算法

注意：1.A 和 1.B 两题任选一题完成即可。

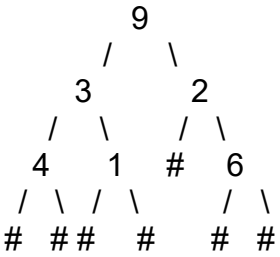
一文本文件 text.txt 内包含多行，每行为一个英文单词。

1.A 用一行 Shell 命令统计（打印）文件中出现频次最高的以字母 t 开头的 10 个单词，和它们出现的次数。

1.B 用你常用的编程语言写一段程序统计（打印）文件中出现频次最高的以字母 t 开头的 10 个单词，和它们出现的次数。

2. 如下例所示，通过前序遍历，我们用一个字符串来表示一棵二叉树（空节点用#表示）。

举例：



该二叉树可以表示为：”9,3,4,#,#,1,#,#,2,#,6,#,#”。

任务：实现一个函数 valid_tree(s)，返回值为 True 或 False，表示 s 是否是一个正确的二叉树表示。其中，参数 s 为一个字符串，其为逗号分隔的整数或者#。

要求：函数不对树进行重构。

你可以假设输入格式总是有效的，例如 s 永远不会包含两个连续的逗号，例如”1,,3”。

注意：3.A 和 3.B 两题任选一题完成即可。

3.A （矩阵降维）给定一大小为(m,n)的矩阵D，用梯度下降法，编程求解大小为(m,k)的矩阵U和大小为(k,n)的矩阵V，使得$U\cdot V$和D尽量接近。其中，$k<m,n$为一个相对较小的参数。设学习率为 lr，梯度下降的轮数为 n_iter，损失函数为$L=\frac{1}{2}\|U\cdot V-D\|_2^2=\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n}(U\cdot V-D)_{ij}^2$。你可以使用向量化编程，例如你可以使用函数 dot(a, b)求解 a 和 b 的内积，trans(a)求解 a 的转置，a+b 或 a-b 求解 a 和 b 的按位求和或求差。

可能有用的公式：

$\|X\|_2^2 = \mathrm{Tr}(X \cdot X^T)$

$\mathrm{Tr}(A) = \sum_i A_{ii}$

$\mathrm{Tr}(A) = \mathrm{Tr}(A^T)$

$\mathrm{Tr}(AB) = \mathrm{Tr}(BA)$

$\mathrm{Tr}(A + B) = \mathrm{Tr}(A) + \mathrm{Tr}(B)$

$\frac{\partial a^T X b}{\partial X} = ab^T$

$\frac{\partial b^T X^T X c}{\partial X} = X(bc^T + cb^T)$

$\frac{\partial}{\partial X}\mathrm{Tr}(AXB) = A^T B^T$

$\frac{\partial}{\partial X}\mathrm{Tr}(AX^T B) = BA$

$\frac{\partial}{\partial X}\mathrm{Tr}(X^T BX) = BX + B^T X$

$\frac{\partial}{\partial X}\mathrm{Tr}(B^T X^T CXB) = C^T XBB^T + CXBB^T$

3.B 给定一个整数数组，实现一个方法找到索引 m 和 n（m < n），使得数组第 m、n 位置之间（包含 m、n 位置）的元素以升序排好序后，整个数组也排好了序，且 n-m 尽可能地小（即找到满足条件的最短子序列）。

