1. 决策树分类树怎么进行分类的?
   ID3：信息增益
   C4.5:信息增益比
   CART：基尼系数
2. L1 与 L2 的区别?
   L1 正则化产生稀疏的权值, L2 正则化产生平滑的权值,

L1 正则化偏向于稀疏，它会自动进行特征选择，去掉一些没用的特征，也就是将这些特征对应的权重置为 0.

L2 主要功能是为了防止过拟合，当要求参数越小时，说明模型越简单，而模型越简单则，越趋向于平滑，从而防止过拟合。

正则化力度： 大： 参数趋近于 0
小： 参数变化小

3. LR 为什么用 sigmoid 函数。这个函数有什么优点和缺点？为什么不用其他函数?
   1、 优点：他的输入范围是$-\infty \rightarrow +\infty$ ，而之于刚好为（0，1），正好满足概率分布为（0，1）的要求。我们用概率去描述分类器，自然比单纯的某个阈值要方便很多，有很好的连续性
   2、 缺点：会导致求解不到结果
4. 生成模型和判别模型:
   生成模型有：朴素贝叶斯，隐马尔可夫模型
   判别模型有：KNN、SVM、LR、决策树
   区别：是否需要先验概率 P(文档词 1，文档词 2|类别)
5. kmeans 的原理，优缺点以及改进。
   原理:
   优缺点：1、K 中心点：由于取值随机，点非常相近,
   改进：1、N< K 个中心点，剩下 K-N 个中心点，选出距离这 N 个中心点较大的样本
6. softmax 公式

7. 为什么要做数据归一化?
   归一化：1、a = (x-min)/(max-min) 2、a(ma-min) + mi, 所有特征的数据缩小到指定的(mi,ma)
8. 信息熵公式
   -(p1logp1+ …+pnlogpn)
9. 决策树原理
   首先从大的角度说怎么解决分类,
   信息增益：H(D，A) = H(D) – H(D|A)
   C4.5：信息增益比，CART：基尼系数
10. 分类模型和回归模型的区别
    输出值：离散和连续
11. 如何防止过拟合
    特征选择：（1）低方差：在对数据不清楚的情况下，不适用
    方差=0，所有数据相等
    （2）决策树减枝，随机森林:
    （3）L2：回归（包括逻辑回归这种分类模型）
    （4）神经网络：池化层，batch normalization, droupout:随机指定部分神经元失效，不是删除（保留 80%， 其余 20%的神经元值变为 0）
12. 常见分类模型（ KNN，决策树，贝叶斯等）的优缺点，适用场景以及如何选型
    KNN：一个比较容易解释，而且不同维度之间影响小的模型的时候
    决策树：适用于一些特征非常清晰的场景，通常也作为更好的分类模型的基础算法，对于缺失数据不敏感

缺点：1、过拟合 2、特征的变化（异常点）

贝叶斯：需要一个比较容易解释，而且不同维度之间相关性较小的模型的时候，需要计算先验概率

计算题

如下表所示的数据集。请写出按属性 A 和 B 划分时的信息增益的计算表达式。不需要计算出最后结果。并回答计算信息增益在分类算法中的作用。

A	B	类标号
T	F	*
T	T	*
T	T	*
T	F	#
T	T	*
F	F	#
F	F	#
F	F	#
T		#
T	F	#

1、请简要介绍下 tensorflow 的计算图

Tensorflow 是一个通过计算图的形式来表述计算的编程系统，计算图也叫数据流图，可以把计算图看做是一种有向图，Tensorflow 中的每一个计算都是计算图上的一个节点，而

