第一章

第一周

第二周

python两个矩阵a和b，a\*b不是矩阵乘法，二是对应位置相乘（element\_wise multiplication）。如果a\*b中有一个维度相同，则会因为广播的重载，而能够进行对应位置相乘

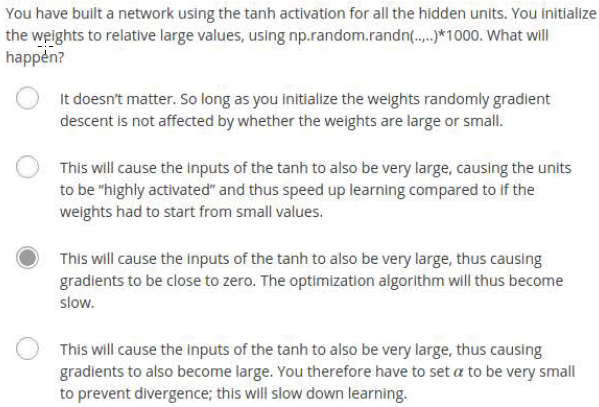
吴恩达课程中，输入x为列向量，输出y为行向量。这点与TensorFlow相反

第三周

激活函数tanh比sigmoid更好，因为其平均输出更接近0，使得输出更为集中，有利于后续层处理

四种激活函数各自优劣和选择理由。各种激活函数的导数是啥？

权重和偏移不能全部初始位0。原因？ Break symmetry是啥



第四周

算法实现中，需要缓存正向传播结果（各层输出A），因为反向传播需要使用

神经网络中，深层计算的输入特征通常比浅层复杂？？？

第二章

第一周

样本划分为训练集，交叉验证集和测试集（train，dev（develop）和test）。样本1w以上时，比例通常为98%，1%和1%

解决偏差和方差的方法？？？

正则方法（正向传播添加额外正则项）通过迭代逐步减小权重数值，从而降低过拟合。此过程称为weight decay

梯度消失和梯度爆炸怎么解决？？？

Data augmentation？？？数据增强，难以直接增加样本量时，对既有样本进行镜像，反转，扭曲等操作，以增大样本数量

将输入归一化可以加快代价函数的收敛？？？输入在各个维度的范围接近，可以使用更大的学习率。圆形碗和扁圆碗帮助理解

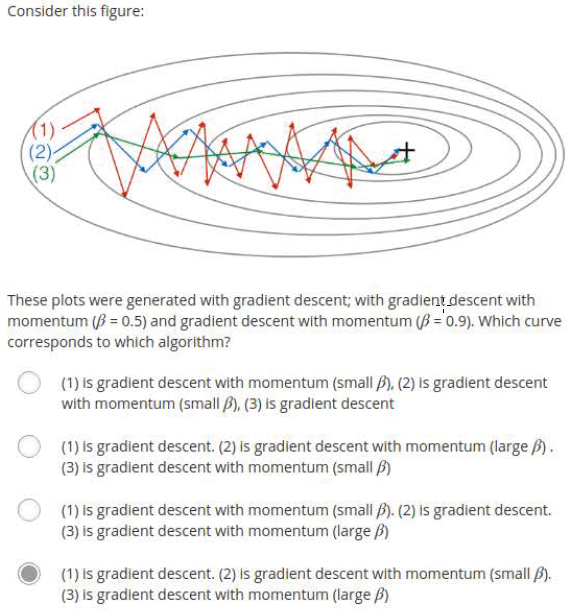
第二周

Mini batch可以加快训练速度，是因为计算机硬件特性的限制。一次计算的样本数量过大，超过了可用内存大小，就需要多次读写内存以完成数据处理。读写内存速度远低于CPU计算速度，导致CPU饥渴。Mini batch相当于引进了流水线，处理数据和计算数据串行同步运行。单纯从梯度收敛有效性来说，batch > mini batch > random 1 sample。一次样本越多，越能反映全局最优梯度，越能减少曲折

移动平均算法的修正参数是啥？？？

学习率衰减有哪些常见公式，总体符合什么原则？？？随着epoch增加，学习率逐渐减小的都可以

移动平均算法中，增加beta将使得拟合曲线更为滞后，因为及时数据的作用比重降低。同时，也将使得拟合曲线更加平滑。



三种优化算法条分缕析！！！

第三周