机器学习

实验指导书

2020.10

山东大学

目录

实验 1

上机练习 2.5 节第 4 题

实验 2 最大似然估计

1、实验目的

(1) 掌握用最大似然估计进行参数估计的原理；
(2) 当训练样本服从多元正态分布时，计算不同高斯情况下的均值和方差。

2、实验数据

样本	类 1			类 2		
	x_1	x_2	x_3	x_1	x_2	x_3
1	0.011	1.03	-0.21	1.36	2.17	0.14
2	1.27	1.28	0.08	1.41	1.45	-0.38
3	0.13	3.12	0.16	1.22	0.99	0.69
4	-0.21	1.23	-0.11	2.46	2.19	1.31
5	-2.18	1.39	-0.19	0.68	0.79	0.87
6	0.34	1.96	-0.16	2.51	3.22	1.35
7	-1.38	0.94	0.45	0.60	2.44	0.92
8	-1.02	0.82	0.17	0.64	0.13	0.97
9	-1.44	2.31	0.14	0.85	0.58	0.99
10	0.26	1.94	0.08	0.66	0.51	0.88

3、实验内容及说明

使用上面给出的三维数据：

(1) 编写程序，对类 1 和类 2 中的 3 个特征x_i分别求解最大似然估计的均值$\hat{\mu}$和方差$\hat{\sigma}^2$。
(2) 编写程序，处理二维数据的情形$p(x)\sim N(\mu,\Sigma)$。对类 1 和类 2 中任意两个特征的组合分别求解最大似然估计的均值$\hat{\mu}$和方差$\hat{\Sigma}$（每个类有3种可能）。
(3) 编写程序，处理三维数据的情形$p(x)\sim N(\mu,\Sigma)$。对类 1 和类 2 中三个特征求解最大似然估计的均值$\hat{\mu}$和方差$\hat{\Sigma}$。
(4) 假设三维高斯模型是可分离的，即$\Sigma=diag\left(\sigma_1^2,\sigma_2^2,\sigma_3^2\right)$，编写程序估计类 1 和类 2 中的均值和协方差矩阵中的参数。
(5) 比较前 4 种方法计算出来的每一个特征的均值μ_i的异同，并加以解释。
(6) 比较前 4 种方法计算出来的每一个特征的方差σ_i的异同，并加以解释。

实验 3 非参数估计

1、实验目的

(1) 掌握用非参数的方法估计概率密度；
(2) 了解parzen 窗方法的原理；
(3) 了解k 近邻方法的原理

2、实验数据

样本	类 1			类 2			类 3		
	x_1	x_2	x_3	x_1	x_2	x_3	x_1	x_2	x_3
1	0.67	0.173	0.85	-0.15	0.84	0.359	1.36	1.86	0.256
2	0.05	-3.04	-3.14	-0.06	0.53	0.23	1.41	1.86	0.75
3	1.55	-0.06	1.96	0.63	0.315	0.235	1.22	-0.15	0.59
4	0.64	0.96	0.5	0.1	0.79	0.281	2.46	-0.19	1.67
5	-1.35	5.56	0.11	-0.1	0.73	0.304	0.68	0.61	3.37
6	0.221	1.14	-4.44	0.42	0.95	0.37	2.51	-0.22	0.38
7	0.02	2.16	2.46	0.239	0.81	0.09	0.6	0.181	0.41
8	0.52	-0.04	-0.6	-0.02	0.87	0.39	0.64	0.04	2.47
9	-1.65	1.02	-1.83	0.185	0.75	0.271	0.85	1.46	-0.19
10	1.12	-0.75	-2.33	0.13	0.314	0.207	0.66	0.15	-0.22

3、实验内容及说明

(1) 问题一：

使用上面表格中的数据进行 Parzen 窗估计和设计分类器。窗函数为一个球形的高斯函数如下所示：

$$\varphi\left(\frac{(x-x_i)}{h}\right) \propto exp[-(x-x_i)^t(x-x_i)/(2h^2)]$$

编写程序，使用 Parzen 窗估计方法对任意一个的测试样本点x进行分类。对分类器的训练则使用表格中的三维数据。令h = 1，分类样本点为$(0.3,1.5,0.4)^t$，$(0.21,0.42,0.18)^t$，$(0.2,0.56,-0.1)^t$ 。

(2) 问题二：

对上面表格中的数据使用k 近邻方法进行概率密度估计：

1) 编写程序，对于一维的情况，当有 n 个数据样本点时，进行k-近邻概率密度估计。对表格中的类 3 的特征x_1，用程序画出当 k=1，3，5 时的概率密度估计结果。
2) 编写程序，对于二维的情况，当有 n 个数据样本点时，进行k-近邻概率密度估计。对表格中的类 2的特征$(x1, x2)^t$，用程序画出当 k=1，3，5 时的概率密度估计结果。
3) 编写程序，对表格中的 3个类别的三维特征，使用k-近邻概率密度估计方法。并且对下列点处的概率密度进行估计： $(0.04,0.62,3.2)^t$，$(-0.7,0.61,-0.28)^t$，$(0.3,1.61,-0.25)^t$。