1、把每部分结构说清楚，图，会话，张量，op

2、节点之间的边描述了计算之间的依赖关系。

6、谈谈判别式模型和生成式模型？

判别方法：由数据直接学习决策函数 Y = f（X），或者由条件分布概率 P（Y|X）作为预测模型，即判别模型。

生成方法：由数据学习联合概率密度分布函数 P（X,Y）,然后求出条件概率分布 P(Y|X)作为预测的模型，即生成模型。

由生成模型可以得到判别模型，但由判别模型得不到生成模型。

常见的判别模型有：K 近邻、SVM、决策树、感知机、线性判别分析（LDA）、线性回归、传统的神经网络、逻辑斯蒂回归、boosting、条件随机场

常见的生成模型有：朴素贝叶斯、隐马尔可夫模型、高斯混合模型、文档主题生成模型（LDA）、限制玻尔兹曼机

7、L1 和 L2 正则化的区别

L1 范数可以使权值稀疏，方便特征提取。

L2 范数可以防止过拟合，提升模型的泛化能力。

9、简单说说 CNN 的原理

卷积层:filter: 5*5, strides1， 多少个 filter

10、为什么朴素贝叶斯如此“朴素”？

因为它假定所有的特征在数据集中的作用是同样重要和独立的。正如我们所知，这个假设在现实世界中是很不真实的，因此，说朴素贝叶斯真的很“朴素”。

12、KNN 中的 K 如何选取的？影响？

KNN 中的 K 值选取对 K 近邻算法的结果会产生重大影响。如李航博士的一书「统计学习方法」上所说：

在实际应用中，K 值一般取一个比较小的数值，例如采用交叉验证法（简单来说，就是一部分样本做训练集，一部分做测试集）来选择最优的 K 值。

a. 如果选择较小的 K 值，就相当于用较小的领域中的训练实例进行预测，“学习”近似误差会减小，只有与输入实例较近或相似的训练实例才会对预测结果起作用，与此同时带来的问题是“学习”的估计误差会增大，换句话说，K 值的减小就意味着整体模型变得复杂，容易发生过拟合；

b. 如果选择较大的 K 值，就相当于用较大领域中的训练实例进行预测，其优点是可以减少学习的估计误差，但缺点是学习的近似误差会增大。这时候，与输入实例较远（不相似的）训练实例也会对预测器作用，使预测发生错误，且 K 值的增大就意味着整体的模型变得简单。

c. K=N，则完全不足取，因为此时无论输入实例是什么，都只是简单的预测它属于在训练实例中最多的累，模型过于简单，忽略了训练实例中大量有用信息。

14、请简要说说一个完整机器学习项目的流程。

1 抽象成数学问题

明确问题是进行机器学习的第一步。机器学习的训练过程通常都是一件非常耗时的事情，胡乱尝试时间成本是非常高的。

这里的抽象成数学问题，指的我们明确我们可以获得什么样的数据，目标是一个分类还是回归或者是聚类的问题，如果都不是的话，如果划归为其中的某类问题。

2 获取数据

数据要有代表性，否则必然会过拟合。

而且对于分类问题，数据偏斜不能过于严重，不同类别的数据数量不要有数个数量级的差距。

而且还要对数据的量级有一个评估，多少个样本，多少个特征，可以估算出其对内存的消耗程度，判断训练过程中内存是否能够放得下。如果放不下就得考虑改进算法或者使用一些降维的技巧了。如果数据量实在太大，那就要考虑分布式了。

3 特征预处理与特征选择

良好的数据要能够提取出良好的特征才能真正发挥效力。

特征预处理、数据清洗是很关键的步骤，往往能够使得算法的效果和性能得到显著提高。归一化、、缺失值处理、去除共线性等，数据挖掘过程中很多时间就花在它们上面。这些工作简单可复制，收益稳定可预期，是机器学习的基础必备步骤。

筛选出显著特征、摒弃非显著特征，需要机器学习工程师反复理解业务。这对很多结果有决定性的影响。特征选择好了，非常简单的算法也能得出良好、稳定的结果。这需要运用特征有效性分析的相关技术，如相关系数、平均互信息、条件熵、后验概率、逻辑回归权重等方法。

