# 机器学习试题

**（一共30题，标有下划线的，如3\_1，3\_2，只用选择其中一题）**

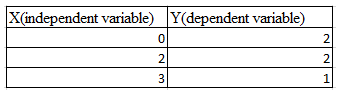
**1. 在回归模型中，下列哪一项在权衡欠拟合（under-fitting）和过拟合（over-fitting）中影响最大？**

A. 多项式阶数

B. 更新权重 w 时，使用的是矩阵求逆还是梯度下降

C. 使用常数项

**2. 假设你有以下数据：输入和输出都只有一个变量。使用线性回归模型（y=wx+b）来拟合数据。那么使用留一法（Leave-One Out）交叉验证得到的均方误差是多少？**



A. 10/27

B. 39/27

C. 49/27

D. 55/27

**3\_1. 下列关于极大似然估计（Maximum Likelihood Estimate，MLE），说法正确的是（多选）？**

A. MLE 可能并不存在

B. MLE 总是存在

C. 如果 MLE 存在，那么它的解可能不是唯一的

D. 如果 MLE 存在，那么它的解一定是唯一的

**3\_2. 下列哪些假设是我们推导线性回归参数时遵循的（多选）？**

A. X 与 Y 有线性关系（多项式关系）

B. 模型误差在统计学上是独立的

C. 误差一般服从 0 均值和固定标准差的正态分布

D. X 是非随机且测量没有误差的

**4\_1. 为了观察测试 Y 与 X 之间的线性关系，X 是连续变量，使用下列哪种图形比较适合？**

A. 散点图

B. 柱形图

C. 直方图

D. 以上都不对

**4\_2. 一般来说，下列哪种方法常用来预测连续独立变量？**

A. 线性回归

B. 逻辑回顾

C. 线性回归和逻辑回归都行

D. 以上说法都不对

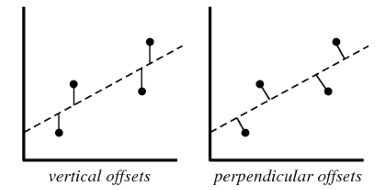
**5. 个人健康和年龄的相关系数是 -1.09。根据这个你可以告诉医生哪个结论？**

A. 年龄是健康程度很好的预测器

B. 年龄是健康程度很糟的预测器

C. 以上说法都不对

**6. 下列哪一种偏移，是我们在最小二乘直线拟合的情况下使用的？图中横坐标是输入 X，纵坐标是输出 Y。**



A. 垂直偏移（vertical offsets）

B. 垂向偏移（perpendicular offsets）

C. 两种偏移都可以

D. 以上说法都不对

**7. 假如我们利用 Y 是 X 的 3 阶多项式产生一些数据（3 阶多项式能很好地拟合数据）。那么，下列说法正确的是（多选）？**

A. 简单的线性回归容易造成高偏差（bias）、低方差（variance）

B. 简单的线性回归容易造成低偏差（bias）、高方差（variance）

C. 3 阶多项式拟合会造成低偏差（bias）、高方差（variance）

D. 3 阶多项式拟合具备低偏差（bias）、低方差（variance）

**8\_1. 假如你在训练一个线性回归模型，有下面两句话：**

**1. 如果数据量较少，容易发生过拟合。**

**2. 如果假设空间较小，容易发生过拟合。**

**关于这两句话，下列说法正确的是？**

A. 1 和 2 都错误

B. 1 正确，2 错误

C. 1 错误，2 正确

D. 1 和 2 都正确

**8\_2. 假如我们使用 Lasso 回归来拟合数据集，该数据集输入特征有 100 个（X1，X2，...，X100）。现在，我们把其中一个特征值扩大 10 倍（例如是特征 X1），然后用相同的正则化参数对 Lasso 回归进行修正。**

