PART4 贝叶斯决策

2020-05-25

河北师范大学 软件学院 要点：

1. 贝叶斯决策规则；
2. 正态分布概率密度函数形式、估计；
3. 朴素贝叶斯分类模型的学习与使用。

**连续的特征空间**

**离散的特征空间**

# 基本内容：

1. 掌握连续特征空间、离散特征空间两种情况下，基于最小错误率的贝叶斯分类决策规则；

最小错误率决策指将样本分错的概率最小，那么换句话说就是在对样本进行分类的时候分为正确类别的概率最大。也就是说，我们把使得后验概率p(wi|x)最大的wi作为x的判定类别。因为p(wi|x)最大即表示x属于wi的概率最大，也就是x被判错的概率最小。

1. 掌握连续特征空间、离散特征空间两种情况下，基于最小风险的贝叶斯分类决策规则；

最小错误率标准在某些决策中是不合理的。比如在癌细胞识别的决策过程中，我们对于不同错误的接受程度不同，因此不同错误的决策要乘上不同的权重，来衡量决策的风险，也就形成了最小风险决策。

1. 掌握正态分布概率密度函数：
2. 一维、多维连续特征空间，正态分布的类条件概率密度函数具体的形式？

**(2)**对于多类别分类：若多维连续特征空间，各个特征类条件独立，正态分布的类条件概率密度函数具体形式？具有什么特点？

1. 已知原始**d**维特征空间的随机向量**x**的期望向量、协方差矩阵，该随机向量经线性变换后的期望、协方差矩阵？

|  |
| --- |
| 随机向量***X*** *X1*,..., *Xd* ***T*** ***X*** ,***X***   换矩阵***A******a***1 ***a***2  ***ak***  *d*行*k*列  若 变   ***T*** 线性变换   ***Y A X***   ***Y***  ***AT X***  则  ***T***  ***Y***  ***A*** ***X A***  *y* ***a XT***    特别地：  ***y*** ***aT X***   2 ***T***  ***y*** ***a*** ***X a*** |

；

1. **(1)**一元、多元正态分布的概率密度函数的参数的最大似然估计结果？

|  |
| --- |
| 参数 、 2的最大似然估计   *\** 1 ***N*** *x =xi*    ***N****i*1    ***N***  2*\**  ***N***1 ***i***1 *x -i* *\** 2   |
| **最大似然参数估计**   *\** 1 ***N*** *x =xi*  **期望**    ***N****i*1    ***N***  **协方差矩阵** *\**  ***N***1 ***i***1 *x -i*  *\** *x -i \** ***T***   | | |

**(2)**特征独立时，多元正态分布概率密度函数的参数的最大似然估计结果？

|  |
| --- |
| 随机变量 的边缘密度***X*** ***i***  1  1  *x* ***i***  ***i*** 2  ***pX*** ***i*** (*x* ***i*** ;  ***i*** ,  ***i*** 2)  2 exp 2  ***i***     ***i***      随机向量 的概率密度函数：***X*** *p****X*** (***x***;，) ***pX*** ***i*** (*x* ***i*** ;  ***i*** ,  ***i*** 2)  ***i***1  **边缘密度*p***  ***i*** (*x* ***i*** ;  ***i*** ,  ***i*** 2)**最大似然参数估计**  ***X***    ***i*** \* 1 ***N***  ***i***  **期望**    *N* ***k***1 ***x****k*     ***N*** \* 2  **方差**  ***i*** 2\*  ***N***1 ***k***1*xk* ***i*** *-* ***i***    **期望向量** \*   1 \* ,  2 \* ,..., *d* \****T***  **协方差矩阵** \*  *diag*  1 2\* ,  2 2\* ,..., ***d*** 2\*   |

1. 给定连续特征空间**/**离散特征空间，已知类别标记的训练样本集，能够基于该样本集，实现两步法的**(**朴素**)**贝叶斯决策。

|  |
| --- |
| **基于样本的两步贝叶斯决策**   1 **利用有限规模训练样本*****x j***,y ***j*** , ***j***1,...,***N*****,**  先验概率 ***P(ω )***  **设计贝叶斯分类器. 估计** ***i***  条件概率密度 ***p(x|ωi )***或条件概率***P(x|ωi )***  ***i***1,2,...,***c***  **利用估计的*P*** (***i***)、***p x***( |***i***)或***P x*** ( |***i***)  2  **对未知样本*x*进行判决**   |