4 训练模型与调优

直到这一步才用到我们上面说的算法进行训练。现在很多算法都能够封装成黑盒供人使用。但是真正考验水平的是调整这些算法的（超）参数，使得结果变得更加优良。这需要我们对算法的原理有深入的理解。理解越深入，就越能发现问题的症结，提出良好的调优方案。

5 模型结果判断

如何确定模型调优的方向与思路呢？这就需要对模型进行诊断的技术。

过拟合、欠拟合 准确度和误差是模型诊断中至关重要的一步。常见的方法如交叉验证。过拟合的基本调优思路是增加数据量，降低模型复杂度。欠拟合的基本调优思路是提高特征数量和质量，增加模型复杂度。

误差分析 也是机器学习至关重要的步骤。通过观察误差样本，全面分析误差产生误差的原因:是参数的问题还是算法选择的问题，是特征的问题还是数据本身的问题……

诊断后的模型需要进行调优，调优后的新模型需要重新进行诊断，这是一个反复迭代不断逼近的过程，需要不断地尝试， 进而达到最优状态。

15、如何解决梯度膨胀

（1）梯度膨胀：

根据链式法则，如果每一层神经元对上一层的输出的偏导乘上权重结果都小于 1 的话，那么即使这个结果是 0.99，在经过足够多层传播之后，误差对输入层的偏导会趋于 0。

可以采用 ReLU 激活函数有效的解决梯度消失的情况。

（2）使用优化的梯度下降算法

adam 等

16、简单说下有监督学习和无监督学习的区别

有监督学习：对具有标记的训练样本进行学习，以尽可能对训练样本集外的数据进行分类预测。（LR,BP,RF）

无监督学习：对未标记的样本进行训练学习，比发现这些样本中的结构知识。(KMeans,DL)

17、了解正则化么？

正则化是针对过拟合而提出的，以为在求解模型最优的是一般优化最小的经验风险，现在在该经验风险上加入模型复杂度这一项（正则化项是模型参数向量的范数），并使用一个 rate 比率来权衡模型复杂度与以往经验风险的权重，如果模型复杂度越高，结构化的经验风险会越大，现在的目标就变为了结构经验风险的最优化，可以防止模型训练过度复杂，有效的降低过拟合的风险。

19、线性分类器与非线性分类器的区别以及优劣

如果模型是参数的线性函数，并且存在线性分类面，那么就是线性分类器，否则不是。

常见的线性分类器有：LR，单层感知机、线性回归。

常见的非线性分类器：决策树、RF、带有激活函数多层感知机。

线性分类器速度快、编程方便，但是可能拟合效果不会很好。

非线性分类器编程复杂，但是效果拟合能力强。

20、简单说说贝叶斯定理。

在引出贝叶斯定理之前，先学习几个定义：

条件概率（又称后验概率）就是事件 A 在另外一个事件 B 已经发生条件下的发生概率。条件概率表示为 P(A|B)，读作“在 B 条件下 A 的概率”。

比如，在同一个样本空间 Ω 中的事件或者子集 A 与 B，如果随机从 Ω 中选出的一个元素属于 B，那么这个随机选择的元素还属于 A 的概率就定义为在 B 的前提下 A 的条件概率，所以：P(A|B) = |A∩B|/|B|

联合概率表示两个事件共同发生的概率。A 与 B 的联合概率表示为　　或　　。

边缘概率（又称先验概率）是某个事件发生的概率。边缘概率是这样得到的：在联合概率中，把最终结果中那些不需要的事件通过合并成它们的全概率，而消去它们（对离散随机变量用求和得全概率，对连续随机变量用积分得全概率），这称为边缘化（marginalization），比如 A 的边缘概率表示为 P(A)，B 的边缘概率表示为 P(B)。

