编程练习6：

支持向量机

机器学习

# 介绍

在本练习中，您将使用支持向量机（SVM）构建垃圾邮件分类器。在开始编程练习之前，我们强烈建议观看视频讲座并完成相关主题的复习问题。

要开始练习，您需要下载起始代码并将其内容解压缩到您希望完成练习的目录中。如果需要，在开始本练习之前，请使用octave/matlab中的cd命令更改到此目录。

您也可以在课程网站的“环境设置说明”中找到安装octave/matlab的说明。

## 本练习中包含的文件

ex6.m-练习前半部分的八度/matlab脚本

ex6data1.mat-示例数据集1

ex6data2.mat-示例数据集2

ex6data3.mat-示例数据集3

svmtrain.m-svm训练函数

svmpredit.m-svm预测函数

plot data.m-plot 2d数据可视化

boundary线性.m-plot线性边界可视化

boundary.m-plot非线性边界线性

arkernel.m-svm线性核

[]gaussian kernel.m-支持向量机的高斯核*？*

[]dataset3params.m-用于数据集3的参数*？*

ex6 spam.m-下半部分练习的八度/matlab脚本

spamtrain.mat-垃圾邮件培训集spam test.mat-垃圾邮件测试集email sample 1.txt-示例电子邮件1

email sample 2.txt-示例电子邮件2

spamsample1.txt-示例垃圾邮件1

spamsample2.txt-示例垃圾邮件2

vocab.txt-词汇表

getvocablist.m-加载词汇表

porterstemmer.m-词干函数

readfile.m-将文件读取到字符串

submit.m-将您的解决方案发送到我们的服务器的提交脚本

[]processemail.m-电子邮件预处理*？*

[]emailfeatures.m-从电子邮件中提取功能*？*

*？*指示需要完成的文件

在整个练习中，您将使用脚本ex6.m。这些脚本为问题设置数据集，并调用将要编写的函数。您只需要按照本分配中的说明修改其他文件中的函数。

## 在哪里寻求帮助

本课程的练习使用Octaveor Matlab，这是一种高级编程语言，非常适合数值计算。如果您没有安装八度音阶或matlab，请参考课程网站“环境设置说明”中的安装说明。[[ 1 ]](" \l "_ftn1" \o ")

在octave/matlab命令行中，键入帮助，然后键入函数名，显示内置函数的文档。例如，帮助图将显示用于绘制的帮助信息。有关八度函数的更多文档，请参见八度文档页面。matlab文档可以在matlab文档页面中找到。

我们还强烈鼓励使用在线讨论与其他学生讨论练习。但是，不要查看其他人编写的任何源代码，也不要与其他人共享源代码。

# 1             支持向量机

在本练习的前半部分，您将使用支持向量机（SVM）和各种示例二维数据集。实验这些数据集将帮助您获得SVM如何工作以及如何将高斯核与SVM结合使用的直觉。在接下来的半个练习中，您将使用支持向量机构建垃圾邮件分类器。

提供的脚本ex6.m将帮助您完成练习的前半部分。

## 1.1           示例数据集1

我们将从二维示例数据集开始，该数据集可以由线性边界分隔。脚本ex6.m将绘制培训数据（图1）。在这个数据集中，正例（用+表示）和负例（用O表示）的位置表明间隙指示的自然分离。但是，请注意，在最左边的大约（0.1,4.1）处有一个离群值正示例+。作为本练习的一部分，您还将看到该异常值如何影响SVM决策边界。

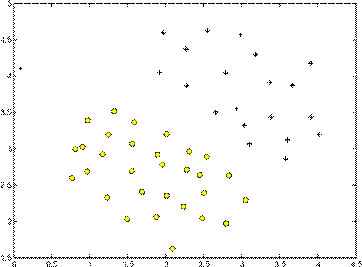


图1：示例数据集1

在练习的这一部分中，您将尝试在SVMS中使用C参数的不同值。非正式地，c参数是一个正值，它控制了错误分类训练示例的惩罚。一个大的C参数告诉SVM要对所有的例子进行正确的分类。c的作用类似于，其中，λ是我们以前用于逻辑回归的正则化参数。

