人工智能基础

实验 2

完成截止时间：2019/5/19/23:59
提交至：ailab_2019@163.com
助教：姜庆彩【QQ：649561941】
林思捷【QQ：419803495】
梁聪【QQ：386166518】

数据描述

EEG 即 Electroencephalography，脑电图数据。实验压缩包内提供 2 个 EEG 数据集，均为被试人员在观看特点的视频片段过程中的脑电图记录，其中 DEAP 数据集给出的标签为被试人 id、视频 id、valence_arousal 标签，MAHNOB-HCI 数据集给出被试人 id、视频 id、valence_arousal 标签和情感类别。

数据文件格式均为 txt，详见其中的 README.txt。

算法简介

Support Vector Machine

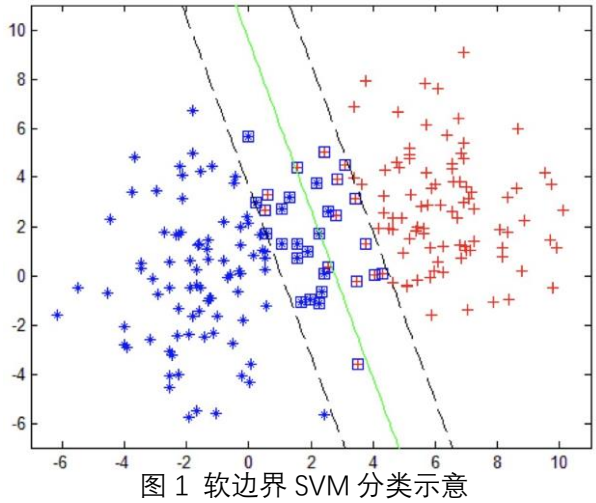

图 1 软边界 SVM 分类示意

在样本不完全线性可分时，为了使 SVM 正常工作，需要引入软边界（soft margin）。如图 1 所示，在 SVM 计算最优分类界面时允许部分样本位于两方边界（margin）之间，计算中需要最大化 margin 并最小化落入 margin 之间的样本点与对应标签的 margin 之间的距离之和。

原始 SVM 要求所有样本满足约束$\boldsymbol{y_i(X_i^T W + b) \geq 1}$，软边界 SVM 允许部分样本不满足

该约束，但是对于不满足约束的点增加惩罚为$\mathbf{1}-\boldsymbol{y_i}(\boldsymbol{X_i^T W}+\boldsymbol{b})$，统一所有样本的惩罚表示为

$$\text{loss}(X_i)=\max(0,1-y_i(X_i^T W+b))$$

则 SVM 优化目标变为

$$\min_{W,b}\frac{1}{2}\|W\|^2+C\sum loss(X_i)$$

其中 C 为惩罚系数，设置得越小则边界越“软”，设置越大则边界越“硬”，C 为无穷大时退化为基础版 SVM。

对于非线性可分数据或高维数据，在使用 SVM 时常引入 Kernel 方法进行降维，即使用某一参数确定的函数对样本数据进行变换，SVM 在变换后的空间对样本进行分类。常用的 Kernel 为高斯核函数（RBF 核函数）$K(x_i,x_j)=\exp(-\frac{\|x_i-x_j\|^2}{2\sigma^2})$、多项式核函数$K(x_i,x_j)=(x_i\cdot x_j+1)^p$等，基础版 SVM 可看作使用线性核函数。

Multilayer Perceptron

多层感知机用于分类任务时，通常将类别转换为独热码（one-hot code），长度为类别总数 m。在 MLP 输出层设置 m 个神经元，每个神经元的输出对应独热码的一个位，即假设当前样本标签为 01000，则期望输出层第二个神经元输出 1，其他输出层神经元输出 0。

可使用的激活函数有 sigmoid、tanh、ReLu、Leaky ReLu 等。

为了防止梯度爆炸、分类结果仅受个别神经元影响，同时防止过拟合，MLP 常引入正则化机制，常用的正则化机制有对权值进行 L1 正则化、L2 正则化、对神经元 drop-out 等。

Naïve Bayes

朴素贝叶斯假设样本的每一维属性互相独立，分类公式为

$$P(c\mid\mathbf{x})=\frac{P(c)P(\mathbf{x}\mid c)}{P(\mathbf{x})}=\frac{P(c)}{P(\mathbf{x})}\prod_{i=1}^{d}P(x_i\mid c)$$

