一、判断题

（1）极大似然估计是无偏估计且在所有的无偏估计中方差最小，所以极大似然估计的风险最小。

（2）回归函数 A 和 B，如果 A 比 B 更简单，则 A 几乎一定会比 B 在测试集上表现更好。

（3）全局线性回归需要利用全部样本点来预测新输入的对应输出值，而局部线性回归只需利用查询点附近的样本来预测输出值。 所以全局线性回归比局部线性回归计算代价更高。

（4） Boosting 的一个优点是不会过拟合。

（5）在回归分析中，最佳子集选择可以做特征选择，当特征数目较多时计算量大；岭回归和 Lasso 模型计算量小，且 Lasso也可以实现特征选择。

（6）梯度下降有时会陷于局部极小值，但 EM 算法不会。

（7）支持向量机是判别模型。 T

（8） ICA 方法对于高斯分布的数据也有效。 F

（9）回归问题属于非监督学习的一种方法。 F

（10）聚类算法中不需要给出标签 y。T

二、考虑一个二分类器问题（ Y 为 1 或 0），每个训练样本 X 有两个特征 X1、X2（0 或 1）。给出 P（Y=0）=P（Y=1） =0.5，条件概率如下表：

$P(X_1\|Y)$	$X_1=0$	$X_1=1$
$Y=0$	0.7	0.3
$Y=1$	0.2	0.8

$P(X_2\|Y)$	$X_2=0$	$X_2=1$
$Y=0$	0.9	0.1
$Y=1$	0.5	0.5

分类器预测的结果错误的概率为期望错误率， Y 是样本类别的实际值， Y'（X1，X2）为样本类别的预测值，那么期望错误率为：

$$P_{\mathcal{D}}\left(Y=1-\hat{Y}(X_1,X_2)\right)=\sum_{X_1=0}^{1}\sum_{X_2=0}^{1}P_{\mathcal{D}}\left(X_1,X_2,Y=1-\hat{Y}(X_1,X_2)\right).$$

（1）给出 X1，X2 的所有可能值，使用贝叶斯分类器预测结果，填写下表：

X1	X2	P（X1,X2,Y=0 ）	P（X1,X2,Y=1 ）	Y (X1,X2)
0	0			
0	1			
1	0			
1	1			

（2）计算给定特征（ X1，X2）预测 Y 的期望错误率，假设贝叶斯分类器从无限的训练样本中学习所得。

（3）下面哪个有更小的期望错误率？

a、仅仅给出 X1，采用贝叶斯分类器预测 Y。

b、仅仅给出 X2，采用贝叶斯分类器预测 Y。

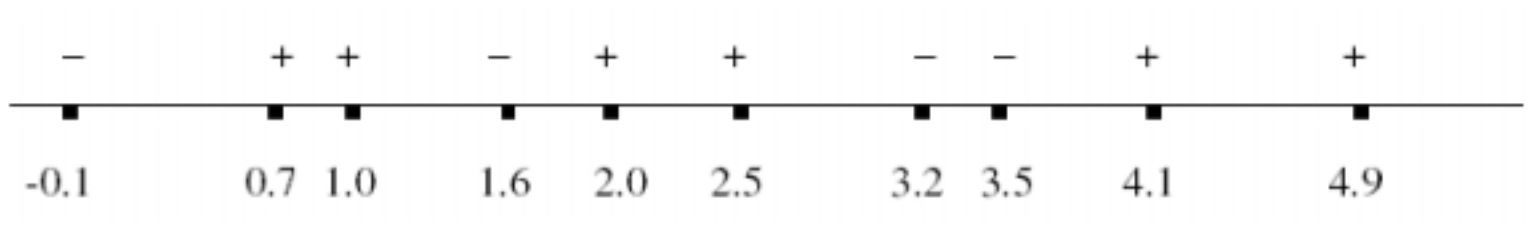
（4）给出一个新的特征 X3，X3 的与 X2 保持完全相同，现在计算给定（ X1，X2，X3）采用贝叶斯分类器预测 Y 的期望错误率，假设分类器从无限的训练数据中学习所得。

（5）使用贝叶斯分类器会产生什么问题，为什么？

三、交叉验证

1、4. 给定如下数据集， 其中 X 为输入变量， Y 为输出变量。 假设考虑采用 k-NN 算法

对 x 对应的 y 进行预测，其中距离度量采用不加权的欧氏距离。 （12 分）



X	Y
-0.1	-
0.7	+
1.0	+
1.6	-
2.0	+
2.5	+
3.2	-
3.5	-
4.1	+
4.9	+

（1）算法 **1-NN** 的训练误差的是多少？（用分类错误的样本数目表示即可，下同）

（2）算法 **3-NN** 的训练误差是多少？

（3）算法 **1-NN** 的 LOOCV （留一交叉验证）估计误差是多少？

（4）算法 **3-NN** 的 LOOCV （留一交叉验证）估计误差是多少？

四、用最大似然估计的方法估计高斯分布的均值和方差，并指出其局限性。

五、随着信息化的发展， 大数据的时代已经到来。 海量的文本、 图像、视频数据存在于互联网上，请结合自己的科研背景和兴趣，探讨机器学习方法如何在大数据分析、处理中应用。 （20 分）