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？

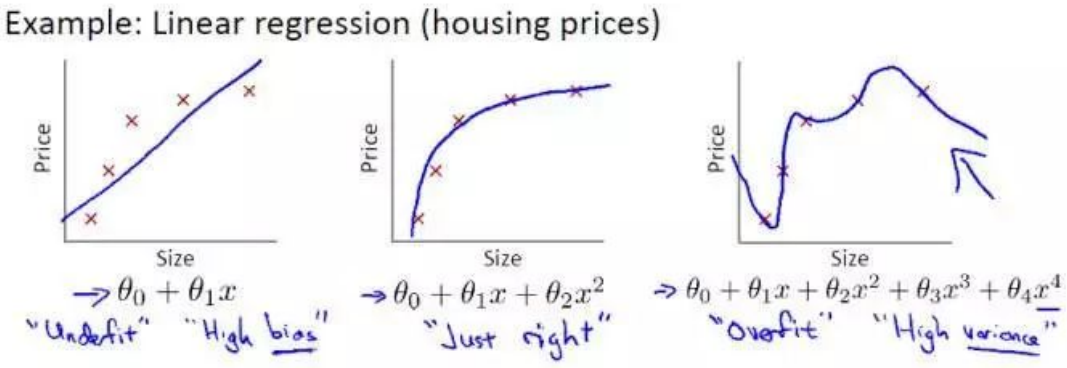
A. 多项式阶数

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

答案：A

解析：选择合适的多项式阶数非常重要。如果阶数过大，模型就会更加复杂，容易发生过拟合；如果阶数较小，模型就会过于简单，容易发生欠拟合。如果有对过拟合和欠拟合概念不清楚的，见下图所示：



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？

X(independent variable)	Y(dependent variable)
0	2
2	2
3	1

A. 10/27

B. 39/27

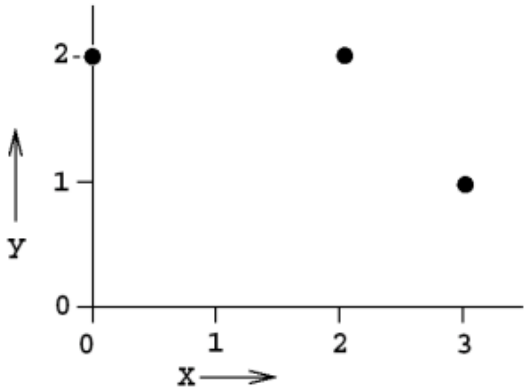
C. 49/27

D. 55/27

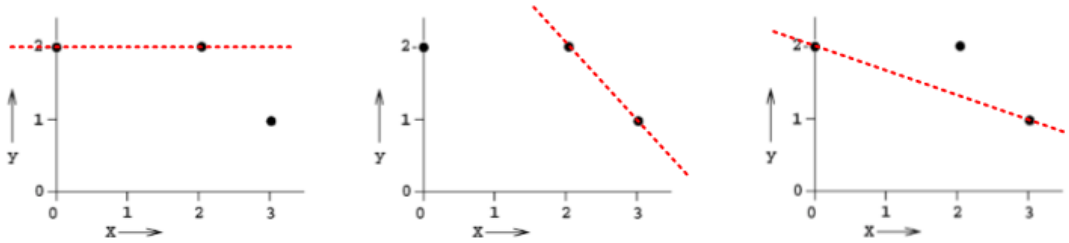
答案：C

解析：留一法，简单来说就是假设有 N 个样本，将每一个样本作为测试样本，其它 N-1 个样本作为训练样本。这样得到 N 个分类器，N 个测试结果。用这 N 个结果的平均值来衡量模型的性能。

对于该题，我们先画出 3 个样本点的坐标：



使用两个点进行线性拟合，分成三种情况，如下图所示：



第一种情况下，回归模型是 y = 2，误差 E1 = 1。

第二种情况下，回归模型是 y = -x + 4，误差 E2 = 2。

第三种情况下，回归模型是 y = -1/3x + 2,误差 E3 = 2/3。

则总的均方误差为：

$$MSE = \frac{1}{3}(E_1^2 + E_2^2 + E_3^2) = \frac{1}{3}[1^2 + 2^2 + (\frac{2}{3})^2] = \frac{49}{27}$$