# 练习：

**1.** 对于二维连续特征空间的两类别分类问题，训练样本集𝑿 𝒙 , y , 𝑖

1, … , 𝑁 ，按照𝑝 𝒙 Pω p 𝒙|ω P ω p 𝒙|ω 独立抽取，其中𝒙

𝑥 ，𝑥 ∈ 𝑅 , 𝒚 ∈ ω , ω . 并且，训练集𝑿中：来自第一类ω 及第二类

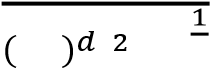
ω 的样本集分别为 𝑿 及𝑿 ，两类样本数目分别𝑁 及𝑁 .

请按要求完成如下工作：

(1)若两个类别的类条件概率密度函数p 𝒙|ω 、p 𝒙|ω 均为正态分布，并且各自期望向量分别为 𝝁 ，𝝁 ，协方差矩阵分别为 ∑ 、∑ ，写出

p 𝒙|ω 、p 𝒙|ω 的具体表达式。

参考答案： p 𝒙|ω exp 𝒙 𝝁 ∑ 𝒙 𝝁 i 1,2

⁄ |∑ |

(2)若基于样本集𝑿 及𝑿 ，采用最大似然法估计上述概率密度函数的参数𝝁 ，𝝁

及∑ 、∑ ，请直接写出参数的估计结果。

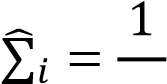
参考答案： 以ω 类的样本集𝑿 估计p 𝒙|ω 的参数𝝁 以及∑ , 𝑖 1,2 其最大似然估计结果为：

1

𝝁 𝒙

𝑁

𝒙∈𝑿

 𝒙 𝝁 𝒙 𝝁 𝑻

𝑁

𝒙∈𝑿

(3)对于每个类别，若二维连续特征空间两种特征相互独立，基于样本集𝑿𝟏及𝑿𝟐，

请直接写出上述概率密度函数的参数𝝁𝟏，𝝁𝟐及∑𝟏、∑𝟐的最大似然估计结果。

|  |  |  |
| --- | --- | --- |
| 参考答案： |  |  |
| 𝝁 𝜇̂ 𝜇̂ |  | 1 1  𝑥 𝑥  𝑁 𝑁 |

𝒙∈𝑿 𝒙∈𝑿

𝟐

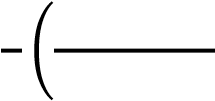
σ 0 ∑𝒙∈𝑿 𝑥 𝜇̂ 0

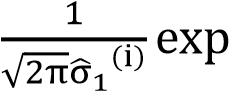
∑ = 𝟐

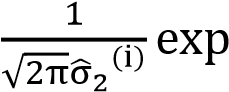
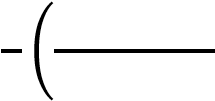
0 σ 0 ∑𝒙∈𝑿 𝑥 𝜇̂

𝑖 1,2

(4)若𝒙 𝑥 ，𝑥 ∈ 𝑅 ，请基于问题(3)的估计结果，分别写出p 𝑥 |ω 及 p 𝑥 |ω 的具体表达式，𝑖 1,2 参考答案：

𝒙 𝒊 𝝁

p 𝑥 |ω  i=1,2

p 𝑥 |ω  𝒙 𝒊 𝝁 i=1,2

(5)请采用上述给定的样本集𝑿 𝑿 ∪ 𝑿 ，分别估计P ω 以及 P ω 。

# 参考答案：

𝑵

𝑃 ω , *i=1,2*

𝑵

(6)请基于上述估计结果，采用最小错误率的朴素贝叶斯分类模型对特征空间任

𝐓 意观测样本𝒙 𝒙 𝟏 ，𝒙 𝟐 ∈ 𝑹𝟐分类。参考答案：对于任意观测样本𝒙 𝑥 ，𝑥 ∈ 𝑅

首先，计算：P ω p 𝒙 𝟏 |ω p 𝒙 𝟐 |ω ， 以及P ω p 𝒙 𝟏 |ω p 𝒙 𝟐 |ω

若P ω p 𝒙 𝟏 |ω p 𝒙 𝟐 |ω Pω p 𝒙 𝟏 |ω p 𝒙 𝟐 |ω ,则将观测样本𝑥判断为ω 类；

若P ω p 𝒙 𝟏 |ω p 𝒙 𝟐 |ω Pω p 𝒙 𝟏 |ω p 𝒙 𝟐 |ω ,则将观测样本𝑥判断为ω 类；否则，拒绝决策𝑥的类别，或结合具体问题，进行决策。

