模式识别作业3

张蔚桐 2015011493 自55

2017 年 3 月 24 日

1

考察采用$\lambda(\hat{\theta},\theta)=(\hat{\theta}-\theta)^2$的期望风险可得

$$R(\hat{\theta}|\mathbf{X})=\int_{\theta}(\hat{\theta}-\theta)^2P(\theta|\mathbf{X})\mathrm{d}\theta \tag{1}$$

对1式求导可得

$$\frac{\mathrm{dR}(\hat{\theta}|\mathbf{X})}{\mathrm{d}\hat{\theta}}=\int_{\theta}2(\hat{\theta}-\theta)P(\theta|\mathbf{X})\mathrm{d}\theta \tag{2}$$

期望风险最小化可以得到2式为0，因此可以得到

$$\int_{\theta}\theta P(\theta|\mathbf{X})\mathrm{d}\theta=\hat{\theta}\int_{\theta}\mathrm{P}(\theta|\mathbf{X})\mathrm{d}\theta \tag{3}$$

4等式右侧系数为归一化系数，因此可直接得到结论

$$\hat{\theta}=\int_{\theta}\theta P(\theta|\mathbf{X})\mathrm{d}\theta=\mathrm{E}(\theta|\mathbf{X}) \tag{4}$$

2

2.1

如图1所示，是样本集分别为10，100,1000时的三次独立测试的MLE参数学习之后的曲线。其中，虚线为原分布的概率密度函数，可以看出学习情况还是比较理想的

2.2

如图2所示，是样本集分别为10，100,1000时的三次独立测试的MLE参数学习之后的曲线。其中，虚线为原分布的概率密度函数，可以看出参数学习结果并不很稳定。

2.3

先讨论学习参数稳定性的问题：如果样本量相同，可以知道选择合适的模型会学习出比较稳定的参数，而不合适的模型学习的参数抖动比较大。在相同的模型下，样本量越大，模型的学习参数越稳定。

再讨论学习得到的分布曲线意义的问题:选用不合适的模型，学习之后的概率密度分布曲线不会趋近于原来的曲线（根据本题提示），这个时候学习的概率密度曲线对之后的判定作用微乎其微。

因此样本量决定学习结果的精度，模型决定学习结果的正确与否

3

3.1

采用高斯窗对样本进行非参数估计之后得到的两种情况下样本的概率密度函数为图3所示

3.2

采用最小错误率对测试样本进行预测的正确率在97%左右

3.3

采用最小风险的贝叶斯决策可以得到，对恶性肿瘤的预测正确率在测试集内部为100%，而对良性肿瘤预测的正确率在92%左右，平均的预测正确率在94%左右

3.4

多次测试可以得到，采用最小错误率的贝叶斯决策正确率普遍大于采用最小风险的贝叶斯决策。采用最小风险的贝叶斯决策通过牺牲风险较小的决策的正确性，提高了风险较大的决策的正确性。

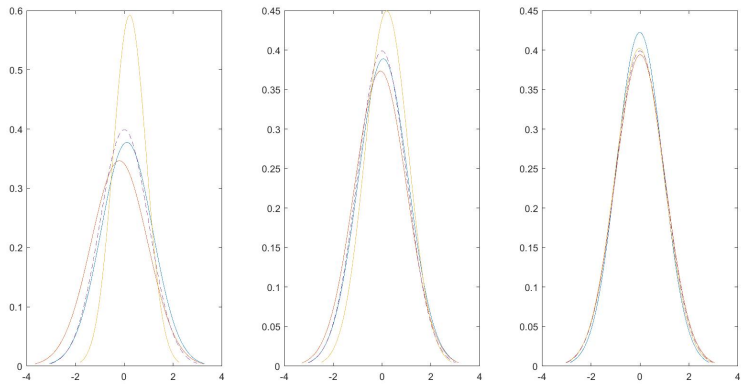

图 1: 用正态分布学习正态分布

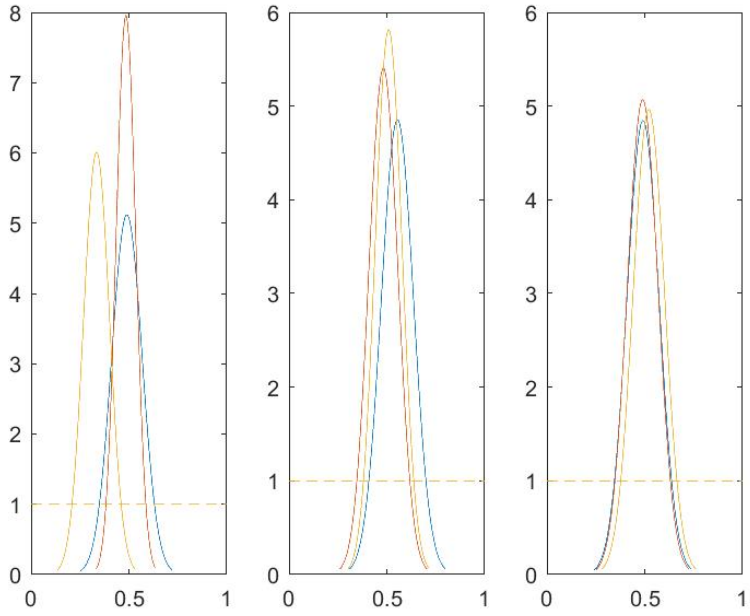

图 2: 用正态分布学习均匀分布

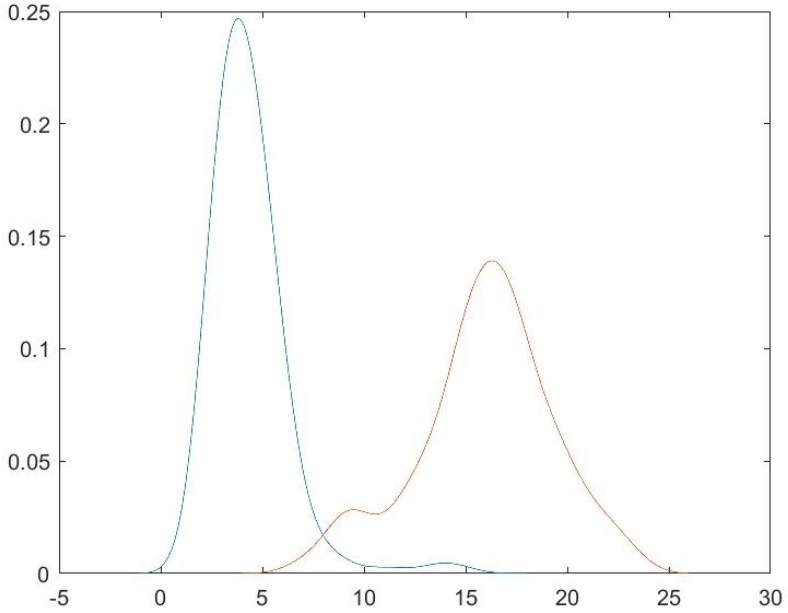

图 3: 采用高斯窗的非参数估计

其中MATLAB代码已经在附件中包含，平台采用的是MATLAB R2015b，执行时，请采用以下顺序

```
% load data, please do first
dataload;

% Fisher method
mainFisher;

% Check for Fisher method
% Please run after Fisher method
checkFisher;%Min Err
checkFisher_1;%Min Risk
```