3. 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

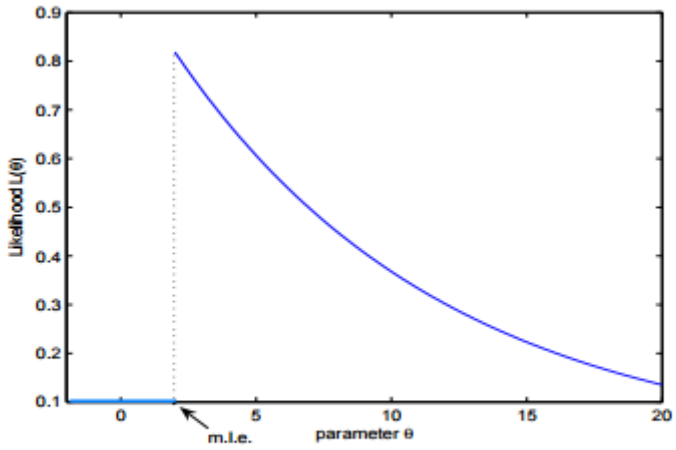
B. MLE 总是存在

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

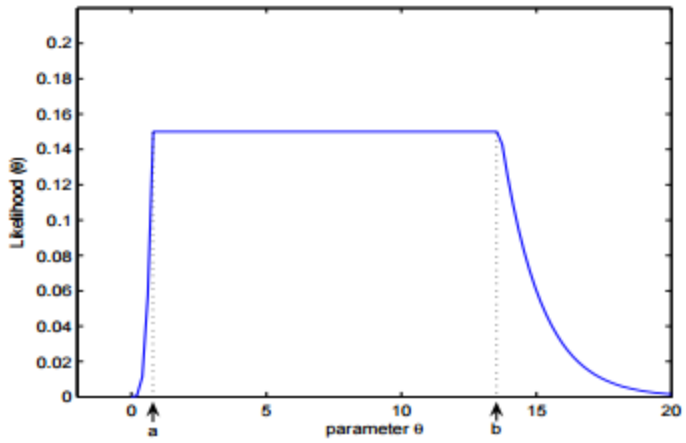
答案：AC

解析：如果极大似然函数 L(θ) 在极大值处不连续，一阶导数不存在，则 MLE 不存在，如下图所示：



The m.l.e. is a boundary point

另一种情况是 MLE 并不唯一，极大值对应两个 θ。如下图所示：



Any point between a and b is a m.l.e

4. 如果我们说“线性回归”模型完美地拟合了训练样本（训练样本误差为零），则下面哪个说法是正确的？

A. 测试样本误差始终为零

B. 测试样本误差不可能为零

C. 以上答案都不对

答案：C

解析：根据训练样本误差为零，无法推断测试样本误差是否为零。值得一提是，如果测试样本样本很大，则很可能发生过拟合，模型不具备很好的泛化能力！

5. 在一个线性回归问题中，我们使用 R 平方（R-Squared）来判断拟合度。此时，如果增加一个特征，模型不变，则下面说法正确的是？

A. 如果 R-Squared 增加，则这个特征有意义

B. 如果 R-Squared 减小，则这个特征没有意义

C. 仅看 R-Squared 单一变量，无法确定这个特征是否有意义。

D. 以上说法都不对

答案：C

解析：线性回归问题中，R-Squared 是用来衡量回归方程与真实样本输出之间的相似程度。其表达式如下所示：

$$R^2=1-\frac{\sum(y-\hat{y})^2}{\sum(y-\bar{y})^2}$$

上式中，分子部分表示真实值与预测值的平方差之和，类似于均方差 MSE；分母部分表示真实值与均值的平方差之和，类似于方差 Var。根据 R-Squared 的取值，来判断模型的好坏：如果结果是 0，说明模型拟合效果很差；如果结果是 1，说明模型无错误。一般来说，R-Squared 越大，表示模型拟合效果越好。R-Squared 反映的是大概有多准，因为，随着样本数量的增加，R-Square 必然增加，无法真正定量说明准确程度，只能大概定量。

