媒体与认知 第四次作业

(V1.1　　2020.5.20)

本次作业为三道习题，其中第三道习题有利用 SVM 工具包进行上机求解的环节。

一、线性判别分析（课件第 4.3.2 节，4 月 23 日课程内容）

已知两类样本数据如下：

正样本： $x_1=(-2,1)^T, x_2=(-1,1)^T, x_3=(0,3)^T$ ；

负样本： $x_4=(0,-2)^T, x_5=(2,0)^T, x_6=(3,0)^T$ 。

正样本 x_1, x_2, x_3 的均值为 μ_1 ， 负样本 x_4, x_5, x_6 均值为 μ_2 。

请计算两类样本的线性判别分析（Linear Discriminant Analysis, LDA）投影方向 $w=S_w^{-1}(\mu_1-\mu_2)$ 。

附：线性判别分析原理

对于多类样本，样本类间散布矩阵(between classes scatter matrix) S_b 表示各类中心围绕总体均值的散布情况：

$$S_b=\sum_{i=1}^{C}\frac{n_i}{N}(\mu_i-\mu)(\mu_i-\mu)^T$$

其中， μ ， N 分别为总体样本的均值和数目； μ_i ， n_i 分别为第 i 类样本的均值和数目，共有 C 类样本。 S_b 的秩是 $C-1$ 。

类内散布矩阵(within classes scatter matrix) S_w 表示样本点 x 围绕各类均值的散布情况， x 为原始 D 维特征向量：

$$S_w=\frac{1}{N}\sum_{i=1}^{C}\sum_{k=1}^{n_i}(x_k^i-\mu_i)(x_k^i-\mu_i)^T \text{ 。}$$

设矩阵 $S_w^{-1}S_b$ 的本征值为 $\lambda_1,\lambda_2,...,\lambda_D$, 按降序排列，取前 d 个本征值对应的本征向量构成 $D\times d$ 维矩阵 W ，即为： $W=[v_1,v_2,...v_d]$ 。

经过特征变换，降维得到 d 特征向量 $y=W^Tx$ 。

两类情况下，$S_w^{-1}S_b$ 的秩是 $C-1=1$，因此，$S_w^{-1}S_b$ 只有一个非零本征值，W 是 $D\times 1$ 维矩阵，$W=w$。为求 $S_w^{-1}S_b$ 的本征值应解方程：

$S_w^{-1}S_b w=\lambda w$

当两类样本数目相同，即为：$\frac{1}{4}S_w^{-1}(\mu_1-\mu_2)(\mu_1-\mu_2)^T w=\lambda w$。

由于 $(\mu_1-\mu_2)^T w$ 为行向量与列向量相乘，所得为标量，因此，投影方向 $w=S_w^{-1}(\mu_1-\mu_2)$。

二、隐含马尔可夫模型 (课件第 5.3 节，5 月 14 日课程内容)

考虑采用隐含马尔可夫模型(HMM)对 DNA 进行分析，DNA 序列的碱基有腺嘌呤 A、胞嘧啶 C、鸟嘌呤 G 和胸腺嘧啶 T，共 4 种基本类型。假设有一个隐含状态 S 控制着 DNA 序列的生成，S 有两个可能的状态 $\{S_1,S_2\}$，假定模型 λ 有如下的状态转移概率：

$P(S_1|S_1)=0.7$，$P(S_2|S_1)=0.3$，$P(S_1|S_2)=0.2$，$P(S_2|S_2)=0.8$

由状态到观测值的发射概率为：

$P(A|S_1)=0.4$，$P(C|S_1)=0.1$，$P(G|S_1)=0.4$，$P(T|S_1)=0.1$

$P(A|S_2)=0.1$，$P(C|S_2)=0.4$，$P(G|S_2)=0.1$，$P(T|S_2)=0.4$

初始状态分布为 $P(S_1)=0.5$，$P(S_2)=0.5$。

假定观测序列 $x=ACT$，请计算：

(1) $P(x|\lambda)$

要求采用前向变量法求解，请写出计算过程。

(2) 最可能的隐含状态序列

要求采用 Viterbi 算法求解，请写出计算过程。

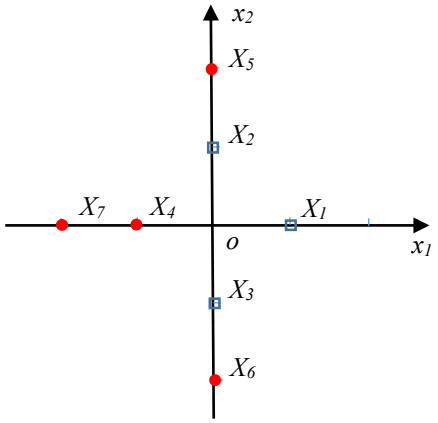
三、支持向量机 (课件第 6.2 节，5 月 21 日课程内容)