接着，考虑一个问题：P(A|B)是在 B 发生的情况下 A 发生的可能性。

1）首先，事件 B 发生之前，我们对事件 A 的发生有一个基本的概率判断，称为 A 的先验概率，用 P(A)表示；

2）其次，事件 B 发生之后，我们对事件 A 的发生概率重新评估，称为 A 的后验概率，用 P(A|B)表示；

3）类似的，事件 A 发生之前，我们对事件 B 的发生有一个基本的概率判断，称为 B 的先验概率，用 P(B)表示；

4）同样，事件 A 发生之后，我们对事件 B 的发生概率重新评估，称为 B 的后验概率，用 P(B|A)表示。

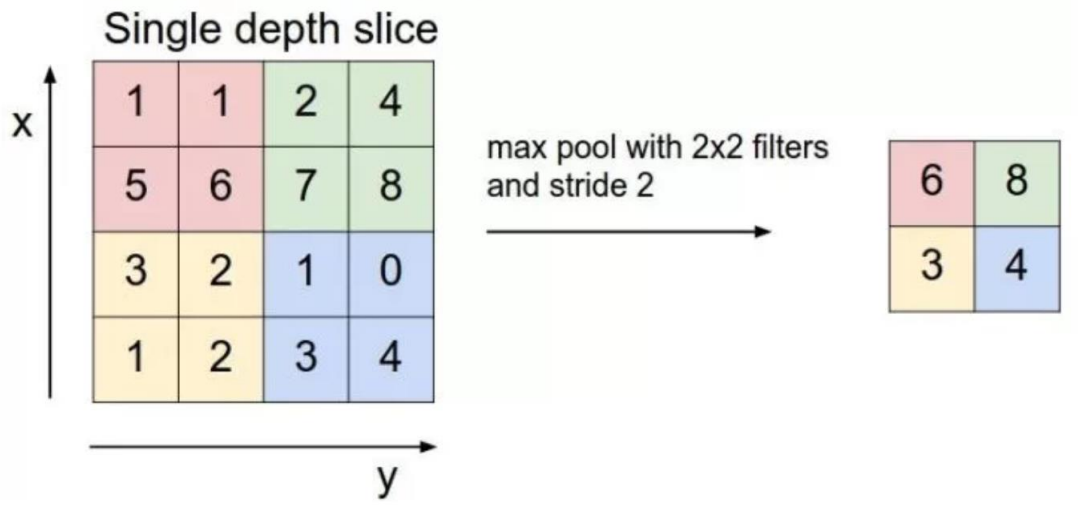
22、什么是卷积?

对图像（不同的数据窗口数据）和滤波矩阵（一组固定的权重：因为每个神经元的多个权重固定，所以又可以看做一个恒定的滤波器 filter）做**内积**（逐个元素相乘再求和）的操作就是所谓的『卷积』操作，也是卷积神经网络的名字来源。

非严格意义上来讲，下图中红框框起来的部分便可以理解为一个滤波器，即带着一组固定权重的神经元。多个滤波器叠加便成了卷积层。

48.什么是CNN的池化pool层？

池化，简言之，即取区域平均或最大，如下图所示（图引自cs231n）



23、上图所展示的是取区域最大，即上图左边部分中 左上角2x2的矩阵中6最大，右上角2x2的矩阵中8最大，左下角2x2的矩阵中3最大，右下角2x2的矩阵中4最大，所以得到上图右边部分的结果：6 8 3 4。很简单不是？

24、哪些机器学习算法不需要做归一化处理？

概率模型不需要归一化，因为它们不关心变量的值，而是关心变量的分布和变量之间的条件概率，如决策树、rf。需LR,KNN、KMeans 之类的最优化问题就需要归一化。

