Why is the Bayes' Classifier called Naıve? What is the Naıve assumption made by the classifier?
Because A Naive Bayes' classifier naively assumes that each of the descriptive features in a domain is conditionally independent of all of the other descriptive features, given the state of the target feature. However, in reality the conditional independence does not exist.

Give the formula for Laplacian Smoothing. Explain why it is used
$P(f=v|t)$ = (count(f = v|t) + k)/(count (f |t) + (k × |Domain(f)|))
Smoothing takes some of the probability from the events with lots of the probability share and gives it to the other probabilities in the set.

假定训练样本很大时，每个分量x的计数加1造成的估计概率变化可以忽略不计，但可以方便有效的避免零概率问题。

应用举例

假设在文本分类中，有3个类，C1、C2、C3，在指定的训练样本中，某个词语K1，在各个类中观测计数分别为0，990，10，K1的概率为0，0.99，0.01，对这三个量使用拉普拉斯平滑的计算方法如下：

1/1003 = 0.001，991/1003=0.988，11/1003=0.011

Explain the idea of inductive bias in machine learning. What is the inductive bias present in the kNN method of classification?
the set of assumptions that define the model selection criteria of an ML algorithm. The inductive bias underpinning similarity-based classification is that things that are similar (i.e., instances that have similar descriptive features) belong to the same class.

Precision gives us the proportion of True Positives among True and False Positives. Why do we then need to calculate Recall? Give an example
recall = TP/(TP+FN) 在两个类别的样本数量差距很大时

Give the formula for Jaccard index. List two domains where you would recommend its usage and explain why?

$$sim(p,q) = \frac{|B_p \cap B_q|}{|B_p \cup B_q|}$$

The accuracy of a new diagnostic test for a particular rare disease is 90%. The incidence of disease in the population is 0.1% A patient, J, has tested positive for the disease. Give the formula for Bayes' Theorem. What is the probability that J actually has the disease?
$P(t) = P(t|d)P(d) + P(t|\neg d)P(\neg d)$=0.9*0.001+0.999*0.1(检测是阳性的概率(有病|测对了+没病|测错了))
P(d|t) = $P(t|d)P(d)$/P(t) = 0.9*0.001/(0.9*0.001+0.999*0.1)

The ID3 algorithm recursively builds a decision tree. Under what conditions does it stop and construct a leaf node?
1.All of the instances in the dataset have the same classification (target feature value) then return a leaf node tree with that classification as its label.
2.The set of features left to test is empty then return a leaf node tree with the majority class of the dataset as its classification.

3.The dataset is empty return a leaf node tree with the majority class of the dataset at the parent node that made the recursive call.

What is meant by binning? Describe two ways of binning data. Which one would you choose on a dataset relating age to income? Why?
classify continuous features
two ways of binning data: equal-width　　equal-frequency
equal-frequency

Explain the difference(s) between k-means and k nearest neighbor algorithm

- The k nearest neighbours model predicts the target level with the majority vote from the set of k nearest neightbors to the query **q**:

$$\mathbb{M}_k(\mathbf{q}) = \underset{l \in levels(t)}{\text{argmax}} \sum_{i=1}^{k} \delta(t_i, l) \qquad (4)$$

- **Goal:** Minimise distances between the items and their nearest centroid - i.e. minimisation of *sum-of-squared error* (SSE):

$$SSE(C) = \sum_{c=1}^{k} \sum_{x_i \in C_c} D(x_i, \mu_c)^2 \quad \text{where} \quad \mu_c = \frac{\sum_{x_i \in C_c} x_i}{|C_c|}$$

- In the standard algorithm, *D* is measured using Euclidean distance:

$$D(x, \mu) = \sqrt{\sum_{l=1}^{m} (x_l - \mu_l)^2}$$

sum of squared difference over all *m* feature values

- *k*-Means tries to reduce SSE via a two step iterative process:
  1) Reassign items to their nearest cluster centroid
  2) Update the centroids based on the new assignments
- Repeatedly apply these two steps until the algorithm converges to a final result.

Explain the F1-Measure and how it is calculated? What kinds of machine learning algorithm are they most useful to evaluate? In what kind of domains would you use it?
F_1-measure = 2 × (precision × recall) / (precision + recall) 作为一个单一metric比较多个模型吧

What is the Receiver Operating Characteristic Curve? How does it help with evaluating a classification method? Explain with reference to the reference line.

A receiver operating characteristic curve, or ROC curve, is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied.
The ROC curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings.

