

姓名\_\_\_\_\_学号\_\_\_\_\_成绩\_\_\_\_\_

一、基础题（共 36 分）

1、请描述极大似然估计 MLE 和最大后验估计 MAP 之间的区别。请解释为什么 MLE 比 MAP 更容易过拟合。（10 分）

2、在年度百花奖评奖揭晓之前，一位教授问 80 个电影系的学生，谁将分别获得 8 个奖项（如最佳导演、最佳男女主角等）。评奖结果揭晓后，该教授计算每个学生的猜中率，同时也计算了所有 80 个学生投票的结果。他发现所有人投票结果几乎比任何一个学生的结果正确率都高。这种提高是偶然的吗？请解释原因。（10 分）

3、假设给定如右数据集，其中  $A$ 、 $B$ 、 $C$  为二值随机变量， $y$  为待预测的二值变量。

- (a) 对一个新的输入  $A=0, B=0, C=1$ ，朴素贝叶斯分类器将会怎样预测  $y$ ? （10 分）
- (b) 假设你知道在给定类别的情况下  $A$ 、 $B$ 、 $C$  是独立的随机变量，那么其他分类器（如 Logistic 回归、SVM 分类器等）会比朴素贝叶斯分类器表现更好吗？为什么？（注意：与上面给的数据集没有关系。）（6 分）

| A | B | C | y |
|---|---|---|---|
| 0 | 0 | 1 | 0 |
| 0 | 1 | 0 | 0 |
| 1 | 1 | 0 | 0 |
| 0 | 0 | 1 | 1 |
| 1 | 1 | 1 | 1 |
| 1 | 0 | 0 | 1 |
| 1 | 1 | 0 | 1 |

二、回归问题。（共 24 分）

现有  $N$  个训练样本的数据集  $D = \{(x_i, y_i)\}_{i=1}^N$ ，其中  $x_i, y_i$  为实数。

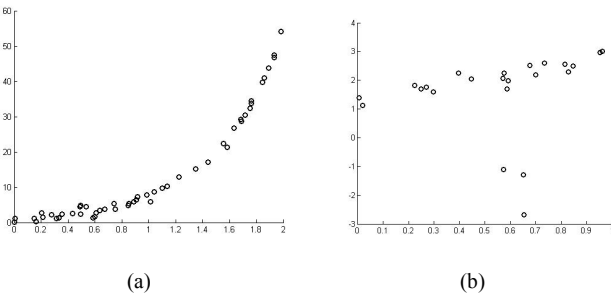
1. 我们首先用线性回归拟合数据。为了测试我们的线性回归模型，我们随机选择一些样本作为训练样本，剩余样本作为测试样本。现在我们慢慢增加训练样本的数目，那么随着训练样本数目的增加，平均训练误差和平均测试误差将会如何变化？为什么？（6 分）

平均训练误差：A、增加          B、减小

平均测试误差：A、增加          B、减小

2. 给定如下图(a)所示数据。粗略看来这些数据不适合用线性回归模型表示。因此我们采用如下模型： $y_i = \exp(wx_i) + \varepsilon_i$ ，其中  $\varepsilon_i \sim N(0,1)$ 。假设我们采用极大似然估计  $w$ ，请给出 log 似然函数并给出  $w$  的估计。（8 分）

3. 给定如下图(b)所示的数据。从图中我们可以看出该数据集有一些噪声，请设计一个对噪声鲁棒的线性回归模型，并简要分析该模型为什么能对噪声鲁棒。（10 分）



三、SVM 分类。（第 1~5 题各 4 分，第 6 题 5 分，共 25 分）

下图为采用不同核函数或不同的松弛因子得到的 SVM 决策边界。但粗心的实验者忘记记录每个图形对应的模型和参数了。请你帮忙给下面每个模型标出正确的图形。

$$1、\min\left(\frac{1}{2}\|\mathbf{w}\|^2 + C\frac{1}{2}\sum_{i=1}^N\xi_i\right), \text{ s.t.}$$

$$\xi_i \geq 0, y_i(\mathbf{w}^T \mathbf{x} + w_0) \geq 1 - \xi_i, i = 1, \dots, N,$$

其中  $C = 0.1$ 。

$$2、\min\left(\frac{1}{2}\|\mathbf{w}\|^2 + C\frac{1}{2}\sum_{i=1}^N\xi_i\right), \text{ s.t.}$$

$$\xi_i \geq 0, y_i(\mathbf{w}^T \mathbf{x} + w_0) \geq 1 - \xi_i, i = 1, \dots, N,$$

其中  $C = 1$ 。

$$3、\max\left(\sum_{i=1}^N\alpha_i - \frac{1}{2}\sum_{i=1}^N\sum_{j=1}^N\alpha_i\alpha_jy_iy_jk(\mathbf{x}_i, \mathbf{x}_j)\right)$$

$$\text{s.t. } \alpha_i \geq 0, i = 1, \dots, N, \quad \sum_{i=1}^N\alpha_iy_i = 0$$

$$\text{其中 } k(\mathbf{x}, \mathbf{x}') = \mathbf{x}^T \mathbf{x}' + (\mathbf{x}^T \mathbf{x}')^2。$$

$$4、\max\left(\sum_{i=1}^N\alpha_i - \frac{1}{2}\sum_{i=1}^N\sum_{j=1}^N\alpha_i\alpha_jy_iy_jk(\mathbf{x}_i, \mathbf{x}_j)\right)$$

$$\text{s.t. } \alpha_i \geq 0, i = 1, \dots, N, \quad \sum_{i=1}^N\alpha_iy_i = 0$$

$$\text{其中 } k(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{1}{2}\|\mathbf{x} - \mathbf{x}'\|^2\right)。$$

