PRML 学习笔记

1 绪论

Beth

1.1 多项式曲线拟合

目的：通过对新数据的预测实现良好的泛化性

用一个多项式函数来拟合数据：

$$y(x,\omega) = \omega_0 + \omega_1 x + \omega_2 x^2 + \cdots + \omega_M x^M = \sum_{j=0}^{M} \omega_J x^J$$

阶数(order)：多项式中的 M

误差函数(error function)：衡量对于任意给定的ω值，函数与训练集数据的差别。

$$E(\omega) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n,\omega) - t_n\}^2$$

模型对比(model comparison) **或者模型选择**（model selection）：通过数据拟合结果选择多项式的阶数M。

过拟合(over-fitting)：对于多项式函数，精确地通过每一个训练数据点，但对于总体而言很差。

根均方(RMS)**误差**：除以N让我们能够以相同的基础对比不同大小的数据集，平方根确保了E_{RMS}与目标 变量t使用相同的规模和单位进行度量。

$$E_{RMS} = \sqrt{2E(\omega^*)/N}$$

增大数据集的规模会减小过拟合问题

正则化(regularization)：给误差函数增加一个惩罚项，使得系数不会达到很大的值。这种惩罚项最简单的形式采用所有系数的平方和的形式。

$$E(\omega) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n,\omega) - t_n\}^2 + \frac{\lambda}{2}||\omega||^2$$

$$||\omega||^2 = \omega^T\omega^* = {\omega_0}^2 + {\omega_1}^2 + {\omega_2}^2 + \cdots + {\omega_M}^2$$

1.2 概率论

概率论让我们能够根据所有能得到的信息做出最优的预测，即使信息可能是不完全的或者是含糊的。

联合概率（joint probability）：X取值x_i且Y 取值$\mathrm{y_j}$的概率被记作$\mathrm{p}(X=x_i, Y=\mathrm{y_j})$，被称为$X=x_i$和$Y=\mathrm{y_j}$的联合概率（joint probability）。

$$\mathrm{p}(X=x_i, Y=\mathrm{y_j})=\frac{\mathrm{n}_{ij}}{N}$$

1.3 模型选择

交叉验证(cross validation)：在有限的数据集中，为了尽可能多的可得到的数据进行训练，交叉检验时让可得到数据的$\frac{\mathrm{S}-1}{S}$用于训练，使用所有的数据来评估。S=N时，称为“留一法”（leave-one-out）。

交叉检验的缺点在于：

1.训练次数随着S而增加，使得训练耗时较大。

2.对于单一模型，可能有多个复杂度参数。

1.4 维度灾难

存在许多输入变量组成的高维空间，是影响识别技术设计的重要因素。

把输入空间划分成小的单元格，当给出测试点，我们要预测类别的时候，我们首先判断它属于哪个单元格，然后我们寻找训练集中落在同一个单元格中的训练数据点。测试点的类别就是测试点所在的单元格中数量最多的训练数据点的类别。 但当输入数据变多，对应的为高维的输入空格键，单元格数量以指数的形式增大。就需要指数量级的训练数据。

1.5 决策论

对于一个病人，要判断他是否为癌症，假设在拍X光前，该病人患癌症的概率为$\mathrm{p}(C_1)$(先验概率)，$\mathrm{p}(C_1|x)$为后验概率，我们的目标是最小化把x分到错误类别中的可能性。所以，应当选择有最大后验概率的类别。

1.5.1 最小化错误分类率

决策区域(decision region) ：**我们需要规则来将每个x**的值分到一个合适的类别，以规则将输入空间切分成不同的区域R_k,这种区域被称为决策区域。（决策区域未必是连续的，可以由若干个分离的区域组成）

决策边界(decision boundary)或者决策面(decision surface): 决策区域间的边界。

如果我们把每个x分配到后验概率$\mathrm{p}(C_k|x)$最大的类别中，那么我们分类错误的概率就会最小。

1.5.2 最小化期望损失

假设对于新的x的值，真实的类别为$\mathrm{C_k}$，我们把x分类为$\mathrm{C_j}$(其中j可能与k相等，也可能不相等)。这样做的结果是，我们 会造成某种程度的损失，记作 $\mathrm{L_{kj}}$ ，它可以看成损失矩阵(loss matrix)的第k, j个元素 。最优解是使损失函数最小的解。

对于一个给定的输入向量x，我们对于真实类别的不确定性通过联合概率分布$\mathrm{p}(x,C_k)$表示。因此，我们转而去最小化平均损失。

$$E[L]=\sum_{\mathrm{k}}\sum_{\mathrm{j}}\int_{R_{\mathrm{j}}}L_{kJ}\mathrm{p}(x,C_k)\mathrm{d}x$$

1.5.3 拒绝选项

拒绝选项(reject option)：类别的归属相对不确定，对于这种困难的情况，避免做出决策是更合适的选择。这样会使得模型的分类错误率降低。这被称为拒绝选项(reject option)。

1.5.4 推断和决策

分类问题划分成了两个阶段:推断(inference)阶段和决策(decision)阶段

推断阶段训练模型，决策阶段根据后验概率进行分类。

判别函数(discriminant function)：简单地学习一个函数， 将输入x直接映射为决策

解决决策问题的三种方法:

(a) **生成式模型(generative model)**：显式地或者隐式地对输入以及输出进行建模的方法 。确定类别的条件密度，推断先验概率，求出后验类概率

缺点：输入多，维度高，需要大量训练数据

优点：能够求出数据的边缘概率密度，对于监测模型中具有低频率的新数据点很有用。

(b) **判别式模型(discriminative models)**：直接对后验概率建模

(c) 通过判别函数直接将输入映射为类别标签

计算后验概率的原因：最小化风险，拒绝选项，补偿类先验概率，组合模型

1.5.5 回归问题的损失函数

闵可夫斯基损失函数(Minkowski loss)

1.6 信息论

随机变量x的熵(entropy)：

$$H[L] = -\sum_{x} p(x) log_2 p(x)$$

1.6.1 相对熵和互信息

分布p(x)和分布q(x)之间的相对熵 (relative entropy)

$$\begin{aligned} \mathrm{KL}(p \,\|\, q) &= -\int p(\boldsymbol{x}) \ln q(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} - \left(-\int p(\boldsymbol{x}) \ln p(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} \right) \\ &= -\int p(\boldsymbol{x}) \ln \left\{ \frac{q(\boldsymbol{x})}{p(\boldsymbol{x})} \right\} \mathrm{d}\boldsymbol{x} \end{aligned}$$