模式识别导论第五次作业

翁晨阳

1. 在短文本分析和图像匹配应用中，使用的典型特征提取方法分别是什么?

短文本分析：TF-IDF、Word2Vec

图像匹配应用：SIFT

2. 简述 PCA 和 LDA 的异同

PCA和LDA都是将高维的数据通过线性变换矩阵转换到低维的空间中，具有数据降维的作用。

但是PCA是非监督的，它的目标是最大程度保持原始数据的样本差异信息（数据方差最大的投影方向），降低问题复杂度，减少数据噪声的影响。

LDA是监督的，它的目的是最大化数据的线性可分性（Fisher分类准则最大的投影方向），提高分类性能。

3. 给定均值为d维0向量的样本集合$X=[x_1,x_2,\dots,x_N]\in R^{d\times N}$，其中d是样本特征的个数，N是样本个数，即$\sum_{i=1}^{N}x_i=0,x_i$是第i个样本。假定使用PCA算法对其进行特征变换，需要保留k个主成分。

1）写出特征变换的流程

由题意$\mu=0$

1. 计算样本离散度矩阵：$S=\sum_{i=1}^{N}(x_i-\mu)(x_i-\mu)^T=XX^T$
2. 对离散度矩阵S进行特征值分解，取前k个最大的特征值对应的特征向量组成线性变换矩阵 W，其中 W 的第i列对应第i个特征向量；
3. 对输入特征进行降维：$y_i=W^T(x_i-\mu)=W^Tx_i$

2）如果特征的维度很高（例如1000维），并且d>>N，应该如何处理?

当维度很高时$\Sigma=XX^T$的特征值分解复杂度很高，可以先计算X^TX的特征值和特征向量V，再将特征向量转化为原始协方差矩阵的特征向量$W=XV$

$$X^TXV=\lambda V\Rightarrow XX^TXV=\lambda XV$$

3）写出KPCA对样本集合进行特征变换的流程。

1. 首先计算样本核矩阵K，对其进行特征值分解，从大到小，得到特征向量V
2. 再转化为非线性变化空间中协方差矩阵的特征向量

$$w^l=\sum_{i=1}^{n}v_i^l\phi(x_i)$$

3. 计算$\phi(x)$在主成分上的投影值

$$w^l\phi(x)=\sum_{i=1}^{n}v_i^l\phi(x_i)\cdot\phi(x)=\sum_{i=1}^{n}v_i^l\kappa(x_i,x)$$

4. 处理模式识别高维数据的两种基本方法是什么？它们有什么不同之处?

处理模式识别高维数据的两种基本方法是**特征变换**和**特征选择**，不同之处在于前者是对已有的特征做变换得到**新的特征**，用新的特征参与运算，而后者并没有产生新的特征，而是从给定的特征集合中选取出与任务相关的**特征子集**参与运算。

5. 根据特征选择与分类器的结合程度，特征选择方法可以分为哪三类？各有什么特性?

根据特征选择与分类器的结合程度，特征选择方法可以分为：过滤式、包裹式、嵌入式。

过滤式方法特征选择过程与分类是单独进行的，先进行特征选择，再训练分类器，特征选择评价判据间接反应分类性能（“选择”与“学习”独立）；

包裹式方法特征选择过程与分类性能相结合，特征评价判据为分类器性能（“选择”依赖“学习”）；

嵌入式方法将分类器学习与特征选择融为一体，分类器训练过程自动完成了特征选择（“选择”与“学习”同时进行）。

6. ID3、C4.5 和 CART 是三种典型的决策树分类算法，请分别说明这三种算法具体使用的决 策特征选择方法（注：CART 分别说明分类和回归两种情况）。

ID3：根据信息增益，确定“最优”特征，构建决策树。

C4.5：根据信息增益率，确定“最优”特征，构建决策树。

CART分类树：使用基尼系数进行特征选择。

CART回归树：使用平方误差最小准则进行特征选择。

7. 简述最优特征选择方法的基本思想。

将所有特征选择组合表示成树的形式，然后采用分枝定界方法进行搜索，使得搜索过程快速到达最优解而不用全部遍历。

8. 简述决策树剪枝的目的，以及两种常用的判断决策树是否需要剪枝的准则。

目的：适当剪除一些不必要的分支，减少过拟合，获得更好的推广能力。

预剪枝：在决策树生成过程中，对每个结点在划分前先进行估计，**若对该结点的划分不会带来决策树泛化性能的提升**，则不进行该划分。

后剪枝：先从训练集生成一颗完整的决策树，然后自底向上地对非叶结点进行考察，**若将该结点对应的子树替换为叶结点能带来决策树泛化性能提升**，则进行替换（剪枝）。

9. 决策树的剪枝策略一般有哪些？各有什么优缺点?