1. 对于二维连续特征空间的两类别分类问题，训练样本集𝑿 𝒙 , y , 𝑖

1, … , 𝑁 按照𝑝 𝒙 Pω p 𝒙|ω P ω p 𝒙|ω 独立抽取，其中𝒙 𝑥 ，

𝑥 ∈ 𝑅 , 𝒚 ∈ ω , ω . 并且，训练集𝑿中：来自第一类ω 及第二类ω 的样本

集分别为𝑿 及𝑿 ，两类样本数目分别𝑁 及𝑁 .请按照要求完成如下工作：

(1)基于训练样本集𝑿设计一个最小错误率高斯朴素贝叶斯分类模型；

(2)基于上述模型，写出面向特征空间任意观测样本𝒙 𝑥 ，𝑥 ∈ 𝑅 分类的决策规则。

1. 类似地，将问题1及2情况扩展至四维连续特征空间的鸢尾花数据集的分类问题。进行求解。
2. 下表所示为来自二维离散特征空间关于两种类别的训练样本.其中：样本特征描述为𝑿 𝑋 ，𝑋 ，并且𝑋 ∈ 1,2 ，𝑋 ∈ 𝑆, 𝑀, 𝐿 ；类别标记𝒀 ∈

𝟏，𝟏 **.**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **1** | **2** | **3** | **4** | **5** |
| 特征𝐗 𝟏 | **1** | **1** | **2** | **2** | **1** |
| 特征𝐗 𝟐 | **S** | **M** | **S** | **M** | **L** |
| 𝐘 | **-1** | **-1** | **-1** | **-1** | **1** |

请结合该表，采用LAPLACE平滑方式，估计朴素贝叶斯分类模型的有关概率信息，并进行给定观测样本的类别决策.

1. 两个类别的先验概率𝑃 𝑌 1 , 𝑃 𝑌 1 ；
2. 𝑃 𝑋 2|𝑌 1 , 𝑃 𝑋 2|𝑌 1
3. 𝑃 𝑋 L|𝑌 1 , 𝑃 𝑋 L|𝑌 1
4. 𝑃 𝑋 2，𝑋 L|𝑌 1
5. 𝑃 𝑋 2，𝑋 L|𝑌 1
6. 对观测样本𝑿 2，𝐿 进行类别决策.

解： 1 先验概率 ***P****Y*1 ***=*** 1 4 5 ***P****Y*1 ***=*** 1 1 2

2 1 5 7  2 1 5 7 

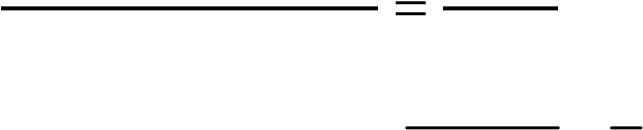
𝜆 ∑ 𝐼 𝑌 1 1 4

𝑃 𝑌 1 

𝜆 ∑ 𝐼 𝑌 1 1 1

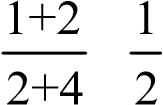
𝑃 𝑌 1

𝐶𝜆 𝑁 2 5

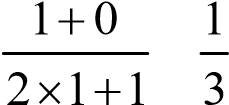
所以：先验概率 ***P****Y*1 ***=*** 1 4 5 ***P****Y*1 ***=*** 1 1 2

2 1 5 7  2 1 5 7 

+ *X*  1  2并且*Y* 1

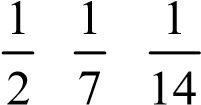
 2 由于***P****X*  1  2 |*Y* 1 ***=*** *i*1 *N* = = *S* 1 +*Y* 1