实验 4 神经网络学习

1、实验目的

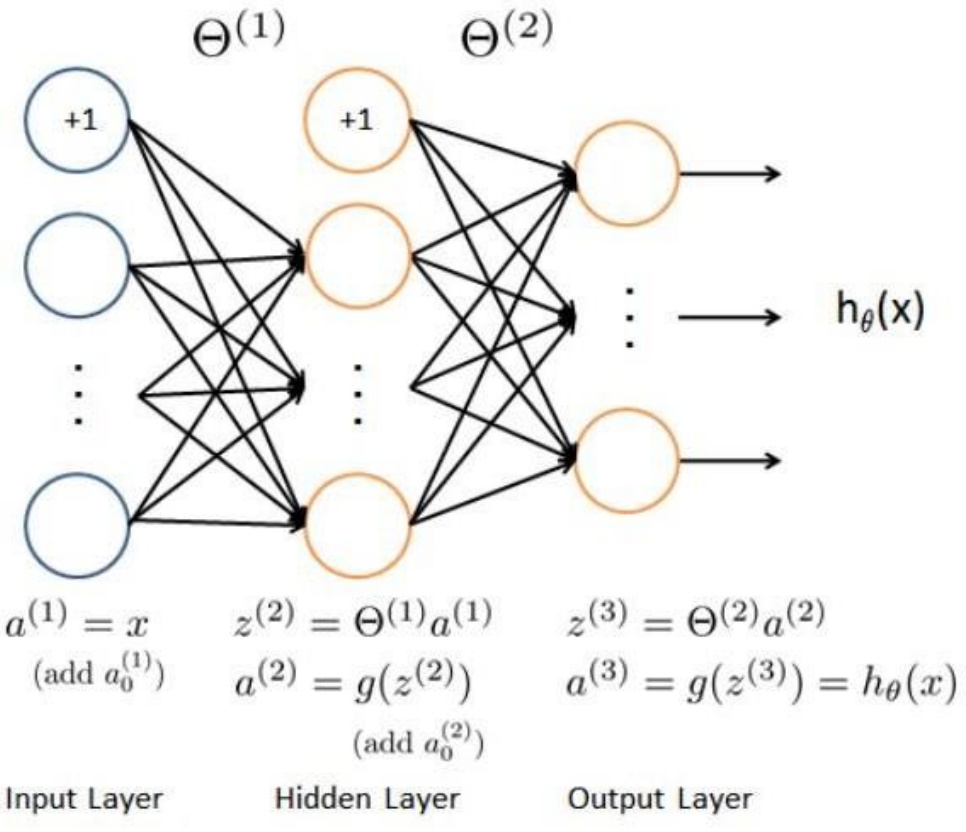
(1) 掌握 BP 神经网络的基本原理和基本的设计步骤；
(2) 了解 BP 算法中各参数的作用和意义。

2、实验数据

CIFAR-10数据集，数据集中包含 50000 张训练样本，10000 张测试样本，可将训练样本划分为49000张样本的训练集和1000 张样本的验证集，测试集可只取1000 张测试样本。其中每个样本都是 32×32 像素的彩色照片，每个像素点包括RGB三个数值，数值范围0～255，所有照片分属10个不同的类别。

3、实验内容及说明

(1) 用神经网络对给定的数据集进行分类，画出loss图，给出在测试集上的精确度；
(2) 不能使用 TensorFlow 等框架，也不能使用库函数，所有算法都要自己实现；
(3) 神经网络结构图如下图所示：



整个神经网络包括 3 层——输入层，隐藏层，输出层。输入层有 32*32*3 个神经元，隐藏层有 50个神经元，输出层有 10 个神经元（对应 10 个类别）。

(4) 附加：可以试着修改隐藏层神经元数，层数，学习率，正则化权重等参数探究参数对实验结果的影响。

实验 5 集成学习

1、 实验目的

用集成方法对数据集进行分类

2、 实验数据

实验 4 中的 CIFAR-10 数据集

3、 实验内容及说明

(1) 利用若干算法，针对同一样本数据训练模型，使用投票机制，少数服从多数，用多数算法给出的结果当作最终的决策依据，对CIFAR-10数据集进行分类，给出在测试集上的精确度；

(2) 所选算法包括：

SVM（核函数为多项式核函数）；

KNN（k=7）；

神经网络。

注：实验 4 中的神经网络模型可以使用，也可以使用框架，SVM和KNN需要自行实现，不可使用框架和库函数。