对于本题来说，单独看 R-Squared，并不能推断出增加的特征是否有意义。通常来说，增加一个特征，R-Squared 可能变大也可能保持不变，两者不一定呈正相关。

如果使用校正决定系数（Adjusted R-Square）：

$$R^2\_adjusted = 1 - \frac{(1-R^2)(n-1)}{n-p-1}$$

其中，n 是样本数量，p 是特征数量。Adjusted R-Square 抵消样本数量对 R-Square 的影响，做到了真正的 0~1，越大越好。

6. 下列关于线性回归分析中的残差（Residuals）说法正确的是？

A. 残差均值总是为零

B. 残差均值总是小于零

C. 残差均值总是大于零

D. 以上说法都不对

答案：A

解析：线性回归分析中，目标是残差最小化。残差平方和是关于参数的函数，为了求残差极小值，令残差关于参数的偏导数为零，会得到残差和为零，即残差均值为零。

7. 下列关于异方差（Heteroskedasticity）说法正确的是？

A. 线性回归具有不同的误差项

B. 线性回归具有相同的误差项

C. 线性回归误差项为零

D. 以上说法都不对

答案：A

解析：异方差性是相对于同方差（Homoskedasticity）而言的。所谓同方差，是为了保证回归参数估计量具有良好的统计性质，经典线性回归模型的一个重要假定：总体回归函数中的随机误差项满足同方差性，即它们都有相同的方差。如果这一假定不满足，即：随机误差项具有不同的方差，则称线性回归模型存在异方差性。

通常来说，奇异值的出现会导致异方差性增大。

8. 下列哪一项能反映出 X 和 Y 之间的强相关性？

A. 相关系数为 0.9

B. 对于无效假设 β=0 的 p 值为 0.0001

C. 对于无效假设 β=0 的 t 值为 30

D. 以上说法都不对

答案：A

解析：相关系数的概念我们很熟悉，它反映了不同变量之间线性相关程度，一般用 r 表示。

$$r(X,Y)=\frac{Cov(X,Y)}{\sqrt{Var[X]\,Var[Y]}}$$

其中，Cov(X,Y) 为 X 与 Y 的协方差，Var[X] 为 X 的方差，Var[Y] 为 Y 的方差。r 取值范围在 [-1,1] 之间，r 越大表示相关程度越高。A 选项中，r=0.9 表示 X 和 Y 之间有较强的相关性。

而 p 和 t 的数值大小没有统计意义，只是将其与某一个阈值进行比对，以得到二选一的结论。例如，有两个假设：

- 无效假设（null hypothesis）H0：两参量间不存在“线性”相关。

- 备择假设（alternative hypothesis）H1：两参量间存在“线性”相关。

  如果阈值是 0.05，计算出的 p 值很小，比如为 0.001，则可以说“有非常显著的证据拒绝 H0 假设,相信 H1 假设。即两参量间存在“线性”相关。p 值只用于二值化判断，因此不能说 p=0.06 一定比 p=0.07 更好。

