Project 文档

刘云飞

Contents

总体分布的非参数估计

已知样本所属的类别，但未知总体概率密度函数的形式，要求我们直接推断概率密度函数本身。统计学中常见的一些典型分布形式不总是能够拟合实际中的分布。此外，在许多实际问题中经常遇到多峰分布的情况，这就迫使我们必须用样本来推断总体分布，常见的总体类条件概率密度估计方法有 Parzen 窗法和 K 近邻法两种。

非参数估计也有人将其称之为无参密度估计，它是一种对先验知识要求最少，完全依靠训练数据进行估计，而且可以用于任意形状密度估计的方法。常见的非参数估计方法有以下几种：

1. 直方图：把数据的值域分为若干相等的区间，数据按照区间分 为若干组，每组形成一个矩形，矩形的高和该组数据的多少成正比， 其底为所属区间，将这些矩形依次排列组成的图形就是直方图。它提 供给数据一个直观的形象，但只适合低维数据的情况，当维数较高时， 直方图所需的空间将随着维数的增加呈指数级增加.

2. 核密度估计：原理和直方图有些类似，是一神平滑的无参密度估计方法。对于一组采样数据，把数据的值域分为若干相等的区间， 每个区间称为一个 bin，数据就按区间分为若干组，每组数据的个数和总参数个数的比率就是每个 bin 的概率值。相对于直方图法，它多了一个用于平滑数据的核函数。和密度估计方法适用于中小规模的数据集，可以很快地产生一个渐进无偏的密度估计，有良好的概率统计性质。

3. K 近邻估计：密度估计的加权是以数据点到 x 的欧式距离为基准来进行的，而 K 近邻估计是无论欧式距离多少，只要是离 x 点的最近的 k 个点的其中之一就可以加权.

Parzen 窗

主要通过概率密度函数的方式来定义. 设概率密度函数 p(x)满足下面条件:

1. x 的概率值是落入范围[a, b]之间的:

$$P(a < x < b) = \int_a^b p(x)dx$$

2. x 是非负实数

3. 概率之和为 1

$$\int_{-\infty}^{\infty} p(x)dx = 1$$

实际生活中大多数分布服从正态分布:

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-c)^2}{2\sigma^2}\right)$$

在这里 c 为样本均值, σ 为样本方差

此处先引入**概率密度估计**，给一个 n 个样本的数据$x_1, \ldots, x_n$，我们需要根据数据来估计密度函数p(x)，通过此函数来预测所有新来的样本 x，最基础的思路是概率密度为 P 为落入区域 R 中的个数：

$$\mathrm{P} = \int_R p(x)dx$$

如果 R 足够小并且p(x)在区域 R 中视为变化比较小，可以得到下面的式子：

$$\mathrm{P} = \int_R p(x)dx \approx p(x)\int_R dx = p(x)V$$

这里用 V 代表区域 R 的容积。

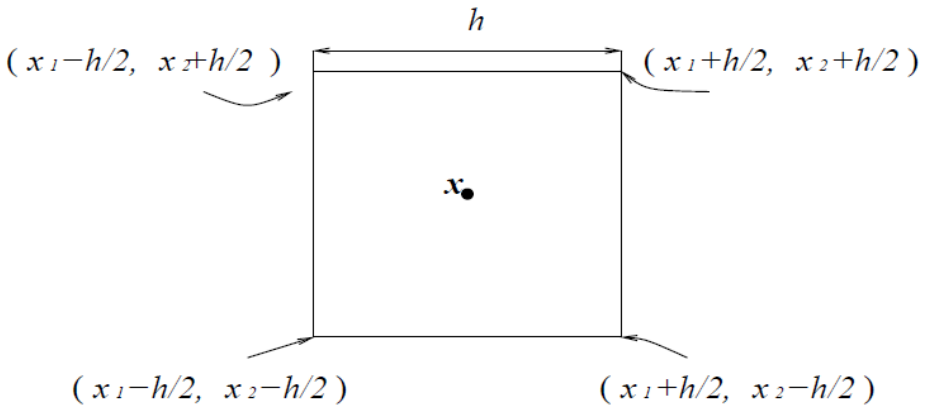
引入样本独立性假设，并且 n 个数据有 k 个落入区域 R 中，因此我们有

$$\mathrm{P} = \mathrm{k/n}$$

进而可以求出概率密度p(x)为：

$$\mathrm{p(x)} = \frac{k/n}{V}$$

如果 R 是一个高维的空间块，此处引用一个二维的方块来辅助理解



把这一区域称为一个窗口，引入下面的定义：

$$\varphi\left(\frac{x_i - x}{h}\right) = \begin{cases} 1 & \frac{|x_{ik} - x_k|}{h} < \frac{1}{2}, \quad k = 1,2 \\ 0 & otherwise \end{cases}$$

