Midge Classify

author:夏华林*

2014 年 8 月 21 日

Contents

Abstract

1 引言

1.1 问题简述

背景 两种蠓Af和Apf已由生物学家W.L.Grongan 和W.W.Wirth(1981年)根据它们的触角长度和翼长加以区分，下表为样本取样数据：

*Email: hua_lin@live.cn Tel: +86–18242360705

样本	触角长(mm)	翼长(mm)	类别	样本	触角长(mm)	翼长(mm)	类别
1	1.1400	1.7800	Af	7	1.2400	1.7200	Apf
2	1.1800	1.9600	Af	8	1.3600	1.7400	Apf
3	1.2000	1.8600	Af	9	1.3800	1.6400	Apf
4	1.2600	2.0000	Af	10	1.3800	1.8200	Apf
5	1.2800	2.0000	Af	11	1.3800	1.9000	Apf
6	1.3000	1.9600	Af	12	1.4000	1.7000	Apf
				13	1.4800	1.8200	Apf
				14	1.5400	1.8200	Apf
				15	1.5600	2.0800	Apf

根据给出的触角长度和翼长识别出一只标本是Af还是Apf是重要的。

问题 已知Af是宝贵的传粉益虫，Apf是某种疾病的载体，要求建立合适的数学模型，正确区分两类蠓虫。

1. 给定一只Af或者Apf族的蠓，你如何建立正确的数学模型来区分它属于哪一族?
2. 用你的模型来区分以下标本数据：

样本	触角长(mm)	翼长(mm)	类别
1	1.2400	1.8000	–
2	1.2800	1.8400	–
3	1.4000	2.0400	–
4	1.1800	1.6600	–
5	1.3200	1.7600	–
6	1.5200	1.8800	–
7	1.5000	2.0600	–

3. 因为Af是宝贵的传粉益虫，Apf是某种疾病的载体，针对这一情景，你的模型如何优化？

2 模型建立

2.1 问题抽象

根据蠓的一系列特征来区分不同的品种，本质上是一个**模式分类**问题。

抽象描述　现我们假定有一组 n 个 d 维[1]的训练样本集 $\chi\{\mathbf{x_1}, \mathbf{x_2}, \mathbf{x_3} \dots \mathbf{x_n}\}$，集合中每一个样本 $\mathbf{x} = (\tau_1, \tau_2, \tau_3 \dots \tau_d)^T$ 是由 d 个特征值[2] 组成的 d 维向量。χ 集合中的每一个元素都有一个类别属性 ω_i，其中 $\omega_i \in \{\omega_1, \omega_2\}$。

目标　在已有的训练集 χ 的基础上，来构建一个足够好的分类器 Ψ，使得 Ψ 能最大性能的区分未知样本数据集 $T\{\mathbf{x_1}, \mathbf{x_2}, \mathbf{x_3} \dots \mathbf{x_m}\}$。

2.2　主体思路

我们考虑把 d 维空间中的数据投影到一条直线上去。当然，即使不同类别的样本点能够在 d 维空间中形成 **相互分离 但各自内部紧凑 的集合**，向任意的直线做投影也有可能把这些不同类别的数据点混合在一起，反而降低了分类的效果。然而，通过适当的选择投影直线，我们还是有可能找到能够最大限度地区分各类数据点的投影方向。而这就是我们要寻找的**最佳分类器** Ψ。