9. 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

答案：ABCD

解析：在进行线性回归推导和分析时，我们已经默认上述四个条件是成立的。

10. 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

答案：A

解析：散点图反映了两个变量之间的相互关系，在测试 Y 与 X 之间的线性关系时，使用散点图最为直观。

11. 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

答案：A

解析：线性回归一般用于实数预测，逻辑回归一般用于分类问题。

12. 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

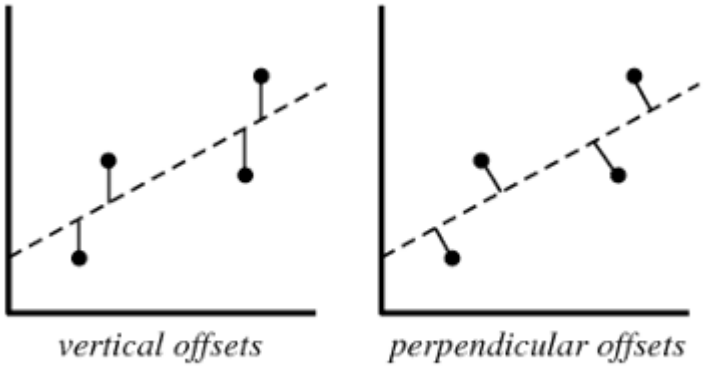
B. 年龄是健康程度很糟的预测器

C. 以上说法都不对

答案：C

解析：因为相关系数的范围是 [-1,1] 之间，所以，-1.09 不可能存在。

13. 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

答案：A

解析：线性回归模型计算损失函数，例如均方差损失函数时，使用的都是 vertical offsets。perpendicular offsets 一般用于主成分分析（PCA）中。