这一函数可以用来判定x_i是否落入上述方块（中心为 x，宽度为 h）中。因此可以用 k 来表示所有 n 个样本落入区域 R 中的个数：

$$\mathrm{k} = \sum_{i=1}^{n} \varphi\left(\frac{x_i - x}{h}\right)$$

结合上述式子可以得到二维的 Parzen 概率密度函数为：

$$\begin{aligned} p(x) &= \frac{k/n}{V} \\ &= \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h^2} \varphi\left(\frac{x_i - x}{h}\right) \end{aligned}$$

这里$\varphi\left(\frac{x_i - x}{h}\right)$被称为窗函数，也因此我们可以自定义不同的窗函数来应对不同的使用情景。因此我们用 Parzen 窗来做分类的时候，来一个新的样本数据x_i，我们可以通过设定窗口大小(h 的值)来判定每一个已有标记好的数据中每一个类别有多少个样本落入这一窗口中，进而可以求出该样本属于每一类别的概率。

相关算法实现与分析

这次通过 Parzen 窗方法来求手写体图片的一个后验概率进而来预测图片属于哪一个数字进而实现图片分类的目的。

数据集使用 MNIST，一个 28×28 的灰度图像，这里将不同数字对应的图片像素$28 \times 28 = 784$维度空间中的一个点，每一个轴对应的是其对应(x, y)位置的像素值。考虑不同数字对应的这么一个空间中的点的分布是呈现不同的聚类关系，基于这样一个模型来进行 Parzen 窗分类算法的实现。

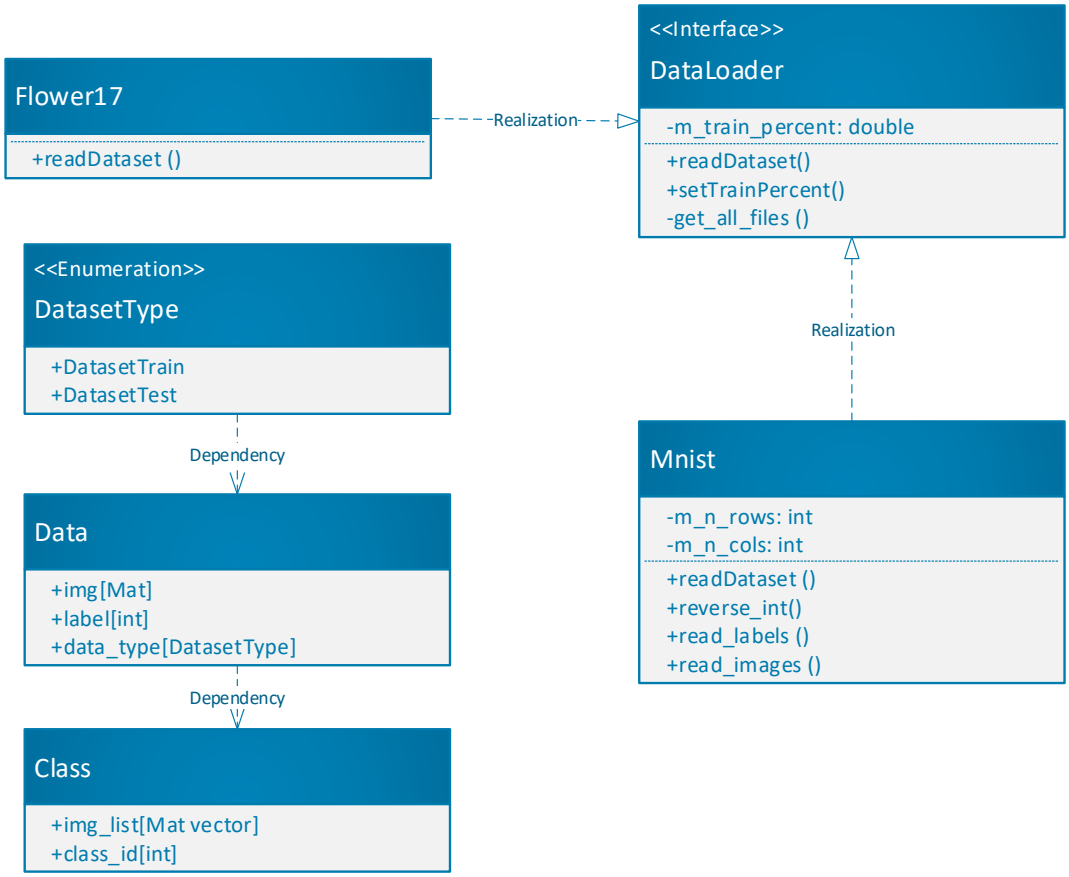
[相关工程和代码]