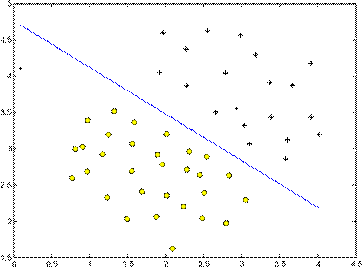


图2:c=1的SVM决策边界（示例数据集1）

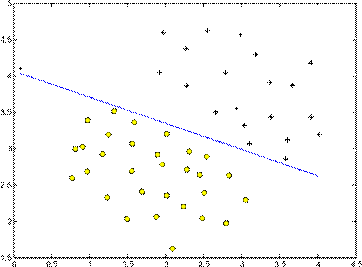


图3:c=100的SVM决策边界（示例数据集1）

ex6.m中的下一部分将使用我们包含在起始代码svmtrain.m中的svm软件来运行svm培训（c=1）。当c=1时，您应该发现svm将决策边界放在两个数据集之间的间隙中，并对最左边的数据点进行错误分类（图2）。[〔2〕](" \l "_ftn2" \o ")

**实施说明：**大多数SVM软件包（包括svmtrain.m）会自动为您添加额外的功能x0=1，并自动学习截距项θ0。因此，当您将培训数据传递给SVM软件时，不需要自己添加这个额外的功能x0=1。尤其是在Octave/Matlab中，您的代码应该使用训练示例x∈r（而不是x∈r）；例如，在第一个示例中，数据集x∈r。*nn*+12

您的任务是在此数据集中尝试不同的C值。具体来说，您应该将脚本中的c值更改为c=100，然后再次运行SVM培训。当c=100时，您应该发现SVM现在能够正确地对每个示例进行分类，但是决策边界似乎不适合数据（图3）。

## 1.2           高斯核支持向量机

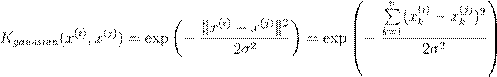
在本部分练习中，您将使用SVM进行非线性分类。特别是，您将在不可线性分离的数据集上使用带有高斯核的SVM。

### 1.2.1          高斯核

为了利用支持向量机寻找非线性决策边界，首先需要实现高斯核。可以将高斯核看作是一个相似函数，它测量一对例子（x（i），x（j））之间的“距离”。高斯核也由带宽参数σ参数化，该参数决定相似性度量降低（到0）的速度，因为示例之间的距离更大。

现在您应该完成gaussiankernel.m中的代码来计算

两个例子之间的高斯核（x（i），x（j））。高斯核函数定义为：

*.*

完成函数gaussiankernel.m后，脚本ex6.m将在提供的两个示例上测试内核函数，您应该看到0.324652的值。

*现在您应该提交解决方案。*

### 1.2.2          示例数据集2



图4：示例数据集2

ex6.m的下一部分将加载并绘制数据集2（图4）。从图中可以看出，没有将此数据集的正示例和负示例分开的线性决策边界。然而，通过将高斯核与支持向量机结合使用，您将能够学习一个非线性决策边界，该边界可以很好地执行数据集。

如果您已经正确地实现了高斯核函数，ex6.m将继续在这个数据集中用高斯核训练SVM。

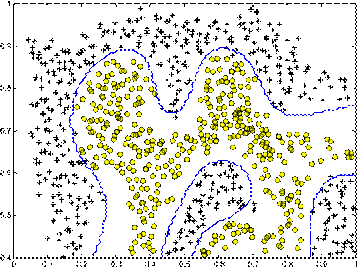


图5:SVM（高斯核）决策边界（示例数据集2）

图5显示了带有高斯核的SVM发现的决策边界。决策边界能够正确区分大部分正、负示例，并能很好地跟踪数据集的轮廓。

### 1.2.3          示例数据集3

在本部分练习中，您将获得更多关于如何使用带有高斯核的SVM的实用技能。ex6.m的下一部分将加载并显示第三个数据集（图6）。您将在这个数据集中使用带有高斯内核的SVM。

在提供的数据集ex6data3.mat中，为您提供了变量x、y、xval、yval。ex6.m中提供的代码使用训练集（x，y）使用从dataset3params.m加载的参数训练SVM分类器。

您的任务是使用交叉验证集xval、yval来确定要使用的最佳C和σ参数。您应该编写任何其他必要的代码来帮助您搜索参数c和σ。对于c和σ，我们建议用乘法步骤尝试值（例如，0.01、0.03、0.1、0.3、1、3、10、30）。注意，您应该尝试所有可能的C和σ值对（例如，C=0.3和σ=0.1）。例如，如果您尝试上面列出的8个值中的每一个，对于C和σ2，您将最终培训和评估（在交叉验证集上）总共82=64个不同的模型。