25、梯度下降法找到的一定是最低点吗？

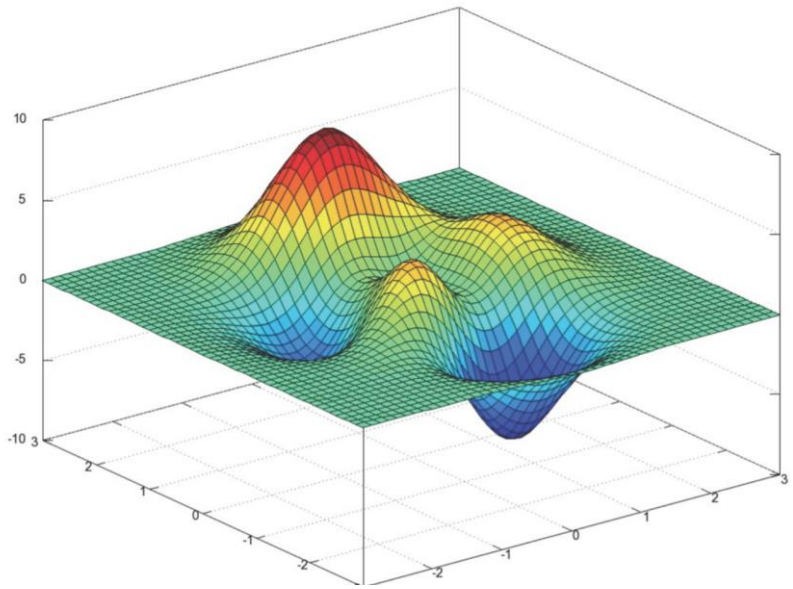
最小二乘法：损失函数：一定只有只有最低点

对数似然损失，交叉熵损失：会有多个局部最低点

会用下山来举例。假设你现在在山顶处，必须抵达山脚下（也就是山谷最低处）的湖泊。但让人头疼的是，你的双眼被蒙上了无法辨别前进方向。换句话说，你不再能够一眼看出哪条路径是最快的下山路径

最好的办法就是走一步算一步，先用脚向四周各个方向都迈出一步，试探一下周围的地势，用脚感觉下哪个方向是下降最大的方向。换言之，每走到一个位置的时候，求解当前位置的梯度，沿着梯度的负方向（当前最陡峭的位置向下）走一步。就这样，每要走一步都根据上一步所在的位置选择当前最陡峭最快下山的方向走下一步，一步步走下去，一直走到我们感觉已经到了山脚。

当然这样走下去，我们走到的可能并不一定是真正的山脚，而只是走到了某一个局部的山峰低处。换句话说，梯度下降不一定能够找到全局的最优解，也有可能只是一个局部最优解。当然，如果损失函数是凸函数，梯度下降法得到的解就一定是全局最优解。



28、什么最小二乘法？

1. 我们口头中经常说：一般来说，平均来说。如平均来说，不吸烟的健康优于吸烟者，之所以要加“平均”二字，是因为凡事皆有例外，总存在某个特别的人他吸烟但由于经常锻炼所以他的健康状况可能会优于他身边不吸烟的朋友。而最小二乘法的一个最简单的例子便是算术平均。

   最小二乘法（又称最小平方法）是一种数学优化技术。它通过最小化误差的平方和寻找数据的最佳函数匹配。利用最小二乘法可以简便地求得未知的数据，并使得这些求得的数据与实际数据之间误差的平方和为最小。用函数表示为：

$$\min_{\vec{x}} \sum_{i=1}^{n} (y_m - y_i)^2.$$

最小二乘使得误差平方和最小，并在各个方程的误差之间建立了一种平衡，从而防止某一个极端误差取得支配地位

- 计算中只要求偏导后求解线性方程组，计算过程明确便捷
- 最小二乘可以导出算术平均值作为估计值

29、说说常见的损失函数

最小二乘法：损失函数：一定只有只有最低点

对数似然损失，交叉熵损失：会有多个局部最低点

31、常见一些图像模型？

图像分类：CNN

图像检测：fast-RCNN， yolo, ssd

选择题：

1 将原始数据进行集成、变换、维度规约、数值规约是在以下哪个步骤的任务？(C)

