyone uses the same link.)
3. If using MNIST dataset, it does not have channels dimension. Refer<https://keras.io/examples/vision/mnist_convnet/> to convert it.
4. We need to use SGD as the main function to shuffle the dataset as in week 6 and pass the choice of optimizer next, Adam or constant (SGD). Essentially, we are comparing model performance on different datasets using different optimizers. And not which model performs best on a given dataset.
5. Optimize Hyper parameters using Global Pop Search as in week 8.
6. For anyone struggling with computing power, use colab and enable GPU. Typically trains on MNIST and CIFAR in less than a minute, 20 epochs.

Write a short report comparing the performance of SGD with (i) Adam and (ii) a constant

step size. To do this you’ll need to make a number of important choices:

写一份简短的报告，比较 SGD 与 (i) Adam 和 (ii) 恒定步长的性能。 为此，您需要做出一些重要的选择：

How to measure performance. You might, for example, plot the ML loss function vs

optimisation iterations and use the lowest value as a performance measure. But this

measures the ML model performance on the training data and so not the generalisation

performance on unseen data. You might also measure the ML loss function of the

trained model on held-out test data, which would measure generalisation performance.

Probably its a good idea to look at both measures.

如何衡量绩效。 例如，您可以绘制 ML 损失函数与优化迭代的关系图，并使用最小值作为性能度量。 但这衡量的是 ML 模型在训练数据上的性能，而不是在未见数据上的泛化性能。 您还可以在保留的测试数据上测量训练模型的 ML 损失函数，这将测量泛化性能。 查看这两种措施可能是一个好主意。

Remember that SGD involves randomisation and so it may be necessary to collect data

from several runs to understand how performance fluctuates from run to run.

请记住，SGD 涉及随机化，因此可能需要从多次运行中收集数据，以了解性能如何随运行而波动。

What hyperparameters (SGD mini-batch size, Adam α, β1, β2 values, constant step size α)

to use and how to choose them. Probably its a good idea to look at the performance both

when using default hyperparameter values and also when using optimised values.

使用哪些超参数（SGD 小批量大小、Adam α、β1、β2 值、恒定步长 α）以及如何选择它们。 在使用默认超参数值和使用优化值时查看性能可能是一个好主意。

What ML model and data to use for the evaluation. Its probably worth considering two

models/datasets since the optimisation algorithm performance may change with different

models/data, and be sure include at least one neural net ML model. Note that you

shouldn’t choose models/datasets which are too large since training will take too

long, and so you need to use some judgement here (your choice will partly depend

on what compute power you have available). Its a good idea to use standard datasets

e.g. MNIST, CIFAR, Imdb.

用于评估的 ML 模型和数据。 它可能值得考虑两个模型/数据集，因为优化算法的性能可能会随着不同的模型/数据而变化，并且确保至少包含一个神经网络 ML 模型。 请注意，您不应该选择太大的模型/数据集，因为训练需要太长时间，因此您需要在这里做出一些判断（您的选择将部分取决于您可用的计算能力）。 使用标准数据集是个好主意，例如 MNIST、CIFAR、Imdb。

Use one of the existing ML libraries e.g keras/tensorflow or pytorch. Don’t

implement Adam etc yourself

使用现有的 ML 库之一，例如 keras/tensorflow 或 pytorch。 不要自己实现 Adam 等

To see some existing examples of performance evaluation you might find it helpful

to read the following two papers:

要查看一些现有的绩效评估示例，您可能会发现阅读以下两篇论文会有所帮助：

Adam: A Method For Stochastic Optimization, https://arxiv.org/pdf/1412.6980.pdf

The Marginal Value of Adaptive Gradient Methods in Machine Learning, https://arxiv.org/pdf/1705.08292.pdf

The first introduces the Adam optimisation algorithm and compares its performance

(in terms of training error) against a range of other algorithms, see Figures 1-3.

