数据来源和处理

这部分主要介绍了数据来源或者搜集过程，数据量，数据的预处理手段等。

2.1 MNIST数据集

MNIST（Mixed National Institute of Standards and Technology database）数据集来自美国国家标准与技术研究所， National Institute of Standards and Technology (NIST)。训练集 (training set) 由来自 250 个不同人手写的数字构成，其中 50% 是高中学生, 50% 来自人口普查局 (the Census Bureau) 的工作人员，测试集(test set) 也是同样比例的手写数字数据。

MNIST 数据集可在 http://yann.lecun.com/exdb/mnist/获取, 它包含了四个部分：

Training set images: train-images-idx3-ubyte.gz (9.9 MB, 解压后 47 MB，包含 60,000 个样本)

Training set labels: train-labels-idx1-ubyte.gz (29 KB, 解压后 60 KB，包含 60,000 个标签)

Test set images: t10k-images-idx3-ubyte.gz (1.6 MB, 解压后 7.8 MB，包含 10,000 个样本)

Test set labels: t10k-labels-idx1-ubyte.gz (5KB, 解压后 10 KB，包含 10,000 个标签)

MNIST数据集包含70000张手写数字的灰度图片，其中每一张图片包含 28 \* 28 个像素点。数据集被分成两部分：60000 行的训练数据集（mnist.train）和10000行的测试数据集（mnist.test）。其中，60000 行的训练集分拆为 55000 行的训练集和 5000 行的验证集。60000行的训练数据集是一个形状为 [60000, 784] 的张量，第一个维度数字用来索引图片，第二个维度数字用来索引每张图片中的像素点。在此张量里的每一个元素，都表示某张图片里的某个像素的强度值，值介于 0 和 1 之间。在此张量里的每一个元素，都表示某张图片里的某个像素的强度值，值介于 0 和 1 之间。

每一张图片都有对应的标签，标签是介于 0 到 9 的数字，也就是图片对应的数字，称为 "one-hot vectors"。 一个 one-hot 向量除了某一位的数字是 1 以外其余各维度数字都是 0。例如，数字 3 将表示成一个只有在第 3 维度（从 0 开始）数字为 1 的 10 维向量。比如，标签 0 将表示成 ( [1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0 ] )。因此，其标签是一个 [60000, 10] 的数字矩阵。

2.2 MNIST数据集处理

MNIST手写体数字库是大小为28px\*28px的图片数据，所以需要将手写体数字图像转换成向量。用numpy向量来表示图像数据，以便训练和测试算法。首先，将图像数据灰度化处理，每个像素点用一个灰度值表示。在这里，我们就将28\*28的像素展开成一个一维的行向量，这些行向量就是图片数组里的行（每行784个值）。然后将其存入numpy数组中。依次解析所有的图片数据，得到了60000\*784的一个二维矩阵。

2.3 IRIS数据集

Iris数据集是常用的分类实验数据集，首次出现在著名的英国统计学家和生物学家[Ronald Fisher](https://en.wikipedia.org/wiki/Ronald_Fisher) 1936年的论文《The use of multiple measurements in taxonomic problems》中，由Fisher收集整理。Iris也称鸢尾花卉数据集，是一类多重变量分析的数据集。数据集包含150个数据集，分为3类，每类50个数据，每个数据包含4个属性。可通过花萼长度，花萼宽度，花瓣长度，花瓣宽度4个属性预测鸢尾花卉属于（Setosa，Versicolour，Virginica）三个种类中的哪一类。可在<http://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data>。