高斯核推导

以下内容来自DeepSeek：

局部加权线性回归（Locally Weighted Linear Regression, LWLR）中高斯核的推导源于一个核心思想：**给训练集中靠近查询点 x 的点赋予更高的权重，远离的点赋予更低的权重。**高斯函数（或正态分布密度函数的形式）因其良好的数学性质（光滑、对称、值域在 [0,1] 之间、随距离单调递减）成为实现这种权重分配的常用选择。

以下是高斯核在 LWLR 中的推导和解释：

1. LWLR 的核心：加权最小二乘法

LWLR 的目标是为每一个查询点 x 拟合一个局部线性模型。它的损失函数是**加权平方误差**：

$$J(\theta) = \sum_{i=1}^{m} w^{(i)} (y^{(i)} - \theta^T x^{(i)})^2$$

其中：

- m 是训练样本数。
- $x^{(i)}, y^{(i)}$ 是第 i 个训练样本。
- θ 是要优化的局部线性模型的参数。
- $w^{(i)}$ 是赋予第 i 个训练样本的**权重**。这个权重 w^{(i)} 是 LWLR 的关键。

2. 权重 $w^{(i)}$ 的设计目标

我们期望权重 $w^{(i)}$ 满足：

- **依赖于距离**：权重应取决于查询点 x 与训练点 $x^{(i)}$ 之间的距离 $d^{(i)} = ||x - x^{(i)}||$ （通常用欧氏距离或其平方）。
- **距离近则权重大**：当 $x^{(i)}$ 非常接近 x 时， $w^{(i)}$ 应接近 1（最大权重）。
- **距离远则权重小**：当 $x^{(i)}$ 远离 x 时， $w^{(i)}$ 应接近 0（最小权重）。
- **光滑递减**：权重应随距离 $d^{(i)}$ 的增加而光滑、单调地减小到 0。

3. 高斯函数作为权重函数

高斯函数（未归一化的形式）完美地符合上述要求：

$$k(d) = e^{-\frac{d^2}{2}}$$

其中 d 是距离。这个函数：

- 当 $d=0$ 时，$k(0)=e^0=1$
- 当 $d\to\infty$ 时，$k(d)\to 0$
- 随 d 增大而光滑、单调递减。

4. 引入带宽参数 τ

直接使用 $k(d)=e^{-\frac{d^2}{2}}$ 有一个问题：权重衰减的速度是固定的。我们需要一个参数来控制“局部”的范围，即决定多远的点被认为是“近邻”。为此引入带宽参数 (Bandwidth Parameter) $\tau(tau)$，也称为平滑参数。

修改权重函数为：

$$k(d;\tau)=e^{-\frac{d^2}{2\tau^2}}$$

- τ **的作用**：
  - 当 τ 很大时，$\frac{d^2}{2\tau^2}$ 整体变小，指数部分衰减变慢，权重下降缓慢。这意味着更远的点也能获得相对显著的权重，模型考虑的范围更广，拟合结果更平滑（可能欠拟合）。
  - 当 τ 很小时，$\frac{d^2}{2\tau^2}$ 整体变大，指数部分衰减变快，权重迅速下降。这意味着只有非常靠近查询点 x 的点才有显著影响，模型更关注局部细节，拟合结果波动更大（可能过拟合）。
- **物理意义**：τ 定义了“邻域”的宽度尺度。可以把它想象成高斯分布的“标准差”，它控制着钟形曲线的宽窄。

5. 关键点

- **非参数**：LWLR 是一种非参数方法，模型形式（线性）虽然简单，但预测结果 $\hat{y}$ 是查询点 x 的复杂函数（因为权重 $w^{(i)}$ 依赖于 x）。
- **计算开销**：对每个查询点 x 都需要重新计算所有权重 $w^{(i)}$ 并求解一次加权最小二乘问题，计算量较大（尤其是大数据集）。
- τ **的选择至关重要**：它直接控制模型的偏差-方差权衡。通常通过交叉验证来选择最优的 τ。
- **核函数选择**：虽然高斯核最常用，但其他满足距离近权重大、距离远权重小的函数（如 Epanechnikov 核、三角核等）也可用作权重函数。高斯核因其无限可微性（非常光滑）而广受欢迎。

总而言之，高斯核在局部加权线性回归中的“推导”本质上是基于“距离近权重大，距离远权重小”的直观要求，选择符合该要求的高斯函数形式 $e^{-d^2/2}$，并引入带宽参数 τ 来控制权重的衰减速度（即邻域大小），最终得到权重公式 $w^{(i)}=\exp\left(-\frac{||x-x^{(i)}||^2}{2\tau^2}\right)$。