The second compares the performance (in terms of test error i.e. generalisation)

of SGD against a range of algorithms including Adam, see Figures 1-2. Some

questions you might reflect on are: do these papers clearly address all the

choices noted above? And if not, might that be important or not?

第一个介绍了 Adam 优化算法，并将其性能（在训练误差方面）与一系列其他算法进行了比较，见图 1-3。

第二个比较了 SGD 与包括 Adam 在内的一系列算法的性能（就测试错误，即泛化而言），见图 1-2。

您可能会思考的一些问题是：这些论文是否清楚地解决了上述所有选择？ 如果不是，那可能重要还是不重要？

100 marks: indicative breakdown (i) experiment design 30 marks, (ii) methodology 30 marks,

(iii) evaluation and critical discussion 30 marks, (iv) report organisation

and presentation 10 marks

100 分：指示性细分 (i) 实验设计 30 分，(ii) 方法 30 分，(iii) 评估和批判性讨论 30 分，(iv)

报告组织和演示 10 分

SGD with adam VS SGD with constant

1. 实验

本次实验是为了比对SGD中adam和constant step的性能。为了找出算法之间的差异，本次实验采用了3个keras中的标准数据集：MNIST、CIFAR、Imdb。本次实验还试验了在不同数据集大小下的性能区别。同样的，还选取了神经网络和逻辑回归两种模型来配合数据集测试adam和constant step的性能。接下来，通过globalsearch算法分别选择adam和constant step的参数和默认的参数做对比，找出最好的一些参数做算法间的对比。通过多次在训练集和测试集的loss值评估优化器的性能。

1. 方法
   1. Dataset 数据集的特点
      1. MNIST 是一个由1-0数字图片组成的数据集，一共10个数字。训练之前将图片每个像素点转化成0-1之间。
      2. CIFAR10 是由不同的bird，deer，dog等10种物体的图片组成的数据集，训练之前需要将图片每个像素点的值转化成0-1之间。
      3. Imdb 是一个二分类的数据集，将上百个标签转化成数字，训练之后预测是正面(1)还是负面(0)评价
   2. model
      1. neural network
         1. MNIST

在为了MNIST数据集的模型中，

输入层是32个filters，kernel为3\*3，滑动步长为2，2激活函数是relu

第二层是64个filters，接下来做flatten操作，

输出层为10，激活函数是softmax，采用L2正则化。

* + - 1. CIFAR10

CIFAR10的CNN模型的输入层是一个Conv2D，拥有16个filters，一个3\*3的kernel，边缘用0填充，激活函数为relu

第二层是也用了16个filters，滑动步长为（2，2），边缘用0填充，激活函数为relu

第三层和第四层都增加到32个filters，其他和第一层和第二层一样

接着模型做dropout和flatten处理

输出层为10，激活函数是softmax，采用L1正则化

* + - 1. Imdb

首先将训练集和测试集转换成数据长度\*maxlen的矩阵

首先是嵌入层，该层是一个整数编码词汇表，通过输入数据训练得到一个嵌入向量，输出给下一层。

第二层以最简单的方式，计算输入序列维度平均数，并且输出

第三层是一个16个filter的全连接层

输出层使sigmoid激活函数输出一维的0-1的数表示概率。

* + 1. logistic regression

由于逻辑回归有良好凸优化性能，不用担心局部最小值问题，适合比较不同的优化器，所以本次实验采用了逻辑回归模型。

这里使用keras实现了逻辑回归，在二分类中使用binary\_crossentropy loss function即可，在多分类任务中使用categorical\_crossentropy.

* 1. 优化器
     1. SGD
     2. Adam
     3. Constant step
  2. 搜索算法

1. 如何评估神经网络，训练集和测试集的loss，其他一个模型训练集和测试集的准确率，多跑几次？
2. 总结。您可能会思考的一些问题是：这些论文是否清楚地解决了上述所有选择？ 如果不是，那可能重要还是不重要？