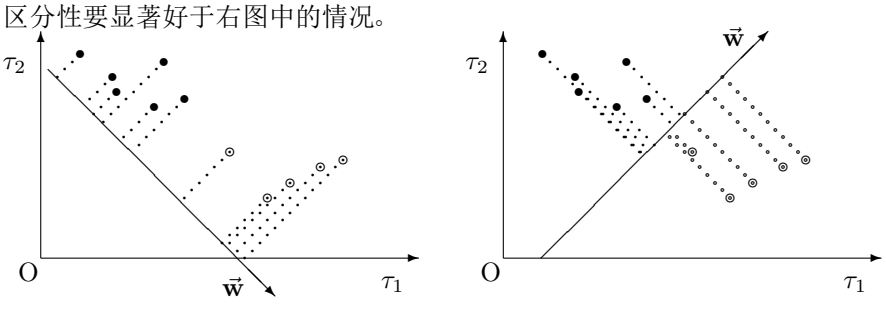
2.3　数学推导

考虑把 d 维数据 $\mathbf{x}$ 投影到方向为 $\mathbf{w}$ 的直线上，我们得到标量

$$y = \mathbf{w}^T\mathbf{x} \tag{1}$$

这样，集合 χ 中所有的 n 个样本 $\mathbf{x_1}, \mathbf{x_2}, \mathbf{x_3} \dots \mathbf{x_n}$ 就产生了 n 个结果 $y_1, y_2, y_3 \dots y_n$，相应的类别属性为 $\{\omega_1, \omega_2\}$。从几何上说，如果 $\|\mathbf{w}\| = 1$，那么每一个 y_i 就是把 $\mathbf{x}_i$ 向方向为 $\mathbf{w}$ 的直线上投影的结果。事实上 $\mathbf{w}$ 的大小并不重要，因为其效果不过是把 y_i 乘以一个标量倍数，我们关心的是 $\mathbf{w}$ 的方向。如果属于类别 ω_1 的样本数据与属于类别 ω_2 的样本数据能够在 d 维空间形成两个**显著分开的聚类**，那么我们希望它们在向直线做投影后应尽量的分开，而不是混在一起。

以下的例子显示相同的样本点在不同的直线上作投影，左图的投影点的可区分性要显著好于右图中的情况。



[1]在蠓的分类这一情景中，$d = 2$
[2]这里的区别于矩阵中的特征值

我们的目标是寻找一条直线能够最大限度地使样本点在直线上的投影能够尽量的分开，以此，我们定义抽象意义上的准则函数，而且，我们希望$min\, J(\bullet)$

$$J(\bullet) = \frac{\text{类内离散程度}}{\text{类间离散程度}} \tag{2}$$

集合χ中类别为ω_i的n_i样本均值

$$\mathbf{m}_i = \frac{1}{n_i}\sum_{\mathbf{x}\in\omega_i}\mathbf{x} \tag{3}$$

样本点$\mathbf{x}$向直线投影后的点的样本均值

$$\begin{aligned}\tilde{m}_i &= \frac{1}{n_i}\sum_{y\in\omega_i} y \\ &= \frac{1}{n_i}\sum_{\mathbf{x}\in\omega_i}\mathbf{w}^T\mathbf{x} = \mathbf{w}^T\mathbf{m}_i\end{aligned} \tag{4}$$

不同类别的样本点的投影点的均值差

$$\mid \tilde{m}_1 - \tilde{m}_2 \mid = \mid \mathbf{w}^T(\mathbf{m}_1 - \mathbf{m}_2) \mid \tag{5}$$

为了计算上的方便，我们定义投影点类间离散程度

$$\mathbf{s}_B = \mid \tilde{m}_1 - \tilde{m}_2 \mid^2 \tag{6}$$

虽然我们可以通过调整$\mathbf{w}$的大小，来调整类间离散度$\mathbf{s}_B$，但，不同类样本投影点的均值之差的大小总是相对而言的，否则问题就失去意义了，因此，我们定义类别ω_i的投影点类内离散程度

$$\tilde{s}_i^2 = \sum_{y\in\omega_i}(y - \tilde{m}_i)^2 \tag{7}$$

故，总的投影点类内离散程度为

$$\tilde{s}_W = \tilde{s}_1^2 + \tilde{s}_2^2 \tag{8}$$

由此，我们开始具体化准则函数$J(\bullet)$

$$J(\mathbf{w}) = \frac{\tilde{s}_W}{\mathbf{s}_B} \tag{9}$$

以上，是我们对样本在直线上投影点的相关分析，为了最小化准则函数$J(\bullet)$，我们还需要对样本自身进行分析。
定义样本类内离散程度矩阵$\mathbf{S}_i$和总的离散程度矩阵$\mathbf{S}_W$

$$\mathbf{S}_i = \sum_{\mathbf{x}\in\omega_i}(\mathbf{x} - \mathbf{m}_i)(\mathbf{x} - \mathbf{m}_i)^T \tag{10}$$

$$\mathbf{S}_W = \mathbf{S}_1 + \mathbf{S}_2 \tag{11}$$

然后我们有

$$\begin{aligned}
\tilde{s}_i^2 &= \sum_{\mathbf{x}\in\omega_i} (y - \tilde{m}_i)^2 \\
&= \sum_{\mathbf{x}\in\omega_i} (\mathbf{w}^T\mathbf{x} - \mathbf{w}^T\mathbf{m}_i)^2 \\
&= \sum_{\mathbf{x}\in\omega_i} \mathbf{w}^T(\mathbf{x} - \mathbf{m}_i)(\mathbf{w}^T\mathbf{x} - \mathbf{w}^T\mathbf{m}_i)^T \\
&= \sum_{\mathbf{x}\in\omega_i} \mathbf{w}^T(\mathbf{x} - \mathbf{m}_i)(\mathbf{x} - \mathbf{m}_i)^T\mathbf{w} \\
&= \mathbf{w}^T\mathbf{S}_i\mathbf{w}
\end{aligned} \tag{12}$$

