中国科学院大学 2012 年《机器学习》试卷及其答案

任课教师：卿来云

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）

MLE：取似然函数最大时的参数值为该参数的估计值，y_{mle}=argmax[p(x|y)]；MAP：取后验函数（似然与先验之积）最大时的参数值为该参数的估计值，y_{map}=argmax[p(x|y)p(y)]。因为 MLE 只考虑训练数据拟合程度没有考虑先验知识，把错误点也加入模型中，导致过拟合。

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）

设 x 为第 i 个学生的猜中率（要么 0 要么 1）x~Ber(θ),E(x)=θ,V(x)=θ(1-θ)
mean(x)~N(θ,θ(1-θ)/N),E(mean(x))=θ,V(mean(x))=θ(1-θ)/N<V(x)

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	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

(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）

y~Ber(θ) p(y=0)=3/7,p(y=1)=4/7
p(y=0|A=0B=0C=1)∝p(y=0)*p(A=0|y=0)*p(B=0|y=0)*p(C=1|y=0)=3/7*2/3*1/3*1/3=2/63
p(y=1|A=0B=0C=1)∝p(y=1)*p(A=0|y=1)*p(B=0|y=1)*p(C=1|y=1)=4/7*1/4*2/4*2/4=1/28，因此属于 y=1 类

(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 Logstic 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）

不会。因为已知独立同分布的前提下 NBC 只用 3 个参数，不用 NBC 则需要 2^3-1=7 个参数。若不独立，则其他基于数据本身的判别式分类器效果较好。

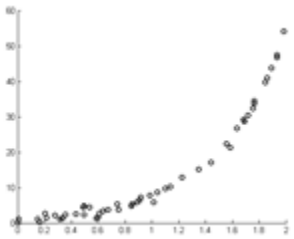
二、回归问题。（共 24 分）

现有 N 个训练样本的数据集 D={(xi,yi)}，其中 xi,yi 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）

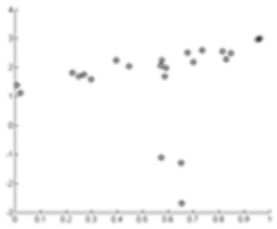
平均训练误差：A、增加　　B、减小
平均测试误差：A、增加　　B、减小
因为当训练样本增多时，模型参数发生改变以拟合新增的样本，因而使得模型原先的拟合程度下降，平均训练误差增加；而训练样本增多，模型越接近真实的分布，因而使得平均测试误差减小。

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）



p(yi|w,xi)~N(exp(wxi),1)
$L(w)=\log p(y|w,x)=-0.5*\Sigma(yi-\exp(wxi))^2+C$
令 $g(w)=\Sigma[(yi-\exp(wxi))*\exp(wxi)*xi]=0$ 求得 w

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）



如图离群点较多（heavy tail），使用鲁棒线性回归模型：$y=w^Tx+\varepsilon\sim Laplace(w^Tx,b)$
因为当 y 服从拉式分布时 $L(\theta)=\log p(D|X,w,b)=\Sigma\log Lap(yi|w^Txi,b)=-N*\log(2b)-\Sigma|yi-w^Txi|/b$，其损失为残差绝对值和，对离群点不敏感；
而当 y 服从正态分布时，
$L(\theta)=\log p(D|X,w,b)=\Sigma\log N(yi|w^Txi,\sigma^2)=-(N/2)*\log(2\pi\sigma^2)-\Sigma(yi-w^Txi)^2/2\sigma^2$，其损失为残差平方和，放大了误差，对离群点敏感。因此使用 Laplace(或 Student)线性回归模型能对噪声鲁棒。

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

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

其中 $C=0.1$。

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

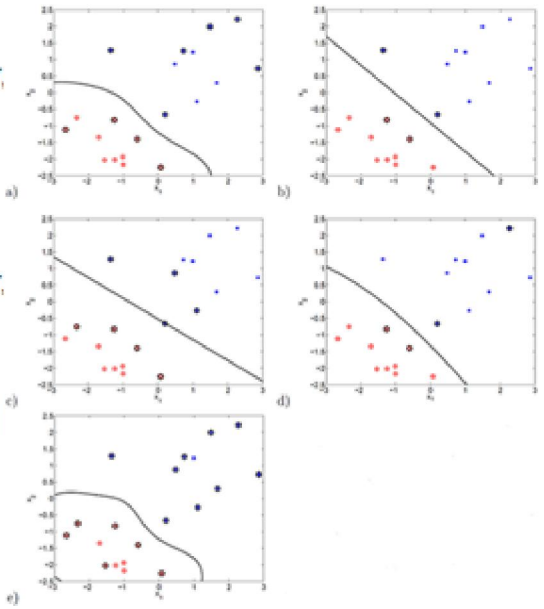
$\xi_i\geq 0,\ y_i\left(\mathbf{w}^T\mathbf{x}+w_0\right)\geq 1-\xi_i,\ i=1,\ldots,N,$

其中 $C=1$。

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

s.t. $\alpha_i\geq 0,\ i=1,\ldots,N,\quad \sum_{i=1}^{N}\alpha_i y_i=0$

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



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

s.t. $\alpha_i\geq 0,\ i=1,\ldots,N,\quad \sum_{i=1}^{N}\alpha_i y_i=0$

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

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

s.t. $\alpha_i\geq 0,\ i=1,\ldots,N,\quad \sum_{i=1}^{N}\alpha_i y_i=0$

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

答：1.c 2.b 3.d 4.a 5.e

6、考虑带松弛因子的线性 SVM 分类器：

$$\min\left(\frac{1}{2}\|\mathbf{w}\|^2+C\frac{1}{2}\sum_{i=1}^{N}\xi_i\right),\ \text{s.t.}\ \xi_i\geq 0,\ y_i\left(\mathbf{w}^T\mathbf{x}+w_0\right)\geq 1-\xi_i,\ i=1,\ldots,N$$

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) w0 不会增大 (不一定)
(2) ||w||增大 (不一定）
(3) ||w||不会减小 （是）
(4) 会有更多的训练样本被分错 （否）
(5) 间隔(Margin)不会增大 （是）

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）

答：1.不能。因为模型参数过多太复杂，训练集上拟合太好，把错误点也考虑进来，因此发生了过拟合，预测误差较大。

2.对之进行 L1 正则，即 Lasso 回归。$y\sim N(w^Tx,\sigma^2)$ $w\sim Lap(0,t)$

$L(\theta)=C-\Sigma(yi-w^Txi)^2/2\sigma^2-\Sigma|wi|/b$，$NLL=RSS+\lambda||w||$

通过调节 L1 正则系数 λ 大小避免模型过拟合，而且估计 w 参数的同时进行了特征选择，使得系数 w 尽可能多的为 0，简化了模型。