举例：

输入： 1, 2, 4, **7, 10, 11, 7, 12, 6, 7**, 16, 18, 19

输出： 3, 9

注意：

(1) 如果输入数组已经排好序，则返回 0, 0。

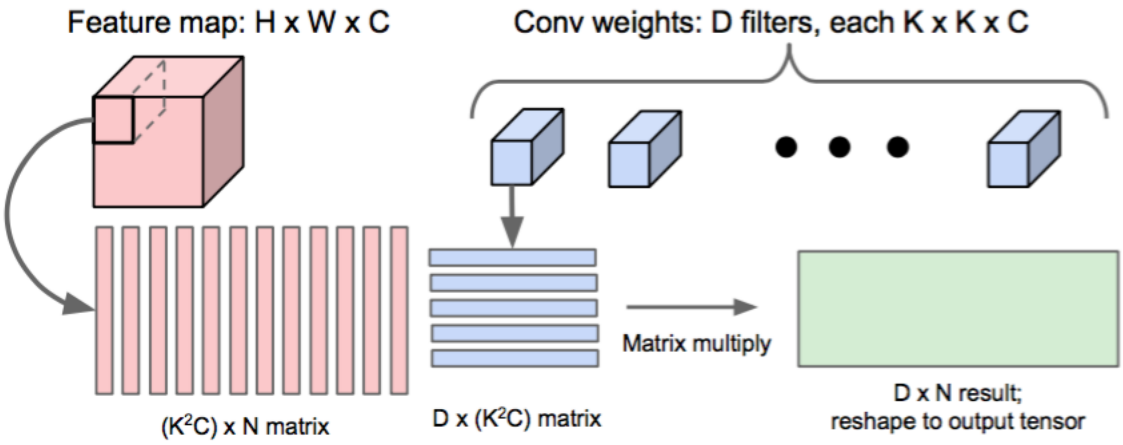
(2) 要求实现的算法复杂度为 O(n)。

机器学习

1. 现在要在包含n个数据点的数据集$\{x_i, y_i\}_{i=1,2,\cdots,n}$，上训练逻辑回归模型（Logistic Regression），模型参数为θ，请写出损失函数（Loss function，有无正则项均可），并推导损失函数的梯度。另外，请简述逻辑回归与支持向量机（SVM）的联系和区别。

2. 请借助公式解释 CART（Classification And Regression Tree）在解决分类问题或者回归问题时对树的某一个节点进行分裂的方法，并简述你所知道的树剪枝的方法。

3. 在很多神经网络中，卷积运算都占有很大的比重。Caffe、PyTorch、TensorFlow 等框架在计算卷积的时候都会用到 im2col（image-to-column）函数，它可以将卷积运算转化为两个矩阵相乘。使用 im2col 进行卷积运算的示意图如下：



图中 im2col 用于将 Feature map 变换为一个二维矩阵，H 为 height，W 为 width，C 为 input_channels，K 为 kernel_size。图中维数仅做示意。

1) 请简述在卷积运算时使用 im2col 的好处和损失。
2) 请实现以下 im2col 函数：
   def im2col(data, kernel_size, stride, padding)
   该函数的功能是将输入的一个多通道图像（即 data）变换为一个矩阵，其输出为二维矩阵 data_col。
   其中 data.shape = (height, width, input_channels)，即输入图像的高和宽，以及通道数；kernel_size 为卷积核的尺寸；stride、padding 为相应操作的尺寸（padding 模式为全零填充）。假设卷积核、stride 和 padding 在高度和宽度两个维度上尺寸相等。

Python 和 Linux

注意：以下所有问题任选 5 题完成即可。

（Linux 部分）

1. 如查看一个文本文件 text 的内容？如何查看该文件的最后 20 行？
2. 如何统计一个文本文件 text 中包含单词 dog 的行数？如何统计该文件中单词 dog 出现在行首的行数？
3. 运行一个程序 python test.py，如何将其标准输出存入一个文本文件 log 中？如不希望覆盖 log 中已有的内容，如何将输出追加至 log 的末尾？
4. 如何删除当前目录下的所有以 tmp 开头，以.csv 结尾的文件？删除目录和删除文件两者的命令主要有何差异？
5. 运行一个程序 python test.py，如何估计其运行时占用的内存大小？
6. 如何让一个脚本 test.sh 在后台运行（从而可以在终端中进行其他工作）？其在后台运行后，如何杀死对应的进程？

（Python 部分）

7. 用一行 Python 代码反转一个字符串（空格分隔的英文单词序列）中的所有单词。举例：
   输入：I love China
   输出：I evol anihC
8. Python 的哪些内置类型是不可变类型？可变类型和不可变类型的区别是什么？
9. 请阐述 python2 和 python3 中 range(100)的区别，并解释哪种更高效。
10. Python 的数组和列表主要区别是？请举例介绍如何用一行代码生成一个长为 10 的列表，其包含元素为长为 10 的列表。
11. 请简述使用 with 关键字打开文件时，with 发挥的主要作用。
12. 什么场景下需要使用* args，** kwargs？它们的主要区别是？