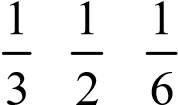
*i*1

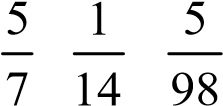
同理： ***P****X*  1  2 |*Y* 1 ***=*** 

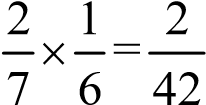
 3 ***P****X*  2 *L Y*| 1 ***=*** 1 0  1 ***P****X*  2 *L Y*| 1 ***=*** 11  1

3 1 4 7 3 1 1 2

 4 ***P****X*  1  2, *X*  2 *L Y*| 1 ***= P****X*  1  2 |*Y* 1***P****X*  2 *L Y*| 1 =  =

 5 ***P****X*  1  2, *X*  2 *L Y*| 1 ***= P****X*  1  2 |*Y* 1***P****X*  2 *L Y*| 1 =  =

  6 ***P*** *Y* 1***P****X*  1  2, *X*  2 *L Y*| 1 =  =

***P****Y* 1***P****X*  1  2, *X*  2 *L Y*| 1 =  所以***P****Y* 1 ***P****X*  1  2, *X*  2 *L Y*| 1 ***< P****Y* 1***P****X*  1  2, *X*  2 *L Y*| 1

1 2 *T T*  观测样本***X*** *X X* ,    2, *L* 决策结果为*Y* 1

5. 给定三个类别分类的训练集。其中：样本特征描述为𝑿 𝑋 ，𝑋 ，并

且𝑋 ∈ 1,2,3 ，𝑋 , ∈ 𝑆, 𝑀, 𝐿, 𝑋𝐿 ；类别标记𝒀 ∈ 𝟏, 𝟐, 𝟑 **.**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** |
| 特征𝐗 𝟏 | **1** | **1** | **1** | **2** | **1** | **2** | **2** | **1** | **2** |
| 特征𝐗 𝟐 | **S** | **M** | **XL** | **XL** | **S** | **S** | **M** | **L** | **L** |
| 𝐘 | **2** | **2** | **1** | **3** | **2** | **2** | **2** | **1** | **1** |

完成如下小题：

1. 三个类别的先验概率𝑃 𝑌 1 , 𝑃 𝑌 2 ，𝑃 𝑌 3

1. 𝑃 𝑋 2|𝑌 1 , 𝑃 𝑋 2|𝑌 2 ，𝑃 𝑋 2|𝑌 3

1. 𝑃 𝑋 L|𝑌 1 , 𝑃 𝑋 L|𝑌 2 ，𝑃 𝑋 L|𝑌 3

1. 𝑃 𝑋 2，𝑋 L|𝑌 1

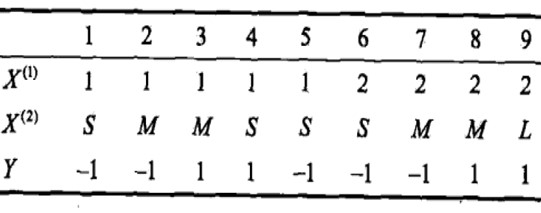
1. 𝑃 𝑋 2，𝑋 L|𝑌 2

## 6 𝑃 𝑋 2，𝑋 L|𝑌 3

7 对观测样本𝑿 2，𝐿 进行类别决策.

6.下表所示为来自二维离散特征空间关于两种类别的训练样本.其中：样本特征描述为𝑿 𝑋 ，𝑋 ，并且𝑋 ∈ 1,2 ，𝑋 , ∈ 𝑆, 𝑀, 𝐿 ；类别标记𝒀 ∈

𝟏，𝟏 .



请基于该表，采用LAPLACE平滑方式，估计朴素贝叶斯分类模型的如下信息：

|  |  |  |  |
| --- | --- | --- | --- |
| (1)两个类别的先验概率𝑃 𝑌 | | 1 , 𝑃 𝑌 | 1 ； |
| 2 𝑃 𝑋 1|𝑌 | 1 , 𝑃 𝑋 | 1|𝑌 | 1 |
| 3 𝑃 𝑋 L|𝑌 | 1 , 𝑃 𝑋 | L|𝑌 | 1 |
| 4 𝑃 𝑋 1，𝑋 | L|𝑌 | 1 |  |
| 5 𝑃 𝑋 1，𝑋 | L|𝑌 | 1 |  |
| 7 对观测样本𝑿 | 1，𝐿 进行类别决策 | | |