在确定了要使用的最佳C和σ参数后，应该修改dataset3params.m中的代码，填充最佳参数。

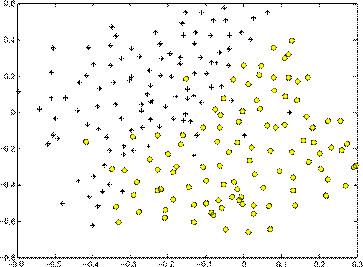


图6：示例数据集3

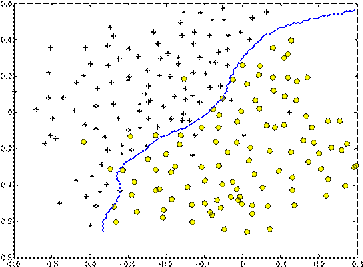


图7:SVM（高斯核）决策边界（示例数据集3）

你找到了。对于我们的最佳参数，SVM返回一个决策边界，如图7所示。

**实施提示：**在执行交叉验证以选择要使用的最佳C和σ参数时，需要评估交叉验证集上的错误。回想一下，对于分类，错误定义为错误分类的交叉验证示例的分数。在octave/matlab中，可以使用mean（double（predictions~=yval））计算此错误，其中predictions是包含来自SVM的所有预测的向量，yval是来自交叉验证集的真正标签。可以使用svmpredit函数为交叉验证集生成预测。

*现在您应该提交解决方案。*



# 2             垃圾邮件分类

如今，许多电子邮件服务都提供垃圾邮件过滤器，能够将电子邮件准确地分为垃圾邮件和非垃圾邮件。在本部分练习中，您将使用SVM构建自己的垃圾邮件过滤器。

您将培训一个分类器来分类给定的电子邮件x是垃圾邮件（y=1）还是非垃圾邮件（y=0）。特别是，您需要将每个电子邮件转换为特征向量x∈r。下面的练习将引导您了解如何从电子邮件构造这样的特征向量。*n*

在本练习的其余部分，您将使用脚本ex6 spam.m。本练习包含的数据集基于SpamAssassin公共语料库的一个子集。在本练习中，您将只使用电子邮件正文（不包括电子邮件头）。[〔3〕](" \l "_ftn3" \o ")

## 2.1           预处理电子邮件

>有人知道托管一个门户网站要花多少钱吗？

>

好吧，这取决于你预计会有多少游客。这可以是任何地方，从每月不到10美元到100美元。如果你经营的是大型企业，你应该去http://www.rackspace.com/或者亚马逊EC2结账。

若要取消订阅此邮件列表，请发送电子邮件至：groupname-unsubscribe@egroups.com

图8：电子邮件示例

在开始机器学习任务之前，从数据集中查看示例通常是有见地的。图8显示了一个包含URL、电子邮件地址（在末尾）、数字和美元金额的示例电子邮件。虽然许多电子邮件包含类似类型的实体（例如，数字、其他URL或其他电子邮件地址），但几乎每个电子邮件中的特定实体（例如，特定URL或特定美元金额）都不同。因此，在处理电子邮件时经常使用的一种方法是“规范化”这些值，以便所有URL都被处理为相同的，所有数字都被处理为相同的，等等。例如，我们可以用唯一的字符串“httpaddr”替换电子邮件中的每个URL，以指示存在一个URL。

这样做的效果是，让垃圾邮件分类器根据是否存在任何URL，而不是特定的URL是否存在作出分类决策。这通常会提高垃圾邮件分类器的性能，因为垃圾邮件发送者通常会随机化URL，因此在新的垃圾邮件中再次看到任何特定URL的几率非常小。

在processemail.m中，我们实现了以下电子邮件预处理和规范化步骤：

•将整个电子邮件转换为小写，以便忽略标题化（例如，将“指示”视为“指示”）。**下套管：**

•从电子邮件中删除所有HTML标记。许多电子邮件通常带有HTML格式；我们删除所有HTML标记，以便只保留内容。**剥离HTML:**

•所有URL都替换为文本“httpaddr”。**规范化URL:**

•所有电子邮件地址均替换为文本“emailaddr”。**规范化电子邮件地址：**

•所有数字均替换为文本**规范化数字：**

“数字”。

•所有美元符号（$）均替换为文本“美元”。**正常化美元：**

•单词被简化为词干形式。例如，“折扣”、“折扣”、“折扣”和“折扣”都替换为“折扣”。有时，词干分析器实际上从结尾处去掉了附加字符，因此“include”、“include”、“included”和“include”都替换为“include”。**词干：**