The k-Means clustering technique is very sensitive to the value of k. How can we convince ourselves that the chosen value of k is good? What other clustering method would you recommend if the user continually wants to change the levels of granularity at which clusters are shown?

1. Compute clustering algorithm (e.g., k-means clustering) for different values of k. For instance, by varying k from 1 to 10 clusters. 为不同的 k 值计算聚类算法(例如，K平均算法)，例如，通过改变 k 从1到10个聚类
2. For each k, calculate the total within-cluster sum of square (wss). 对于每个 k，计算簇内平方和(wss)的总和
3. Plot the curve of wss according to the number of clusters k. 根据集群 k 的数量绘制 wss 曲线
4. The location of a bend (knee) in the plot is generally considered as an indicator of the appropriate number of clusters. 一般认为，地块中弯曲(拐点)的位置是适当的集群数量的指示器

1. compute clustering algorithm for different values of k. For instance, by varying k from 1 to 10 clusters.
2. For each k, calculate the total within-cluster sum of square(wss)
3. plot the curve of wss according to the number of cluster k
4. the location of a bend(knee) in the plot is generally considered as an indicator of the appropriate number of cluster.

You have a medical dataset containing symptoms for which patients have tested either positive or negative (binary values). While trying to find similar patients, what metric would you use. Why? Give the formula used for the metric that you recommend.
Euclidean distance

Why does the kNN classification encounter difficulties when the input data has an extremely high number of features?
(是因为knn都是在坐标系里面画，没办法建立太多元的n-dimensional space, if we build too many dimensions, the performance will decrease dramatically)

Why is Machine-Learning considered an ill-posed problem?
适定问题是指定解满足下面三个要求的问题：① 解是存在的；② 解是唯一的；③ 解连续依赖于定解条件，即解是稳定的。这三个要求中，只要有一个不满足，则称之为不适定问题

What is the Analytics Base Table? Describe its structure.
The basic structure in which we capture historical datasets is the analytics base table (ABT) The general structure of an **analytics base table**—descriptive features and a target feature.

What is a data quality issue, in the context of an ABT? List three common data quality issues and give strategies for dealing with any two?
anything *unusual* about the data in an ABT.

missing values — 1. drop any features having missing values 2. apply imputation
irregular cardinality
outliers —**clamp transformation** that clamps all values above an upper threshold and below a lower threshold to these threshold values, thus removing the offending outliers

Hold-out sampling is easy to understand and makes intuitive sense. However, it is not perfect. Give two reasons which may cause hold-out sampling to result in a poor evaluation of your machine learning technique. Suggest a technique that tries to address these issues.
切数据的时候切的不整齐，比如把上半年全部切进training 把下半年全部切进了validation 可以用k fold解决这个问题

1.hold-out method hugely depends on how data is split，应该多做几次
2.hold-out method may let some data's information never be used/trained

A metric that critically affects Decision Trees is the feature selection metric. List two ways of feature selection and their respective formula?

$$H(t) = -\sum_{i=1}^{I}(P(t=i) \times log_s(P(t=i)))$$

Computing information gain involves the following 3 equations:

$$H(t,\mathcal{D}) = -\sum_{l \in levels(t)}(P(t=l) \times log_2(P(t=l))) \quad (2)$$

$$rem(d,\mathcal{D}) = \sum_{l \in levels(d)} \underbrace{\frac{|\mathcal{D}_{d=l}|}{|\mathcal{D}|}}_{\text{weighting}} \times \underbrace{H(t,\mathcal{D}_{d=l})}_{\substack{\text{entropy of} \\ \text{partition } \mathcal{D}_{d=l}}} \quad (3)$$

$$IG(d,\mathcal{D}) = H(t,\mathcal{D}) - rem(d,\mathcal{D}) \quad (4)$$

entropy information gain

为什么梯度下降可以不需要负号：因为in the error surface，we need go down，to get the aim point就是往错误最少的方向走
error surface is continuously differentiable and convex
when the slope of the gradient is 0 we have reached the minima of the error surface
When the partial derivative is non-zero we want to move in the opposite direction of the gradient, hence we multiple the entire term by minus.

The fundamentals of similarity-based learning are:
- Feature space
- Similarity metrics

Bayesian networks use a graph-based representation to encode the structural relationships —such as direct influence and conditional independence—between subsets of features in a domain.
Consequently, a Bayesian network representation is generally more compact than a full joint distribution, yet is not forced to assert global conditional independence between all descriptive features.