14. 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）

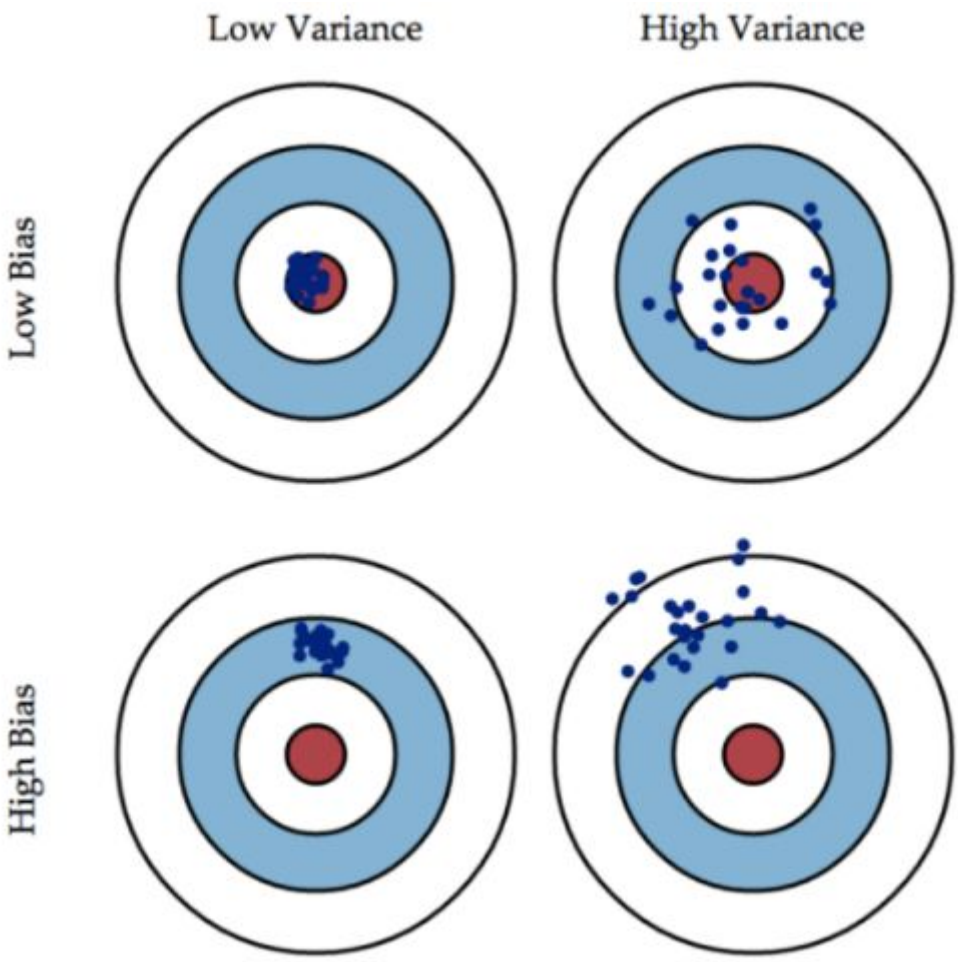
C. 3 阶多项式拟合会造成低偏差（bias）、高方差（variance）

D. 3 阶多项式拟合具备低偏差（bias）、低方差（variance）

答案：AD

解析：偏差和方差是两个相对的概念，就像欠拟合和过拟合一样。如果模型过于简单，通常会造成欠拟合，伴随着高偏差、低方差；如果模型过于复杂，通常会造成过拟合，伴随着低偏差、高方差。

用一张图来形象地表示偏差与方差的关系：



偏差（bias）可以看成模型预测与真实样本的差距，想要得到 low bias，就得复杂化模型，但是容易造成过拟合。方差（variance）可以看成模型在测试集上的表现，想要得到 low variance，就得简化模型，但是容易造成欠拟合。实际应用中，偏差和方差是需要权衡的。若模型在训练样本和测试集上都表现的不错，偏差和方差都会比较小，这也是模型比较理想的情况。

15. 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

答案：B

解析：先来看第 1 句话，如果数据量较少，容易在假设空间找到一个模型对训练样本的拟合度很好，容易造成过拟合，该模型不具备良好的泛化能力。

再来看第 2 句话，如果假设空间较小，包含的可能的模型就比较少，也就不太可能找到一个模型能够对样本拟合得很好，容易造成高偏差、低方差，即欠拟合。

16. 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

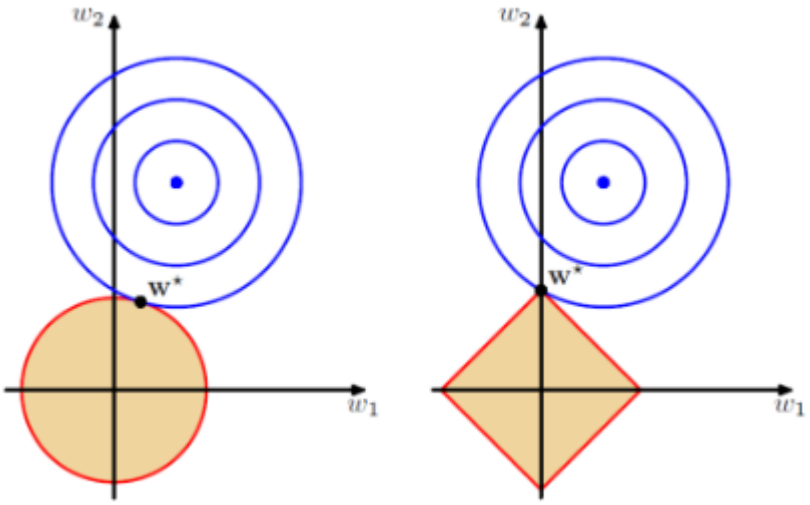
D. 以上说法都不对

答案： B

解析：Lasso 回归类似于线性回归，只不过它在线性回归的基础上，增加了一个对所有参数的数值大小约束，如下所示：

$$min \frac{1}{m} \sum_{i=1}^{m} (y_i - \beta_0 - x_i^T \beta)^2 \quad subject\ to \ ||\beta||_1 \leq t$$

其中，t 为正则化参数。Lasso 回归其实就是在普通线性回归的损失函数的基础上增加了个 β 的约束。那么 β 的约束为什么要使用这种形式，而不使用 β 的平方约束呢？原因就在于第一范数的约束下，一部分回归系数刚好可以被约束为 0。这样的话，就达到了特征选择的效果。如下图所示：



左边是第二范式，右边是第一范式。第一范数约束下，β 更有可能被约束成 0。这点非常类似于 L1 和 L2 正则化的区别，

因此，Lasso 回归适用于样本数量较少，特征维度较大的情形，便于从较多特征中进行特征选择。例如 DNA 数据，特征维度很大，我们只希望通过 Lasso 回归找出与某些疾病有关的 DNA 片段。