预剪枝：在决策树生成的过程进行剪枝，只有泛化性能提升的结点划分才被采纳。**计算速度快，有欠拟合的风险。**

后剪枝：先生成决策树，再自底向上对所有结点进行检验，判断剪枝前后是否会带来泛化性能的提升，以此决定是否剪枝。**欠拟合风险低，泛化能力强，计算量大。**

给出信息增益、信息增益率、Gini指数的计算公式，并对如下数据，

1）计算属性“色泽”的信息增益

$$
\begin{aligned}
\text{Gain}(D,a) &= \text{Ent}(D) - \text{Ent}(D|a) \\
&= -\sum_{k=1}^{c} p_k \log_2 p_k - \sum_{i=1}^{V} \frac{|D^i|}{|D|} \text{Ent}(D^i) \\
&= -\sum_{k=1}^{c} p_k \log_2 p_k - \sum_{i=1}^{V} \frac{|D^i|}{|D|} \times (-\sum_{k=1}^{c_i} p_k \log_2 p_k)
\end{aligned}
$$

$$
\begin{aligned}
\text{Ent}(D) &= -\frac{8}{17}\log_2\frac{8}{17} - \frac{9}{17}\log_2\frac{9}{17} = 0.998 \\
\text{Ent}(D^{\text{青绿}}) &= -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2} = 1 \\
\text{Ent}(D^{\text{乌黑}}) &= -\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3} = 0.918 \\
\text{Ent}(D^{\text{浅白}}) &= -\frac{1}{5}\log_2\frac{1}{5} - \frac{4}{5}\log_2\frac{4}{5} = 0.212 \\
\text{Gain}(D,\text{色泽}) &= \text{Ent}(D) - \text{Ent}(D|a) \\
&= \text{Ent}(D) - \sum_{i=1}^{V} \frac{|D^i|}{|D|} \text{Ent}(D^i) \\
&= 0.998 - \frac{6}{17} \times 1 - \frac{6}{17} \times 0.918 - \frac{5}{17} \times 0.212 \\
&= 0.108
\end{aligned}
$$

2）计算属性“触感”的信息增益率

$$
\text{Gain\_ratio}(D,a) = \frac{\text{Gain}(D,a)}{\text{IV}(a)}
$$

$$
IV(a) = -\sum_{i=1}^{V} \frac{|D^i|}{|D|} \log_2 \frac{|D^i|}{|D|}
$$

$$
\begin{aligned}
\text{Ent}(D) &= -\frac{8}{17}\log_2\frac{8}{17} - \frac{9}{17}\log_2\frac{9}{17} = 0.998 \\
\text{Ent}(D^{\text{硬滑}}) &= -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2} = 1 \\
\text{Ent}(D^{\text{软粘}}) &= -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5} = 0.971 \\
\text{Gain}(D,\text{触感}) &= \text{Ent}(D) - \text{Ent}(D|a) \\
&= \text{Ent}(D) - \sum_{i=1}^{V} \frac{|D^i|}{|D|} \text{Ent}(D^i) \\
&= 0.998 - \frac{12}{17} \times 1 - \frac{5}{17} \times 0.971 \\
&= 0.006 \\
\text{IV}(\text{触感}) &= -\frac{12}{17}\log_2\frac{12}{17} - \frac{5}{17}\log_2\frac{5}{17} = 0.874 \\
\text{Gain\_ratio}(D,\text{触感}) &= \frac{0.006}{0.874} = 0.007
\end{aligned}
$$

3）计算属性“脐部”的Gini指数

$$\text{Gini}(D)=\sum_{i=1}^{C}\sum_{j\neq i}p_ip_j=1-\sum_{i=1}^{C}p_i^2$$

$$\text{Gini\_index}(D,a)=\sum_{i=1}^{V}\frac{|D^i|}{|D|}\text{Gini}(D^i)$$

$$\begin{aligned}\text{Gini\_index}(D,\text{脐部})&=\sum_{i=1}^{V}\frac{|D^i|}{|D|}\text{Gini}(D^i)\\&=\frac{7}{17}\times(1-(\frac{5}{7})^2-(\frac{2}{7})^2)+\frac{6}{17}\times(1-(\frac{1}{2})^2-(\frac{1}{2})^2)+\frac{4}{17}\times(1-1^2)\\&=0.345\end{aligned}$$