其中 c 为某一标签，d 为属性数量，x_i为当前样本的属性 i 的取值。对所有标签c_k进行计算后，取概率最大的c_k作为分类结果。

如果测试集某个属性的某个取值没有在训练样本中出现过，那么分类时仅使用基础公式会导致结果概率为 0，为了防止这种情况需要对朴素贝叶斯公式进行拉普拉斯修正。

Decision Tree

设某一样本集 S 包含 N 个样本，其中属于每一类别c_i的样本个数为n_i，则该样本集合的信息熵为

$$\mathrm{H(S)}=\mathrm{E}\left[\log\left(\frac{1}{P(c_i)}\right)\right]=-\sum P(c_i)\log(P(c_i))$$

其中$P(c_i)=\frac{n_i}{N}$。

决策树分类时，每次需选取一个属性a_i，并选取一个切分值将样本集分为S_1和S_2，计算信息熵增益。信息熵增益定义为

$$Gain(a_i) = H(S) - \frac{\|S_1\|}{\|S\|}H(S_1) - \frac{\|S_2\|}{\|S\|}H(S_2)$$

对所有属性的最优切分界面进行计算后，选取信息熵增益最大的属性进行切分。

K Nearst Neighbor

KNN 作为查找算法，在样本集过大时运行效果不佳（虽然本次实验样本集很小），为了加快运行速度可对样本进行结构化存储，减少对新样本进行分类时所需查找的样本数量。

K 的大小设置将显著影响结果。

Logistic Regression（LR）

可以 kernel，类似 SVM 的 Kernel。

数据处理

将 2 个数据集分别随机平均分为 5 份并保存，要求保证每个被试人员出现在每份样本集中的数量均匀。后续测试时使用算法在每个数据集的 5 份数据上进行 5 折交叉检验（5-fold cross validation），记录平均结果。

实验要求

1. 数据预处理，要求如上。可以对数据属性进行任意加工处理，比如删减、降维、组合等，在实验中测试其影响。**（5%）**
2. 使用非线性 kernel 的软边界 SVM 对 2 个数据集的 valence_arousal_label 进行分类，分析结果；**（30%）**
3. 使用带有 drop-out 的 MLP 对 2 个数据集的 valence_arousal_label 进行分类，对 2 个数据集的被试人员（subject id）进行分类，对 MAHNOB-HCI 数据集的情感标签进行分类，分析结果；**（30%）**
4. 使用带拉普拉斯修正的 Naïve Bayes 对 2 个数据集的被试人员（subject id），对 MAHNOB-HCI 数据集的情感标签进行分类，分析结果；**（20%）**
5. 选取决策树、KNN、LR 其中之一，或者任意其他一个你感兴趣的分类器，对 2 个数据集的 valence_arousal_label 进行分类，或者任选你感兴趣的标签进行分类，分析结果；**（15%）**
6. 编程语言不限，可以使用任何现有工具包，手动实现 SVM、MLP 或 Naïve Bayes 的对应部分成绩×1.2，最后课程实验及作业总分不超过总评的 40%（即能补贴实验及作业扣的分）。要求对代码进行详细注释，推荐注释行数与代码行数比值为 30%~60%，注释太少会影响成绩。
7. 实验报告中需要记录实验语言及版本，对程序代码关键部分进行说明，说明数据处理方法，记录测试方法及测试结果，分析结果。
8. 严禁抄袭，批改时会进行代码查重，抄袭者 2 人（或多人）均为 0 分。实验分为上述 5 部分，缺少的部分扣对应分数。

实验提交

在实验截止时间之前将源代码和实验报告打包压缩后发送到 ailab_2019@163.com，邮邮件主题为**学号_姓名_lab02**，如果需要重复提交则主题为**学号_姓名_lab02_重交 n**，其中 n 为重交次数。如果有重交则之前提交作废。如果没有收到自动回复请联系助教。

压缩包命名为**学号_姓名_lab02**（.压缩包后缀），仅限包含一个同名文件夹**学号_姓名_lab02**。一共 4 个分类器各需要 1 份 pdf 格式实验报告，每个分类器单独建立一个子文件夹存放源代码和实验报告。

逾期提交则本次实验得分×0.7。**5 月 26 日 23:59:59** 之后不接受补交。

对实验有任何问题请尽快联系助教