本题中，将特征 X1 数值扩大 10 倍，他对应的回归系数将相应会减小，但不为 0，以此来保证仍然满足 β 的正则化约束。

17. 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

答案：B

解析：上一题我们已经介绍过，Lasso 回归会让一部分回归系数刚好可以被约束为 0，起到特征选择的效果。

Ridge 回归又称岭回归，它是普通线性回归加上 L2 正则项，用来防止训练过程中出现的过拟合。L2 正则化效果类似上一题左图，限定区域是圆，这样，得到的回归系数为 0 的概率很小，很大概率是非零的。因此，比较来说，Lasso 回归更容易得到稀疏的回归系数，有利于舍弃冗余或无用特征，适用于特征选择。

18. 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

答案：AB

解析：线性回归问题中，R-Squared 是用来衡量回归方程与真实样本输出之间的相似程度。其表达式如下所示：

$$R^2=1-\frac{\sum(y-\hat{y})^2}{\sum(y-\bar{y})^2}$$

上式中，分子部分表示真实值与预测值的平方差之和，类似于均方差 MSE；分母部分表示真实值与均值的平方差之和，类似于方差 Var。一般来说，R-Squared 越大，表示模型拟合效果越好。R-Squared 反映的是大概有多准，因为，随着样本数量的增加，R-Squared 必然增加，无法真正定量说明准确程度，只能大概定量。

单独看 R-Squared，并不能推断出增加的特征是否有意义。通常来说，增加一个特征特征，R-Squared 可能变大也可能保持不变，两者不一定呈正相关。

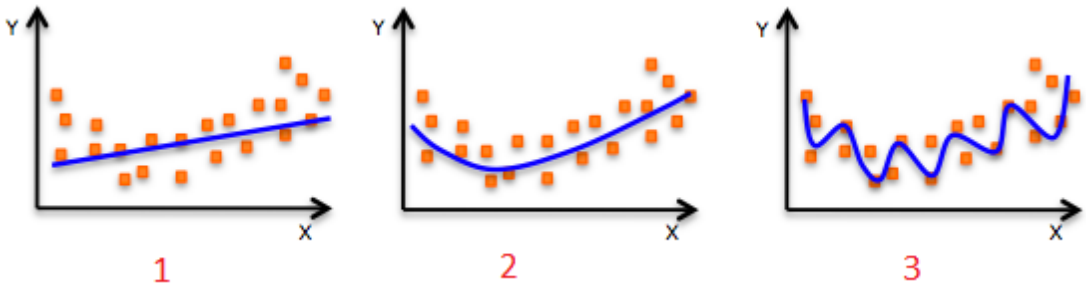
如果使用校正决定系数（Adjusted R-Squared）：

$$R^2\_adjusted=1-\frac{(1-R^2)(n-1)}{n-p-1}$$

其中，n 是样本数量，p 是特征数量。Adjusted R-Squared 抵消样本数量对 R-Squared 的影响，做到了真正的 0~1，越大越好。

增加一个特征变量，如果这个特征有意义，Adjusted R-Square 就会增大，若这个特征是冗余特征，Adjusted R-Squared 就会减小。

19. 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

答案：ACD

解析：1、2、3 模型分别对应的多项式阶数由小到大，即模型由简单到复杂。模型越简单，容易发生欠拟合；模型越复杂，容易发生过拟合。第 1 个模型过于简单，出现欠拟合；第 3 个模型过于复杂，对训练样本拟合得很好，但在测试样本上效果会很差，即过拟合；第 2 个模型最为“健壮”，在训练样本和测试样本上拟合效果都不错！

20. 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

答案：ABCD

解析：R-Squared 和 Adjusted R-Squared 的概念，我们在 Q3 有过介绍，它们都可以用来评估线性回归模型。F Statistics 是指在零假设成立的情况下，符合 F 分布的统计量，多用于计量统计学中。