**那么，下列说法正确的是？**

A. 特征 X1 很可能被排除在模型之外

B. 特征 X1 很可能还包含在模型之中

C. 无法确定特征 X1 是否被舍弃

D. 以上说法都不对

**9\_1. 关于特征选择，下列对 Ridge 回归和 Lasso 回归说法正确的是？**

A. Ridge 回归适用于特征选择

B. Lasso 回归适用于特征选择

C. 两个都适用于特征选择

D. 以上说法都不对

**9\_2. 如果在线性回归模型中增加一个特征变量，下列可能发生的是（多选）？**

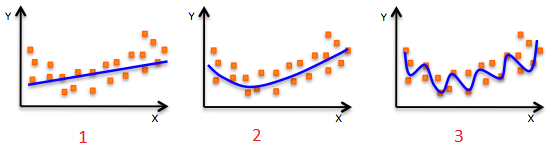
A. R-squared 增大，Adjust R-squared 增大

B. R-squared 增大，Adjust R-squared 减小

C. R-squared 减小，Adjust R-squared 减小

D. R-squared 减小，Adjust R-squared 增大

1. **下面三张图展示了对同一训练样本，使用不同的模型拟合的效果（蓝色曲线）。那么，我们可以得出哪些结论（多选）？**



A. 第 1 个模型的训练误差大于第 2 个、第 3 个模型

B. 最好的模型是第 3 个，因为它的训练误差最小

C. 第 2 个模型最为“健壮”，因为它对未知样本的拟合效果最好

D. 第 3 个模型发生了过拟合

E. 所有模型的表现都一样，因为我们并没有看到测试数据

**11. 下列哪些指标可以用来评估线性回归模型（多选）？**

A. R-Squared

B. Adjusted R-Squared

C. F Statistics

D. RMSE / MSE / MAE

**12\_1. 线性回归中，我们可以使用正规方程（Normal Equation）来求解系数。下列关于正规方程说法正确的是？**

A. 不需要选择学习因子

B. 当特征数目很多的时候，运算速度会很慢

C. 不需要迭代训练

**12\_2. 如果 Y 是 X（X1，X2，...，Xn）的线性函数：**

**Y = β0 + β1X1 + β2X2 + ··· + βnXn**

**则下列说法正确的是（多选）？**

A. 如果变量 Xi 改变一个微小变量 ΔXi，其它变量不变。那么 Y 会相应改变 βiΔXi。

B. βi 是固定的，不管 Xi 如何变化

C. Xi 对 Y 的影响是相互独立的，且 X 对 Y 的总的影响为各自分量 Xi 之和

**13. 构建一个最简单的线性回归模型需要几个系数（只有一个特征）？**

A. 1 个

B. 2 个

C. 3 个

D. 4 个

**14. 如果两个变量相关，那么它们一定是线性关系吗？**

A. 是

B. 不是

**15. 两个变量相关，它们的相关系数 r 可能为 0。这句话是否正确？**

A. 正确

B. 错误

**16\_1. 加入使用逻辑回归对样本进行分类，得到训练样本的准确率和测试样本的准确率。现在，在数据中增加一个新的特征，其它特征保持不变。然后重新训练测试。则下列说法正确的是？**

A. 训练样本准确率一定会降低

B. 训练样本准确率一定增加或保持不变

C. 测试样本准确率一定会降低

D. 测试样本准确率一定增加或保持不变

**16\_2. 假设一个公司的薪资水平中位数是 $35,000，排名第 25% 和 75% 的薪资分别是21000和  53,000。如果某人的薪水是 $1，那么它可以被看成是异常值（Outlier）吗？**

A. 可以

B. 不可以

C. 需要更多的信息才能判断

D. 以上说法都不对

**13. 关于“回归（Regression）”和“相关（Correlation）”，下列说法正确的是？注意：x 是自变量，y 是因变量。**

A. 回归和相关在 x 和 y 之间都是互为对称的

B. 回归和相关在 x 和 y 之间都是非对称的

C. 回归在 x 和 y 之间是非对称的，相关在 x 和 y 之间是互为对称的

D. 回归在 x 和 y 之间是对称的，相关在 x 和 y 之间是非对称的

**14. 仅仅知道变量的均值（Mean）和中值（Median），能计算的到变量的偏斜度（Skewness）吗？**

A. 可以

B. 不可以

**15. 观察样本次数如何影响过拟合（多选）？注意：所有情况的参数都保持一致。**

A. 观察次数少，容易发生过拟合

B. 观察次数少，不容易发生过拟合

C. 观察次数多，容易发生过拟合

D. 观察次数多，不容易发生过拟合

**16. 假如使用一个较复杂的回归模型来拟合样本数据，使用 Ridge 回归，调试正则化参数 λ，来降低模型复杂度。若 λ 较大时，关于偏差（bias）和方差（variance），下列说法正确的是？**

A. 若 λ 较大时，偏差减小，方差减小

B. 若 λ 较大时，偏差减小，方差增大

C. 若 λ 较大时，偏差增大，方差减小