由此，投影点的各类别离散程度之和可以写成

$$\tilde{s}_W = \tilde{s}_1^2 + s_2^2 = \mathbf{w}^T\mathbf{S}_W\mathbf{w} \tag{13}$$

类似的，样本投影点的均值之差可以展开为

$$\begin{aligned}
(\tilde{m}_1 - \tilde{m}_2)^2 &= (\mathbf{w}^T\mathbf{m}_1 - \mathbf{w}^T\mathbf{m}_2)^2 \\
&= \mathbf{w}^T(\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^T\mathbf{w} \\
&= \mathbf{w}^T\mathbf{S}_B\mathbf{w}
\end{aligned} \tag{14}$$

其中

$$\mathbf{S}_B = (\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^T \tag{15}$$

故，得到准则函数 $J(\bullet)$ [3]

$$J(\mathbf{w}) = \frac{\mathbf{w}^T\mathbf{S}_W\mathbf{w}}{\mathbf{w}^T\mathbf{S}_B\mathbf{w}} \tag{16}$$

总类内离散程度矩阵 $\mathbf{S}_W$ 与全部样本的协方差矩阵成正比，且是对称半正定。当 $n > d$ 时， $\mathbf{S}_W$ 通常是非奇异的。而 $\mathbf{S}_B$ 是两个向量的外积，因此，其秩最多为 1。

容易证明，要使此时的准则函数 $J(\bullet)$ 最小化， $\mathbf{w}$ 需满足

$$\mathbf{S}_B\mathbf{w} = \lambda\mathbf{S}_W\mathbf{w} \tag{17}$$

这是一个广义特征值的问题，如果此时的 $\mathbf{S}_W$ 是非奇异的，我们就能得到通常的特征值问题

$$\mathbf{S}_W^{-1}\mathbf{S}_B\mathbf{w} = \lambda\mathbf{w} \tag{18}$$

[3]这是一个广义瑞利商，设 A 和 B 为复数域上的两个 n 阶方阵，给定一个 $0 \neq x \in \mathbb{C}^n$ 称

$$R(x) \equiv \frac{x^*Ax}{x^*Bx}$$

为矩阵对 A, B 的广义瑞利商，其中 x^* 表示 x 的共轭转置

而事实上我们并不需要求出矩阵 $\mathbf{S}_W^{-1}\mathbf{S}_B$ 的特征值和特征向量，注意到

$$\begin{aligned}\mathbf{S}_B\mathbf{w} &= (\mathbf{m}_1-\mathbf{m}_2)(\mathbf{m}_1-\mathbf{m}_2)^T\mathbf{w}\\ &= (\mathbf{m}_1-\mathbf{m}_2)(\mathbf{w}^T(\mathbf{m}_1-\mathbf{m}_2))^T\\ &= (\mathbf{m}_1-\mathbf{m}_2)(\tilde{m}_1-\tilde{m}_2)\end{aligned} \tag{19}$$

所以 $\mathbf{S}_B\mathbf{w}$ 与 $(\mathbf{m}_1-\mathbf{m}_2)$ 是同方向的，因为我们关心的是 $\mathbf{w}$ 的方向，而非大小，故从公式18得到

$$\mathbf{w}=\mathbf{S}_W^{-1}(\mathbf{m}_1-\mathbf{m}_2) \tag{20}$$

由此，我们已经完成了所有的理论推导，并且从准则函数 $J(\bullet)$ 得到了最佳的投影方向 $\mathbf{w}$。

接下来，我们将完成构建分类器 Ψ 的最后一步，判决边界：

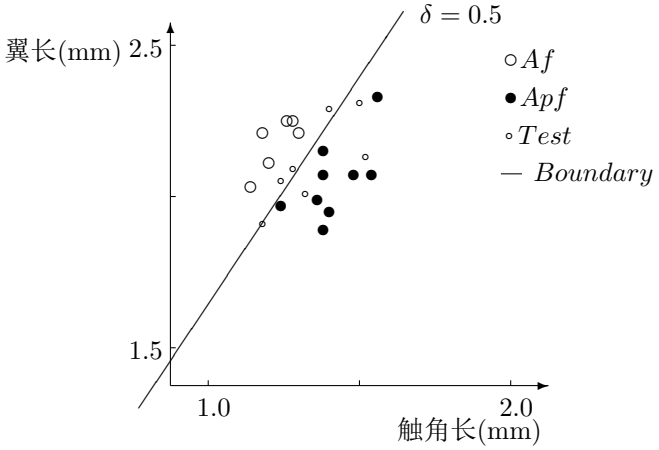
$$\mathbf{w}^T\left(\mathbf{x}-\mathbf{w}(\delta\tilde{m}_1+(1-\delta)\tilde{m}_2)\right)=0 \tag{21}$$