RMSE 指的是均方根误差：

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(y^{(i)} - \hat{y}^{(i)})^2}$$

MSE 指的是均方误差：

$$MSE = \frac{1}{m}\sum_{i=1}^{m}(y^{(i)} - \hat{y}^{(i)})^2$$

MAE 指的是评价绝对误差：

$$MAE = \frac{1}{m}\sum_{i=1}^{m}|y^{(i)} - \hat{y}^{(i)}|$$

以上指标都可以用来评估线性回归模型。

21. 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

答案：ABC

解析：求解线性回归系数，我们一般最常用的方法是梯度下降，利用迭代优化的方式。除此之外，还有一种方法是使用正规方程，原理是基于最小二乘法。下面对正规方程做简要的推导。

已知线性回归模型的损失函数 Ein 为：

$$E_{in} = \frac{1}{m}(XW - Y)^2$$

对 Ein 计算导数，令 ∇Ein=0：

$$\nabla E_{in} = \frac{2}{m}(X^T XW - X^T Y) = 0$$

然后就能计算出 W：

$$W = (X^T X)^{-1} X^T Y$$

以上就是使用正规方程求解系数 W 的过程。可以看到，正规方程求解过程不需要学习因子，也没有迭代训练过程。当特征数目很多的时候，XTX 矩阵求逆会很慢，这时梯度下降算法更好一些。

如果 XTX 矩阵不可逆，是奇异矩阵怎么办呢？其实，大部分的计算逆矩阵的软件程序，都可以处理这个问题，也会计算出一个逆矩阵。所以，一般伪逆矩阵是可解的。

22. 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

答案：ABC

解析：这题非常简单，Y 与 X（X1，X2，...，Xn）是线性关系，故能得出 ABC 结论。

23. 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

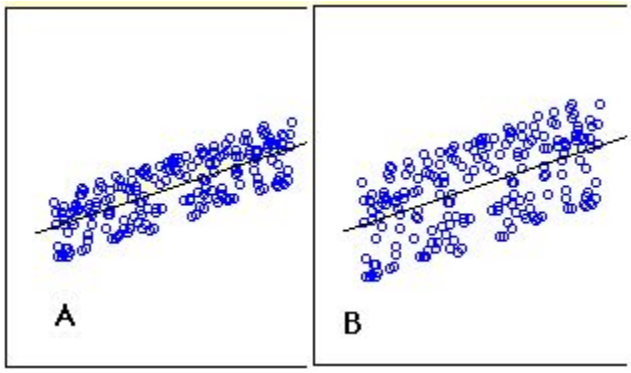
C. 3 个

D. 4 个

答案：B

解析：最简单的线性回归模型，只有一个特征，即 Y = aX + b，包含 a 和 b 两个系数。

24. 下面两张图展示了两个拟合回归线（A 和 B），原始数据是随机产生的。现在，我想要计算 A 和 B 各自的残差之和。注意：两种图中的坐标尺度一样。



关于 A 和 B 各自的残差之和，下列说法正确的是？

A. A 比 B 高

B. A 比 B 小

C. A 与 B 相同

D. 以上说法都不对

答案：C

解析：A 和 B 中各自的残差之和应该是相同的。线性回归模型的损失函数为：

$$J = \frac{1}{m}(XW - Y)^2$$

对损失函数求导，并令 ∇J=0，即可得到 XW-Y=0，即残差之和始终为零。

25. 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

答案：B

解析：相关不一定是线性关系，也有可能是非线性相关。

26. 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

答案：A

解析：一般来说，相关系数 r=0 是两变量相互独立的必要不充分条件。也就是说，如果两个变量相互独立，那么相关系数 r 一定为 0，如果相关系数 r=0，则不一定相互独立。相关系数 r=0 只能说明两个变量之间不存在线性关系，仍然可能存在非线性关系。