•删除了非文字和标点符号。所有空格（制表符、换行符、空格）都已修剪为一个空格字符。**删除非文字：**

这些预处理步骤的结果如图9所示。虽然预处理留下了单词片段和非单词，但这种形式在执行特征提取时更容易使用。

任何人都知道网站的运营成本，这取决于你对Mani访问者的期望，无论是从每月不到几块钱到几块钱，你都应该通过httpaddr或perhap amazon ecnumb结账，如果你的运营规模大到取消订阅，请通过邮件列表向emailaddr发送电子邮件。

图9：预处理的示例电子邮件

|  |  |  |  |
| --- | --- | --- | --- |
| |  | | --- | | 1           AA  2           抗体  3           阿比尔…86安永…916知道…  1898    零  1899    拉链 | | |  | | --- | | 86 916 794 1077 883  370 1699 790 1822  1831 883 431 1171  794 1002 1893 1364  592 1676 238 162 89  688 945 1663 1120  1062 1699 375 1162 479 1893 1510 799  1182 1237 810 1895  1440 1547 181 1699  1758 1896 688 1676  992 961 1477 71 530  1699 531 | |

图10：词汇表图11：示例电子邮件的词汇索引

### 2.1.1          词汇表

在预处理电子邮件之后，我们为每个电子邮件都有一个单词列表（如图9）。下一步是选择要在分类器中使用的单词，以及要遗漏的单词。

对于本练习，我们只选择最常见的单词作为考虑的单词集（词汇表）。因为很少出现在训练集中的单词只出现在少数电子邮件中，它们可能会导致模型过度适合我们的训练集。完整的词汇表位于vocab.txt文件中，如图10所示。我们的词汇表是通过选择垃圾邮件语料库中出现至少100次的所有单词来选择的，结果得到了1899个单词的列表。在实践中，常用的词汇表大约有10000到50000个单词。

根据词汇表，我们现在可以将预处理电子邮件（如图9）中的每个单词映射到包含词汇表中单词索引的单词索引列表中。图11显示了示例电子邮件的映射。具体来说，在示例电子邮件中，单词“anyone”首先被规范化为“anyon”，然后映射到词汇表中的索引86。

现在您的任务是完成processemail.m中的代码以执行此映射。在代码中，会给您一个字符串str，它是处理过的电子邮件中的一个单词。您应该在词汇表词汇表中查找该单词，并找出该单词是否存在于词汇表中。如果单词存在，则应将单词的索引添加到单词索引变量中。如果单词不存在，因此不在词汇表中，则可以跳过该单词。

一旦您实现了processemail.m，脚本ex6 spam.m将在电子邮件示例上运行您的代码，您将看到类似于图9和11的输出。

**八度/matlab提示：**在octave/matlab中，可以使用strcmp函数比较两个字符串。例如，只有当两个字符串相等时，strcmp（str1，str2）才会返回1。在提供的起始代码中，vocalBlist是一个包含词汇表中单词的“单元数组”。在Octave/Matlab中，单元格数组就像一个普通数组（即向量），只是它的元素也可以是字符串（在普通的Octave/Matlab矩阵/向量中它们不能），并且可以使用大括号而不是方括号对其进行索引。具体来说，要在索引I中获得单词，可以使用vocalist i。您还可以使用length（词汇表）来获取词汇表中的单词数。

*现在您应该提交解决方案。*

## 2.2           从电子邮件中提取功能

现在，您将实现将每个电子邮件转换为r中的向量的功能提取。对于本练习，您将在词汇表中使用n=单词。具体而言，电子邮件的特征X{{0,1}对应于字典中的第i个词是否出现在电子邮件中。即，如果在电子邮件中不存在第i个字，则如果第i个字在电子邮件中且XI＝0，则XI＝1。*n*

因此，对于典型的电子邮件，此功能如下所示：

（0）

.