D. 若 λ 较大时，偏差增大，方差增大

**17. 假设使用逻辑回归进行 n 多类别分类，使用 One-vs-rest 分类法。下列说法正确的是？**

A. 对于 n 类别，需要训练 n 个模型

B. 对于 n 类别，需要训练 n-1 个模型

C. 对于 n 类别，只需要训练 1 个模型

D. 以上说法都不对

**18. 在 n 维空间中（n > 1），下列哪种方法最适合用来检测异常值？**

A. 正态概率图

B. 箱形图

C. 马氏距离

D. 散点图

**19\_1. 逻辑回归与多元回归分析有哪些不同之处？**

A. 逻辑回归用来预测事件发生的概率

B. 逻辑回归用来计算拟合优度指数

C. 逻辑回归用来对回归系数进行估计

D. 以上都是

**19\_2. 下列关于 bootstrap 说法正确的是？**

A. 从总的 M 个特征中，有放回地抽取 m 个特征（m < M）

B. 从总的 M 个特征中，无放回地抽取 m 个特征（m < M）

C. 从总的 N 个样本中，有放回地抽取 n 个样本（n < N）

D. 从总的 N 个样本中，无放回地抽取 n 个样本（n < N）

**20. "监督式学习中存在过拟合，而对于非监督式学习来说，没有过拟合"，这句话是否正确？**

A. 正确

B. 错误

**21. 关于 k 折交叉验证，下列说法正确的是？**

A. k 值并不是越大越好，k 值过大，会降低运算速度

B. 选择更大的 k 值，会让偏差更小，因为 k 值越大，训练集越接近整个训练样本

C. 选择合适的 k 值，能减小验方差

D. 以上说法都正确

**22. 如果回归模型中存在多重共线性（multicollinearity），应该如何解决这一问题而不丢失太多信息（多选）？**

A. 剔除所有的共线性变量

B. 剔除共线性变量中的一个

C. 通过计算方差膨胀因子（Variance Inflation Factor，VIF）来检查共线性程度，并采取相应措施

D. 删除相关变量可能会有信息损失，我们可以不删除相关变量，而使用一些正则化方法来解决多重共线性问题，例如 Ridge 或 Lasso 回归。

**23. 评估完模型之后，发现模型存在高偏差（high bias），应该如何解决？**

A. 减少模型的特征数量

B. 增加模型的特征数量

C. 增加样本数量

D. 以上说法都正确

**24. 在构建一个决策树模型时，我们对某个属性分割节点，下面四张图中，哪个属性对应的信息增益最大？**

A. outlook

B. humidity

C. windy

D. temperature

**25. 在决策树分割结点的时候，下列关于信息增益说法正确的是（多选）？**

A. 纯度高的结点需要更多的信息来描述它

B. 信息增益可以用”1比特-熵”获得

C. 如果选择一个属性具有许多特征值, 那么这个信息增益是有偏差的

**26. 如果一个 SVM 模型出现欠拟合，那么下列哪种方法能解决这一问题？**

A. 增大惩罚参数 C 的值

B. 减小惩罚参数 C 的值

C. 减小核系数（gamma参数）

**27. 我们知道二元分类的输出是概率值。一般设定输出概率大于或等于 0.5，则预测为正类；若输出概率小于 0.5，则预测为负类。那么，如果将阈值 0.5 提高，例如 0.6，大于或等于 0.6 的才预测为正类。则准确率（Precision）和召回率（Recall）会发生什么变化（多选）？**

A. 准确率（Precision）增加或者不变

B. 准确率（Precision）减小

C. 召回率（Recall）减小或者不变

D. 召回率（Recall）增大

**28\_1. 点击率预测是一个正负样本不平衡问题（例如 99% 的没有点击，只有 1% 点击）。假如在这个非平衡的数据集上建立一个模型，得到训练样本的正确率是 99%，则下列说法正确的是？**

A. 模型正确率很高，不需要优化模型了

B. 模型正确率并不高，应该建立更好的模型

C. 无法对模型做出好坏评价

D. 以上说法都不对

**28\_2. 假设我们使用 kNN 训练模型，其中训练数据具有较少的观测数据（下图是两个属性 x、y 和两个标记为 “+” 和 “o” 的训练数据）。现在令 k = 1，则图中的 Leave-One-Out 交叉验证错误率是多少？**

A. 0%

B. 20%

C. 50%

D. 100%

**29. 如果在大型数据集上训练决策树。为了花费更少的时间来训练这个模型，下列哪种做法是正确的？**

A. 增加树的深度

B. 增加学习率

C. 减小树的深度

D. 减少树的数量

**30. 关于神经网络，下列说法正确的是？**

A. 增加网络层数，可能会增加测试集分类错误率

B. 增加网络层数，一定会增加训练集分类错误率

C. 减少网络层数，可能会减少测试集分类错误率

D. 减少网络层数，一定会减少训练集分类错误率