非线性变换习题解答

1，分类时进行非线性变换的目的是什么？它可能会导致什么现象发生？一般会采取什么措施减低其影响？

解：

线性可分的场景大多是理想情况，我们面临的常常是线性不可分的场景，非线性变换的目的是将样本从线性不可分的特征空间映射到另一维度的特征空间中，使其在该维度空间的样本呈现出线性可分的特性。大部分非线性变换的结果是将低维特征空间映射到更高维的特征空间中，这将增加模型的复杂度，导致模型在训练样本集上可以获得优秀的性能，但在测试样本集上性能却急剧下降，即发生过拟合现象，尤其是训练样本集越小、模型越复杂时，过拟合越容易发生。为此，常采取的措施包括：（1）增加训练样本集的样本数（含数据增强、数据清洗等）；（2）通过正则化的手段，控制模型复杂度；（3）模型验证（交叉验证）等手段。

2，证明在正态分布假设下，最大后验估计等价于岭回归（即线性回归的正则化操作），并说明在怎样的情况下最大后验估计将退化为最大似然估计。

解：

岭回归的表达式为：

$$\hat{w}_{ridge} = \underset{\boldsymbol{w}}{\text{argmin}} \left(\frac{1}{N}\sum_{n=1}^{N}(y_n - \boldsymbol{w}^T\mathbf{x}_n)^2 + \frac{\lambda}{N}\boldsymbol{w}^T\boldsymbol{w}\right)$$

$$= \underset{\boldsymbol{w}}{\text{argmin}} \left(\frac{1}{N}\sum_{n=1}^{N}(y_n - \boldsymbol{w}^T\mathbf{x}_n)^2 + \frac{\lambda}{N}\sum_{i=1}^{d} w_i^2\right)$$

上式的第一项为线性回归的平方误差项，只有第一项时就是利用最小二乘法得到最佳解。第二项为正则化项，目的是防止过拟合发生，两项合在一起寻找最佳解时称为岭回归算法。

假设残差$\varepsilon = y - \boldsymbol{w}^T\mathbf{x}$，为正态分布的随机变量：$\varepsilon \sim \mathcal{N}(0, \sigma^2)$

则概率密度函数为：$p(y|\mathbf{x}, \boldsymbol{\theta}) = \mathcal{N}(y|\mu(\mathbf{x}), \sigma^2) = \mathcal{N}(y|\boldsymbol{w}^T\boldsymbol{x}, \sigma^2)$

表示给定输入$\mathbf{x}$和模型中的各项参数，y是正态分布，其均值是对$\mathbf{x}$的线性预测。

当假定$\boldsymbol{w}$是未知常数时，我们是用点估计方式，也就是对上式的第一项通过最小二乘法（已证明过等价于最大化似然$\mathcal{L}(\mathbf{w})$）得到最佳解，但有可能导致过拟合现象发生。

当假定$\boldsymbol{w}$是随机变量，且该d维向量中的各个分量w_i相互独立，且满足正态分布$\mathcal{N}(0, \tau^2)$时，则：$\boldsymbol{w}$的先验概率为：

$$P(\boldsymbol{w}) = \prod_{i=1}^{d} \mathcal{N}(0, \tau^2)$$

最大后验概率为：

$$\hat{\boldsymbol{w}}_{MAP} = \underset{\boldsymbol{w}}{\text{argmax}} \left(logP(\boldsymbol{w}) + \sum_{n=1}^{N} log\, p(\mathbf{x}_n|y_n, \boldsymbol{w})\right)$$

$$= \underset{\boldsymbol{w}}{\text{argmax}} \left(logP(\boldsymbol{w}) + \sum_{n=1}^{N} log\, p(y_n|\mathbf{x}_n, \boldsymbol{w})\right)$$

$$= \underset{\boldsymbol{w}}{\text{argmax}} \left(log \prod_{i=1}^{d} \mathcal{N}(0, \tau^2) + \sum_{n=1}^{N} log\, \mathcal{N}(y|\boldsymbol{w}^T\mathbf{x}, \sigma^2)\right)$$

$$= \underset{\boldsymbol{w}}{\text{argmax}}\,(\,log\prod_{i=1}^{d}\frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{1}{2\tau^2}w_i^2)$$

$$+\sum_{n=1}^{N}log\frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{1}{2\sigma^2}(y_n-\boldsymbol{w}^T\mathbf{x}_n)^2)$$

$$= \underset{\boldsymbol{w}}{\text{argmax}}\,(-\frac{1}{2\tau^2}\sum_{i=1}^{d}w_i^2-\frac{d}{2}\log(2\pi\tau^2)$$

$$-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(y_n-\boldsymbol{w}^T\mathbf{x}_n)^2-\frac{N}{2}log\,(2\pi\sigma^2))$$

$$= \underset{\boldsymbol{w}}{\text{argmin}}\,(\frac{1}{2\tau^2}\sum_{i=1}^{d}w_i^2+\frac{d}{2}\log(2\pi\tau^2)$$

$$+\frac{1}{2\sigma^2}\sum_{n=1}^{N}(y_n-\boldsymbol{w}^T\mathbf{x}_n)^2+\frac{N}{2}log\,(2\pi\sigma^2))$$

上式中的第二项和第四项为常数，因此，对比岭回归的参数估计表达式与上述参数的最大后验估计表达式一致，即两者等价。

如果$\boldsymbol{w}$是随机变量，且该d维向量中的各个分量w_i相互独立，但其满足的是均匀分布时，则$\boldsymbol{w}$的先验概率为：$P(\boldsymbol{w})=\frac{1}{d}$，则最大后验估计表达式的第一项也为常数，因此，最大后验估计就退化为最大似然估计。