…γ

γ

（1）

γ

（0）

x=…阿尔*n.名词*

γ

（1）

γ

（0）

小精灵…小精灵

小精灵

γ

0

现在您应该完成emailfeatures.m中的代码，根据单词索引为电子邮件生成一个特征向量。

一旦您实现了emailfeatures.m，ex6的下一部分spam.m将在电子邮件示例上运行您的代码。您应该看到特征向量有长度1899和45个非零条目。

*现在您应该提交解决方案。*

## 2.3           为垃圾邮件分类培训SVM

完成特征提取功能后，ex6 spam.m的下一步将加载一个预处理的训练数据集，该数据集将用于训练SVM分类器。spamtrain.mat包含4000个垃圾邮件和非垃圾邮件的培训示例，而spam test.mat包含1000个测试示例。使用processemail和emailfeatures函数处理每个原始电子邮件，并将其转换为向量x（i）∈r。1899

加载数据集后，ex6 spam.m将继续训练SVM，以便在垃圾邮件（y=1）和非垃圾邮件（y=0）之间进行分类。培训完成后，您应该看到分类器的培训精度约为99.8%，测试精度约为98.5%。

## 2.4           垃圾邮件的主要预测因素

我们的点击删除保证访问basenumb dollar将为大多数lo ga dollarnumb定价。

图12：垃圾邮件的预测因素

为了更好地理解垃圾邮件分类器是如何工作的，我们可以检查参数，看看分类器认为哪些单词对垃圾邮件最具预测性。ex6 spam.m的下一步是在分类器中查找具有最大正值的参数，并显示相应的单词（图12）。因此，如果一封电子邮件包含“guarantee”、“remove”、“dollar”和“price”（图12中显示的最重要的预测值）等词，则很可能被归类为垃圾邮件。

## 2.5           可选（未分级）练习：尝试自己的电子邮件

既然你已经训练了一个垃圾邮件分类器，你就可以开始在自己的电子邮件中尝试它了。在起始代码中，我们包含了两个电子邮件示例（emailsample1.txt和emailsample2.txt）和两个垃圾邮件示例（spamsample1.txt和spamsample2.txt）。ex6的最后部分m在第一个垃圾邮件示例上运行垃圾邮件分类器，并使用所学的SVM对其进行分类。现在，您应该尝试我们提供的其他示例，看看分类器是否正确。您还可以尝试自己的电子邮件，用自己的电子邮件替换示例（纯文本文件）。

*您不需要为此可选（未分级）练习提交任何解决方案。*

## 2.6           可选（未分级）练习：构建自己的数据集

在这个练习中，我们提供了一个预处理的训练集和测试集。这些数据集是使用您现在已经完成的相同功能（processemail.m和emailfeatures.m）创建的。对于这个可选（未分级）练习，您将使用来自SpamAssassin公共语料库的原始电子邮件构建自己的数据集。

在这个可选（未分级）练习中，您的任务是从公共语料库下载原始文件并将其提取出来。在提取它们之后，您应该在每个电子邮件上运行processemailand emailfeatures函数，从每个电子邮件中提取一个特征向量。这将允许您构建示例的数据集x，y。然后您应该随机地将数据集分为一个训练集、一个交叉验证集和一个测试集。[〔4〕](" \l "_ftn4" \o ")

当您构建自己的数据集时，我们还鼓励您尝试构建自己的词汇表（通过选择数据集中出现的高频词）并添加您认为可能有用的任何其他功能。

最后，我们还建议尝试使用高度优化的SVM工具箱，如libsvm。

*您不需要为此可选（未分级）练习提交任何解决方案。*

# 提交和评分

完成任务的各个部分后，请务必使用提交功能系统将您的解决方案提交给我们的服务器。下面是这个练习的每个部分的评分方式的细目。

|  |  |  |
| --- | --- | --- |
| **部分** | **提交的文件** | **点** |
| 高斯核 | Gaussiankernel.m.先生 | 25点 |
| 数据集3的参数（c，σ） | 数据集3参数.m | 25点 |
| 电子邮件预处理 | 处理电子邮件.m | 25点 |
| 电子邮件功能提取 | 电子邮件功能.m | 25点 |
| 总分 |  | 100点 |

您可以多次提交解决方案，我们只考虑最高分数。

[[ 1 ]](" \l "_ftnref1" \o ")八度是matlab的免费替代品。对于编程练习，您可以自由使用八度或matlab。

[〔2〕](" \l "_ftnref2" \o ")为了确保与Octave/Matlab的兼容性，我们已经包含了SVM学习算法的实现。然而，这个特定的实现被选择来最大化兼容性，并且不是非常有效。如果您正在针对实际问题培训SVM，特别是需要扩展到更大的数据集时，我们强烈建议您使用高度优化的SVM工具箱，如libsvm。

[〔3〕](" \l "_ftnref3" \o ")http://spamasassin.apache.org/old/publiccorpus/

[〔4〕](" \l "_ftnref4" \o ")原始电子邮件将包含您可能希望忽略的电子邮件头。我们在processemail中包含了有助于删除这些头的代码。