$\delta\in(0,1)$，为具体案例中的量化因子。

整个过程中的计算复杂度主要由计算总类内离散程度矩阵 $\mathbf{S}_W$ 和其逆矩阵所决定，其复杂度为 $O(d^2n)$。

3 具体运用

根据已建立的模型，我们来看看在具体案例中实施的效果，调用附录中按所建立的模型而编写的matlab代码，根据已知的训练样本集 χ 我们得到一条最佳的分界线，且能良好的区分测试集合 T 中的数据。



然而，在具体案例中，我们的 Af 是宝贵的传粉益虫， Apf 是某种疾病的载体。显而易见的是，当我们的分类器识别错一只 Apf 所带来的**成本风险**等要**数倍甚至十倍百倍** 识别错一只 Af，因此，我们倾向于要尽可能少的识别

错 Apf，因为 Apf 给我们带来的成本风险较高导致。由此，公式21中我们的量化因子 δ 的范围将缩小至 $\delta \in (0.5, 1)$。至此，我们可以综合考虑成本风险及分类效果两个因素来权衡 δ 的取值。

4 模型优化

4.1 现有问题

1. 无法应对训练样本中过多的低质量的样本数据。
2. 仅适用于二元分类问题。

4.2 解决方案

1. 针对样本数据中的过多低质量数据，采取的措施也唯有尽量的提高样本数据来源的可靠性，不然，我们得到的分类器 Ψ 也是无意义的。
2. 针对多元分类问题，需要在已建立的模型的基础上作一些相应的延伸，具体如下：

   对于 c 元分类问题，我们就需要 $c-1$ 个分类器 Ψ。也就是说，投影问题实际上是从 d 维空间向 $c-1$ 维空间作投影，并且已经假设 $d \geq c$。类别内的离散程度矩阵的推广也是明显的：

$$\mathbf{S}_W = \sum_{i=1}^{c} \mathbf{S}_i \tag{22}$$

   其中，就像以前一样，

$$\mathbf{S}_i = \sum_{\mathbf{x} \in \omega_i} (\mathbf{x} - \mathbf{m}_i)(\mathbf{x} - \mathbf{m}_i)^T \tag{23}$$

   和

$$\mathbf{m}_i = \frac{1}{n_i} \sum_{\mathbf{x} \in \omega_i} \mathbf{x} \tag{24}$$

   对 $\mathbf{S}_B$ 的推广并不那么显而易见。假设我们定义总体均值向量 $\mathbf{m}$ 和样本总体离散程度矩阵 $\mathbf{S}_T$ 为

$$\mathbf{m} = \frac{1}{n} \sum_{\mathbf{x} \in \chi} \mathbf{x} = \frac{1}{n} \sum_{i=1}^{c} n_i \mathbf{m}_i \tag{25}$$

$$\mathbf{S}_T = \sum_{\mathbf{x} \in \chi} (\mathbf{x} - \mathbf{m})(\mathbf{x} - \mathbf{m})^T \tag{26}$$

于是有

$$\begin{aligned}\mathbf{S}_T &= \sum_{i=1}^{c}\sum_{\mathbf{x}\in\omega_i}(\mathbf{x}-\mathbf{m}_i+\mathbf{m}_i-\mathbf{m})(\mathbf{x}-\mathbf{m}_i+\mathbf{m}_i-\mathbf{m})^T \\ &= \sum_{i=1}^{c}\sum_{\mathbf{x}\in\omega_i}(\mathbf{x}-\mathbf{m}_i)(\mathbf{x}-\mathbf{m}_i)^T+\sum_{i=1}^{c}\sum_{\mathbf{x}\in\omega_i}(\mathbf{m}_i-\mathbf{m})(\mathbf{m}_i-\mathbf{m})^T \\ &= \mathbf{S}_W+\sum_{i=1}^{c}n_i(\mathbf{m}_i-\mathbf{m})(\mathbf{m}_i-\mathbf{m})^T \end{aligned} \tag{27}$$

