贝叶斯决策与（非）参数估计

1. 贝叶斯估计[20%]

在贝叶斯估计中，给定有限样本集合 $\mathbf{X}=\{\mathbf{x}_1,\mathbf{x}_2,\cdots,\mathbf{x}_n\}$ 情况下，贝叶斯估计量可通过最小化期望风险获得，即

$$\theta^*=\arg\min_{\hat{\theta}} R(\hat{\theta}\mid\mathbf{X})=\int_\theta \lambda(\hat{\theta},\theta)p(\theta\mid\mathbf{X})d\theta$$

其中，$\lambda(\hat{\theta},\theta)$ 是定义的损失函数，当损失函数为平方误差损失函数，即

$\lambda(\hat{\theta},\theta)=(\hat{\theta}-\theta)^2$ 时，请证明，在给定样本集 $\mathbf{X}$ 下，θ 的贝叶斯估计量是

$\theta^*=E[\theta\mid\mathbf{X}]=\int_\theta \theta p(\theta\mid\mathbf{X})d\theta$。参见张学工老师模式识别（第三版）49 页。

2. 最大似然估计[40%]（提交作业请提供源代码）

(1) 请从正态分布$N(0,1)$中分别抽取 10，100，1000 个样本，利用最大似然估计正态分布假设下的模型参数，分别重复三次实验，将同一抽样量下的三次重复实验估计的概率密度分布曲线绘制在一张图片内，并与正态分布$N(0,1)$的概率密度分布曲线比较。

(2) 请从均匀分布$U(0,1)$中抽取 100 个样本，利用最大似然估计正态分布假设下的模型参数，绘制出估计得到的概率密度分布曲线图，并与均匀分布$U(0,1)$的概率密度分布曲线图比较。

(3) 通过上述实验，讨论模型的选择、样本量对参数估计的影响。

3. 非参数估计与贝叶斯决策[40%]（提交作业请提供源代码）

在第二次作业中，大家已经通过 Fisher 线性判别将多维特征投影到一维，请利用投影得到的一维特征完成下列要求，其中训练集依然随机选取其中 70%的数据，测试集选取剩余 30%的数据。

(1) 利用高斯窗的方法对训练集中的良性肿瘤和恶性肿瘤样本分布进行非参数估计。推荐使用$\sigma=1$（前提条件是，Fisher 投影方向 w^* 先标准化（模为 1），然后对原始特

征投影）。感兴趣者可以尝试不同σ的取值。

(2) 利用(1)中非参数估计的概率密度，请使用最小错误率的贝叶斯决策对测试集样本进行预测，给出测试集的错误率；

(3) 在同一测试集、训练集中，利用(1)中非参数估计的概率密度，请使用最小风险的贝叶斯决策对测试集样本进行预测，其中，将良性预测为恶性的风险为 1，将恶性预测为良性的风险为 10。

(4) 请随机划分训练集和测试集样本三次，重复上述步骤，比较最小风险下的预测结果与最小错误率下的预测结果有什么不同。