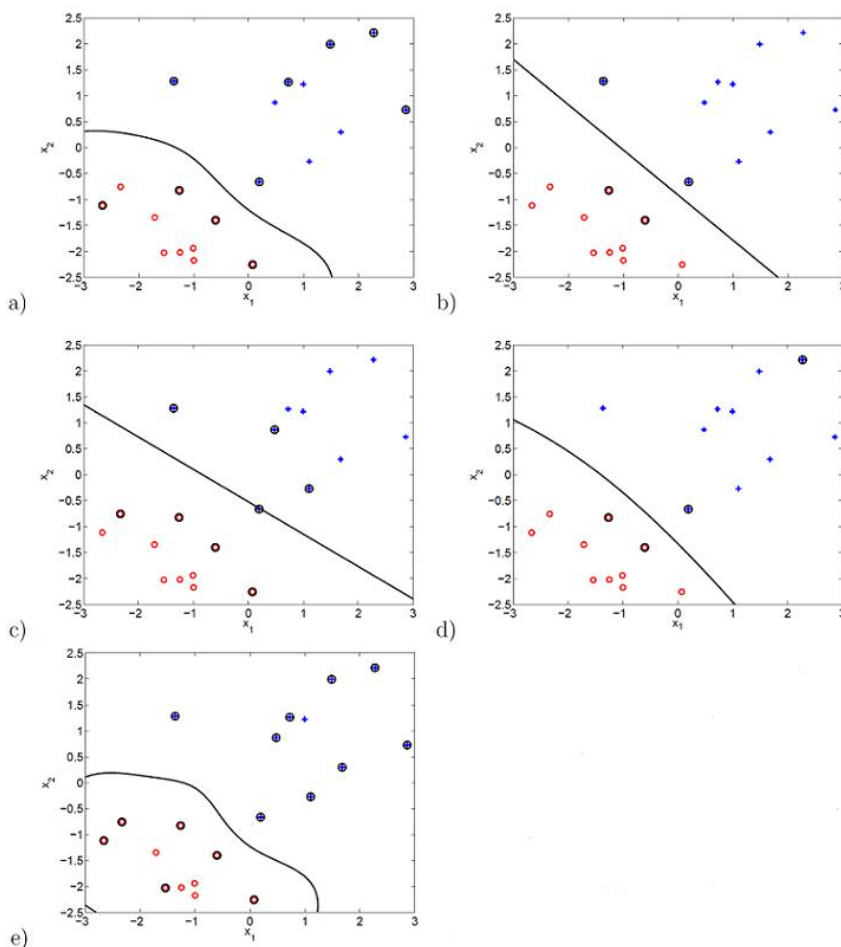
$$5、\max\left(\sum_{i=1}^N\alpha_i - \frac{1}{2}\sum_{i=1}^N\sum_{j=1}^N\alpha_i\alpha_jy_iy_jk(\mathbf{x}_i, \mathbf{x}_j)\right)$$

$$\text{s.t. } \alpha_i \geq 0, i = 1, \dots, N, \quad \sum_{i=1}^N\alpha_iy_i = 0$$

$$\text{其中 } k(\mathbf{x}, \mathbf{x}') = \exp(-\|\mathbf{x} - \mathbf{x}'\|^2)。$$

$$6、\text{考虑带松弛因子的线性 SVM 分类器: } \min\left(\frac{1}{2}\|\mathbf{w}\|^2 + C\frac{1}{2}\sum_{i=1}^N\xi_i\right), \text{ s.t. } \xi_i \geq 0, y_i(\mathbf{w}^T \mathbf{x} + w_0) \geq 1 - \xi_i, i = 1, \dots, N,$$

下面有一些关于某些变量随参数  $C$  的增大而变化的表述。如果表述总是成立，标示“是”；如果表述总是不成立，标示“否”；如果表述的正确性取决于  $C$  增大的具体情况，标示“不一定”。



- (1)  $w_0$  不会增大
- (2)  $\|\hat{\mathbf{w}}\|$  增大
- (3)  $\|\hat{\mathbf{w}}\|$  不会减小
- (4) 会有更多的训练样本被分错
- (5) 间隔(Margin)不会增大

四、一个初学机器学习的朋友对房价进行预测。他在一个  $N=1000$  个房价数据的数据集上匹配了一个有 533 个参数的模型，该模型能解释数据集上 99% 的变化。

- 1、请问该模型能很好地预测来年的房价吗？简单解释原因。（5 分）
- 2、如果上述模型不能很好预测新的房价，请你设计一个合适的模型，给出模型的参数估计，并解释你的模型为什么是合理的。（10 分）