环境要求

1. Ubuntu 16.04
2. Cmake 2.8+
3. OpenCV 3.2

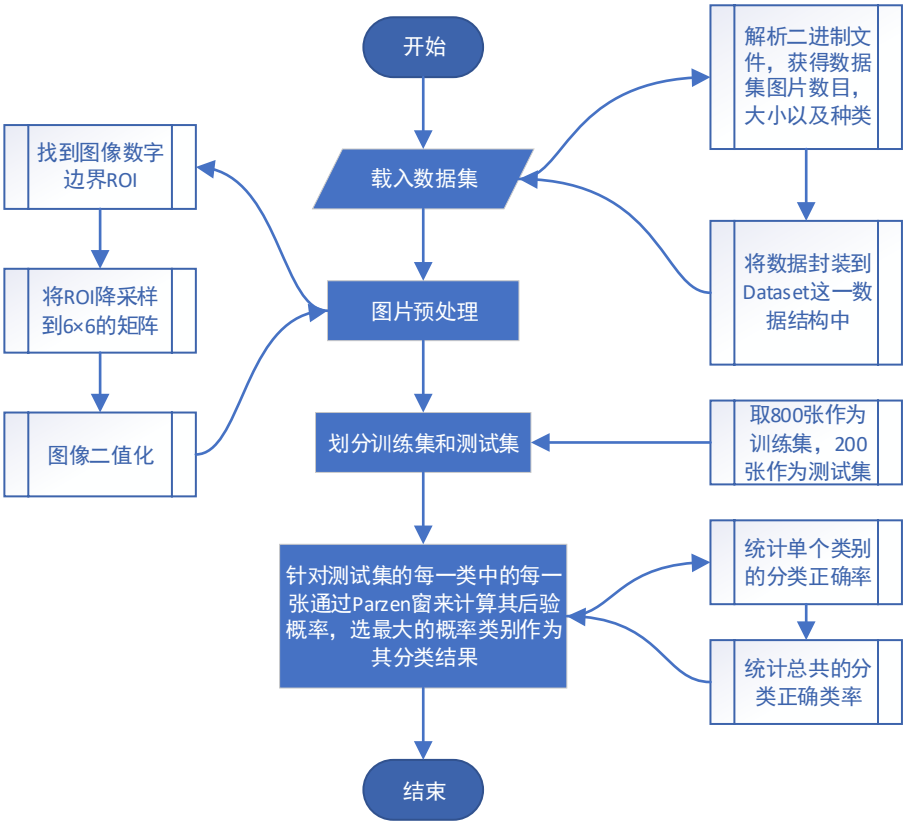
数据结构

数据结构如下图所示：



算法流程

Parzen 窗算法的流程如下：

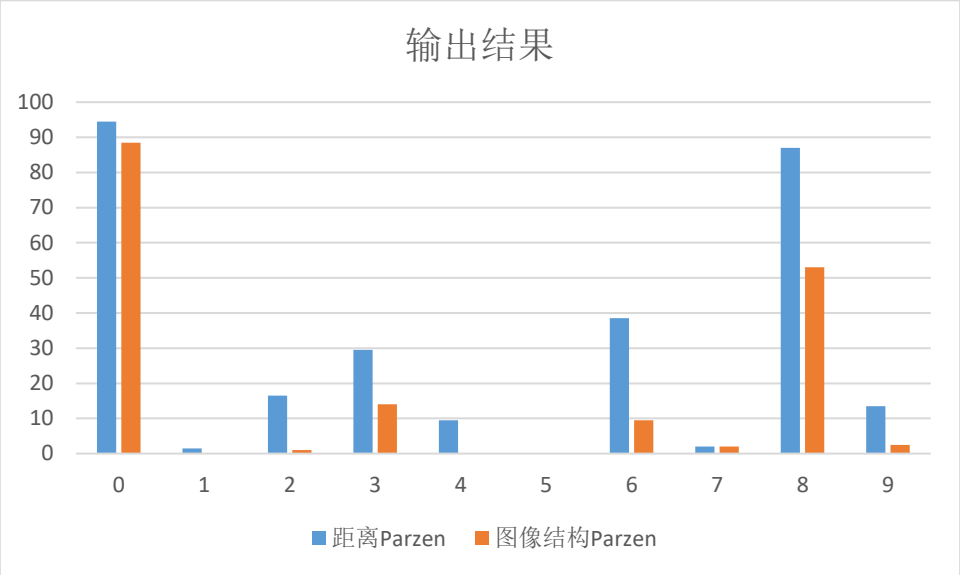


实验结果

实验相关过程图示如下：

分类结果的可视化，绿色边框表示预测正确，红色表示错误

实验分类预测准确率结果如下：



注：距离 parzen 是根据二值图像之间像素的欧式距离来设定的窗口；图像结构是根据图像像素周围设置一个二维的窗口来搜索该窗口里是否有相同的像素来判断两张图片的距离。

结果很明显，使用距离的方法要优于使用图像结构的方法，原因为图像结构方法设计的不够优导致距离估计错误率过大。

Parzen 窗口大小选取讨论

实验结果是经过调整 parzen 窗口大小，得到稍微合适的结果的。这次没有花费大量功夫在调整 parzen 窗口，着重在 parzen 的物理模型的理解以及相关实现。这里面分两个极端进行讨论：

1. 窗口过小。由于图片分类本身就是一个高维度数据的信息，窗口过小直接导致几乎没有其他训练集中的高维点落入这一窗口，从而不能根据落入点的信息来判断对应样本属于哪一类（没有参考）；
2. 窗口过大。窗口过大会导致其他距离样本点远的训练集点落入窗口内从而“误导”对样本的分类。