A. 频繁模式挖掘　　B. 分类和预测　　C. 数据预处理　　D. 数据流挖掘

2 当不知道数据所带标签时，可以使用哪种技术促使带同类标签的数据与带其他标签的数据相分离？(B)

A. 分类　　B. 聚类　　C. 关联分析　　D. 隐马尔可夫链

3 以下哪种方法不属于特征选择的标准方法？(D)

A 嵌入　B 过滤　　C 包装　　D 抽样

4Nave Bayes 是一种特殊的 Bayes 分类器,特征变量是 X,类别标签是 C,它的一个假定是:()

A.各类别的先验概率 P(C)是相等的

B.以 0 为均值，sqr(2)/2 为标准差的正态分布

C.特征变量 X 的各个维度是类别条件独立随机变量

D.P(X|C)是高斯分布

正确答案：C

@BlackEyes_SGC：朴素贝叶斯的条件就是每个变量相互独立。

5 假定某同学使用 Naive Bayesian（NB）分类模型时，不小心将训练数据的两个维度搞重复了，那么关于 NB 的说法中正确的是：

A.这个被重复的特征在模型中的决定作用会被加强

B.模型效果相比无重复特征的情况下精确度会降低

C.如果所有特征都被重复一遍，得到的模型预测结果相对于不重复的情况下的模型预测结果一样。

D.当两列特征高度相关时，无法用两列特征相同时所得到的结论来分析问题

E.NB 可以用来做最小二乘回归

F.以上说法都不正确

正确答案：BD

NB 的核心在于它假设向量的所有分量之间是独立的。在贝叶斯理论系统中，都有一个重要的条件独立性假设：假设所有特征之间相互独立，这样才能将联合概率拆分

6 以下哪些方法不可以直接来对文本分类？

A、Kmeans

B、决策树

C、支持向量机

D、KNN 正确答案:

A 分类不同于聚类。

A：Kmeans 是聚类方法，典型的无监督学习方法。分类是监督学习方法，BCD 都是常见的分类方法。

7 输入图片大小为 200×200，依次经过一层卷积（kernel size 5×5，padding 1，stride 2），pooling（kernel size 3×3，padding 0，stride 1），又一层卷积（kernel size 3×3，padding 1，stride 1）之后，输出特征图大小为： 95

96

97

98

99

100

正确答案：C

计算尺寸不被整除只在 GoogLeNet 中遇到过。卷积向下取整，池化向上取整。

本题 （200-5+2*1）/2+1 为 99.5，取 99

（99-3）/1+1 为 97

（97-3+2*1）/1+1 为 97

研究过网络的话看到 stride 为 1 的时候，当 kernel 为 3 padding 为 1 或者 kernel 为 5 padding 为 2 一看就是卷积前后尺寸不变。

计算 GoogLeNet 全过程的尺寸也一样。

8 假定某同学使用 Naive Bayesian（NB）分类模型时，不小心将训练数据的两个维度搞重复了，那么关于 NB 的说法中正确的是：

- 这个被重复的特征在模型中的决定作用会被加强
- 模型效果相比无重复特征的情况下精确度会降低
- 如果所有特征都被重复一遍，得到的模型预测结果相对于不重复的情况下的模型预测结果一样。
- 当两列特征高度相关时，无法用两列特征相同时所得到的结论来分析问题
- NB 可以用来做最小二乘回归
- 以上说法都不正确

决策树中不包含以下哪种结点 （）

A. 根结点

B. 内部结点

C. 叶结点

D. 外部结点

下面有关分类算法的准确率，召回率，F1 值的描述，错误的是?

A. 准确率是检索出相关文档数与检索出的文档总数的比率，衡量的是检索系统的查准率

B. 召回率是指检索出的相关文档数和文档库中所有的相关文档数的比率，衡量的是检索系统的查全率

C. 正确率、召回率和 `F` 值取值都在0和1之间，数值越接近0，查准率或查全率就越高

D. 为了解决准确率和召回率冲突问题，引入了F1分数