很自然，把上式右边的第二项定义为类别间样本离散程度矩阵 $\mathbf{S}_B$，因此就有

$$\mathbf{S}_B=\sum_{i=1}^{c}n_i(\mathbf{m}_i-\mathbf{m})(\mathbf{m}_i-\mathbf{m})^T \tag{28}$$

及

$$\mathbf{S}_T=\mathbf{S}_W+\mathbf{S}_B \tag{29}$$

从 d 维空间向 $c-1$ 维空间的投影是通过下列的 $c-1$ 个分类方程来进行的：

$$y_i=\mathbf{w}_i^T\mathbf{x}\qquad i=1,\ldots,c-1 \tag{30}$$

如果我们把 y_i 看成一个 $c-1$ 维向量 $\mathbf{y}$ 的分量，把 $\mathbf{w}_i$ 看成一个 $d*(c-1)$ 矩阵 $\mathbf{W}$ 的列向量，那么，公式30中的投影方程组就可以表达为简单的矩阵方程

$$\mathbf{y}=\mathbf{W}^T\mathbf{x} \tag{31}$$

对原始样本 $\mathbf{x}_1,\ldots,\mathbf{x}_n$ 进行投影后，得到了新的样本 $\mathbf{y}_1,\ldots,\mathbf{y}_n$。这些新的样本本身具有它们自己的均值向量和离散程度矩阵，这样，我们定义

$$\tilde{\mathbf{m}}_i=\frac{1}{n_i}\sum_{\mathbf{y}\in\omega_i}\mathbf{y} \tag{32}$$

$$\tilde{\mathbf{m}}=\frac{1}{n}\sum_{i=1}^{c}n_i\tilde{\mathbf{m}}_i \tag{33}$$

$$\tilde{\mathbf{S}}_W=\sum_{i=1}^{c}\sum_{\mathbf{y}\in\omega_i}(\mathbf{y}-\tilde{\mathbf{m}}_i)(\mathbf{y}-\tilde{\mathbf{m}}_i)^T \tag{34}$$

$$\tilde{\mathbf{S}}_B=\sum_{i=1}^{c}n_i(\tilde{\mathbf{m}}_i-\tilde{\mathbf{m}})(\tilde{\mathbf{m}}_i-\tilde{\mathbf{m}})^T \tag{35}$$

容易证明

$$\tilde{\mathbf{S}}_W=\mathbf{W}^T\mathbf{S}_W\mathbf{W} \tag{36}$$

$$\tilde{\mathbf{S}}_B=\mathbf{W}^T\mathbf{S}_B\mathbf{W} \tag{37}$$

上述各个方程说明了从高维空间向低维空间的投影过程，我们的目的是寻找一个变换矩阵 $\mathbf{W}$，能够在某种意义上，使得类内离散程度和类间离散程度的比值最小。使用这样的度量方式，我们得到准则函数如下：

$$J(\mathbf{W}) = \frac{|\ \tilde{\mathbf{S}}_W\ |}{|\ \tilde{\mathbf{S}}_B\ |} = \frac{|\ \mathbf{W}^T\mathbf{S}_W\mathbf{W}\ |}{|\ \mathbf{W}^T\mathbf{S}_B\mathbf{W}\ |} \tag{38}$$

求解使得 $J(\bullet)$ 最小化的矩阵 $\mathbf{W}$ 的过程需要上文同样的技巧，但最后的解的形式确实比较简洁的—最优矩阵 $\mathbf{W}$ 的列向量是下列等式中的最大特征值对应的特征向量：

$$\mathbf{S}_B\mathbf{W}_i = \lambda_i\mathbf{S}_W\mathbf{W}_i \tag{39}$$

如果 $\mathbf{S}_W$ 是非奇异的，那么，求解方程39就是一个普通的特征值问题，用求解特征多项式的根的方法来求解特征值：

$$|\ \mathbf{S}_B - \lambda_i\mathbf{S}_W\ | = 0 \tag{40}$$

然后，我们通过求解：

$$(\mathbf{S}_B - \lambda_i\mathbf{S}_W)\mathbf{w}_i = 0 \tag{41}$$

计算出 $\mathbf{w}_i$。

References

[1] Parlett,B.(1974). *The Rayleigh Quotient Iteration.* Mathematics of Computation.

[2] Richard O. Duda Peter E. Hart David G. Stork. *Pattern Classification*

[3] Donald E. Knuth. *LaTeX Referance*