考虑如下训练集样本，

负样本： $X_1=(1,0),Y_1=-1$； $X_2=(0,1),Y_2=-1$； $X_3=(0,-1),Y_3=-1$；

正样本： $X_4=(-1,0),Y_4=+1$； $X_5=(0,2),Y_5=+1$； $X_6=(0,-2),Y_6=+1$；

$X_7=(-2,0),Y_7=+1$。



(1). 用如下的非线性变换，先将输入的样本 $X=(x_1,x_2)$ 变换为向量 $Z=(\phi_1(X),\phi_2(X))$：

$\phi_1(X)=x_2^2-2x_1+3$， $\phi_2(X)=x_1^2-2x_2-3$；

再通过作图观察或利用 libSVM 等工具包求解，求得 Z 空间中两类样本线性分类问题的支持向量机模型参数，再写出 X 空间中的决策函数 $g_m(X)=g_m(x_1,x_2)$。

(2). 在不对原训练集样本进行从 X 到 Z 空间的显式映射情况下，采用如下核函数：

$$K(X_i,X_j)=(1+X_i^TX_j)^2=(1+x_{i1}x_{j1}+x_{i2}x_{j2})^2$$

写出采用该核函数的支持向量机 Lagrangian 对偶问题的目标函数（拉格朗日乘子 $\alpha_i\geq 0,i=1,...,7$）。通过上机求解，列出 $\alpha_i,i=1,...,7$ 的取值，由此指出哪些数据是支持向量，并写出决策函数 $g_k(X)=g_k(x_1,x_2)$。

注： $g_k(X)=\mathbf{w}^T\phi(X)+b=\sum_{j\in SV}y_j\alpha_jK(X_j,X)+b$

(3). 分析比较第（1）步和第（2）步中求解得到的决策函数 $g_m(x_1,x_2)$ 和 $g_k(x_1,x_2)$，说明核函数的作用。

(4). 分别利用 $g_m(x_1,x_2)$ 和 $g_k(x_1,x_2)$，对样本 $X_8=(0,0)$， $X_9=(2,2)$ 的类别进行预测，得到 Y_{8m}， Y_{8k}， Y_{9m}， Y_{9k}。

附：上机说明

请选择一种支持向量机程序工具（如 libsvm），阅读其源代码，在了解算法原理和技术实现细节基础上，完成习题中的上机求解环节。

作业附件中的 \libsvm 目录为 LIBSVM 工具包 (http://www.csie.ntu.edu.tw/~cjlin/libsvm/)。

具体使用说明可参见其中的 README。

一、Windows 平台程序运行方式如下：

1. 将附件程序解压至本机上，比如，c:\libsvm。
2. 运行 Windows 命令行终端程序 cmd，先运行如下命令进入可执行程序所在目录：
   cd c:\libsvm\windows
   运行如下命令，得到习题第(1)步 Z 空间中模型参数：
   svm-train -t 0 -c 1000 with-mapping.txt

   运行如下命令，得到习题第(2)步模型参数：
   svm-train -t 1 -d 2 -g 1 -r 1 with-kernel.txt

   svm-train 相关参数说明为：

```
-t kernel_type : set type of kernel function (default 2)
  0 -- linear: u'*v
  1 -- polynomial: (gamma*u'*v + coef0)^degree
  …
-d degree : set degree in kernel function (default 3)
-g gamma : set gamma in kernel function (default 1/num_features)
-r coef0 : set coef0 in kernel function (default 0)
```

   其中，with-mapping.txt 和 with-kernel.txt 为样本数据，格式为：

   <label> <index1>:<value1> <index2>:<value2> ...

<label> 是训练数据集的目标值，对于分类，它是类别标号(支持多个类)；对于回归，是任意实数。<index> 是以 1 开始的整数，表示特征维数序号；<value>为特征向量对应该维数序号的数值。当特征向量数据是稀疏数据，<index>可以不连续，即省略特征数值为 0 的维数序号及数值。

3. 输出的模型参数保存在文件 with-mapping.txt.model 和 with-kernel.txt.model 中，可用文本编辑器打开*.model 文件查看，其中：
   rho 是偏置量 (-b)，即 b = -rho。
   SV列表为支持向量，格式为：
   <sv_coef> <index1>:<value1> <index2>:<value2> ...
   第 j 行中，<sv_coef> 为第 j 个支持向量对应的系数 $y_j\alpha_j$；<index1>:<value1> <index2>:<value2> ... 为该支持向量的样本数据，格式与训练样本数据相同。

二、Linux或MAC平台程序运行方式如下：

请先解压附件，在命令行终端进入libsvm目录，用Make命令编译生成可执行程序；再执行如下命令：

./svm-train -t 1 -d 2 -g 1 -r 1 with-kernel.txt

“with-kernel.txt”需要先从 libsvm\windows 子目录拷贝到 libsvm 目录中。

本次作业责任助教为闫睿劼（Email: yrj17@mails.tsinghua.edu.cn）。

【关于作业迟交的说明】 请同学们争取按时提交作业，迟交会酌情扣分。