那么，若两个变量相关，存在非线性关系，那么它们的相关系数 r 就为 0。

27. 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

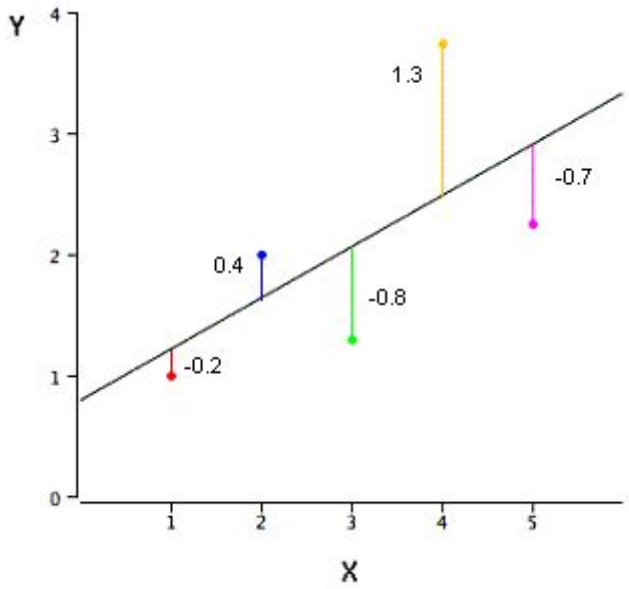
D. 测试样本准确率一定增加或保持不变

答案：B

解析：在模型中增加更多特征一般会增加训练样本的准确率，减小bias。但是测试样本准确率不一定增加，除非增加的特征是有效特征。

这题对应的知识点也包括了增加模型复杂度，虽然会减小训练样本误差，但是容易发生过拟合。

28. 下面这张图是一个简单的线性回归模型,图中标注了每个样本点预测值与真实值的残差。计算 SSE 为多少？



A. 3.02

B. 0.75

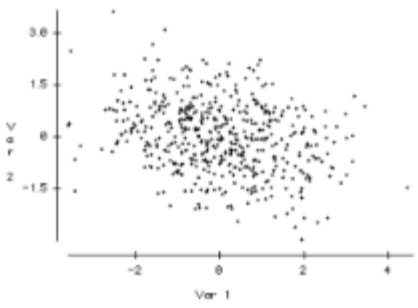
C. 1.01

D. 0.604

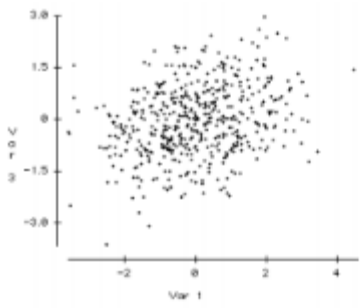
答案：A

解析：SSE 是平方误差之和（Sum of Squared Error），SSE = (-0.2)^2 + (0.4)^2 + (-0.8)^2 + (1.3)^2 + (-0.7)^2 = 3.02

29. 假设变量 Var1 和 Var2 是正相关的，那么下面那张图是正确的？图中，横坐标是 Var1，纵坐标是 Var2，且对 Var1 和 Var2 都做了标准化处理。



Plot 1



Plot 2

A. Plot 1

B. Plot 2

答案：B

解析：显然，Plot 2 显示出 Var2 与 Var1 是正相关的，例如 Var2 = Var1。Plot 1 显示出 Var2 与 Var1 是负相关的，例如 Var2 = -Var1。

30. 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 $21,000 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

答案：C

解析：异常值（Outlier）指样本中的个别值，其数值明显偏离它（或他们）所属样本的其余观测值，也称异常数据，离群值。目前人们对异常值的判别与剔除主要采用物理判别法和统计判别法两种方法。

所谓物理判别法就是根据人们对客观事物已有的认识，判别由于外界干扰、人为误差等原因造成实测数据值偏离正常结果，在实验过程中随时判断，随时剔除。

统计判别法是给定一个置信概率，并确定一个置信限，凡超过此限的误差，就认为它不属于随机误差范围，将其视为异常值剔除。当物理识别不易判断时，一般采用统计识别法。

该题中，所给的信息量过少，无法肯定一定是异常值。