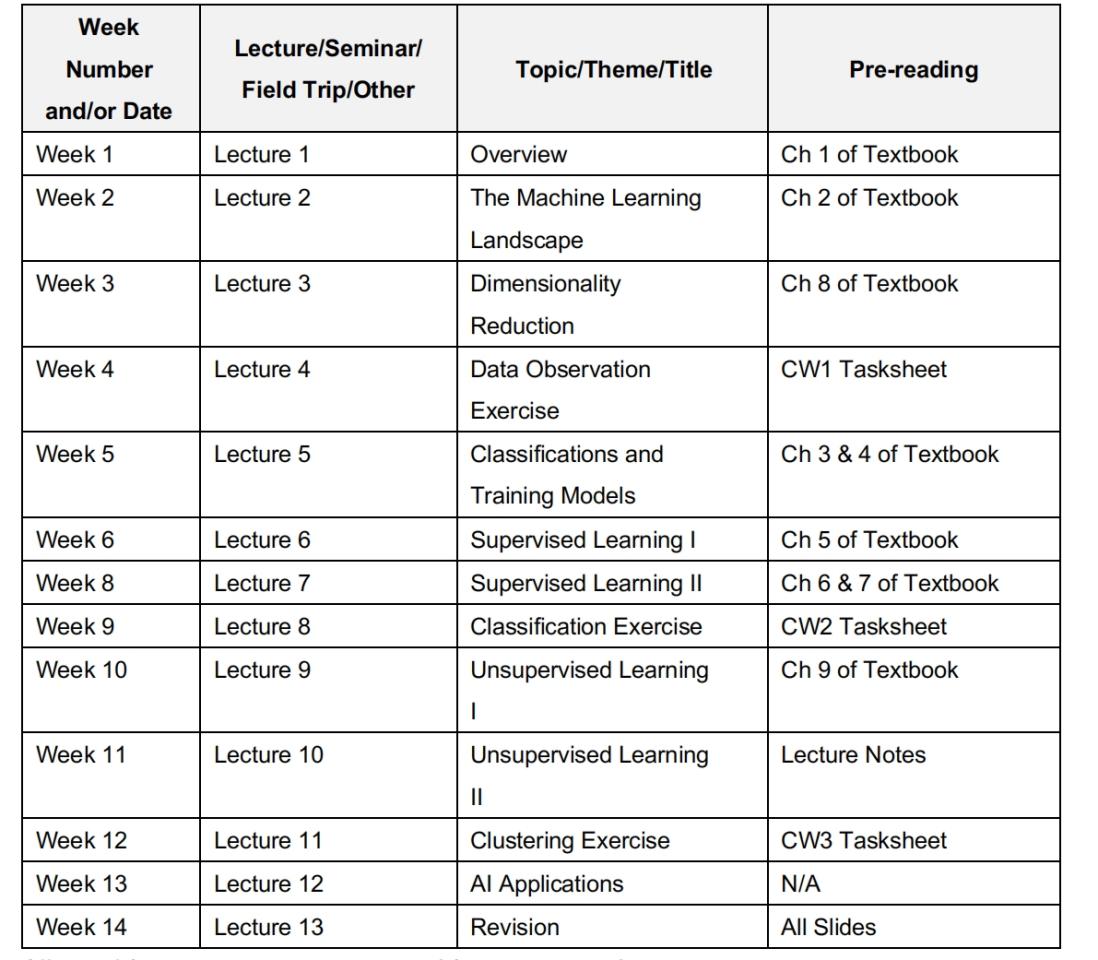
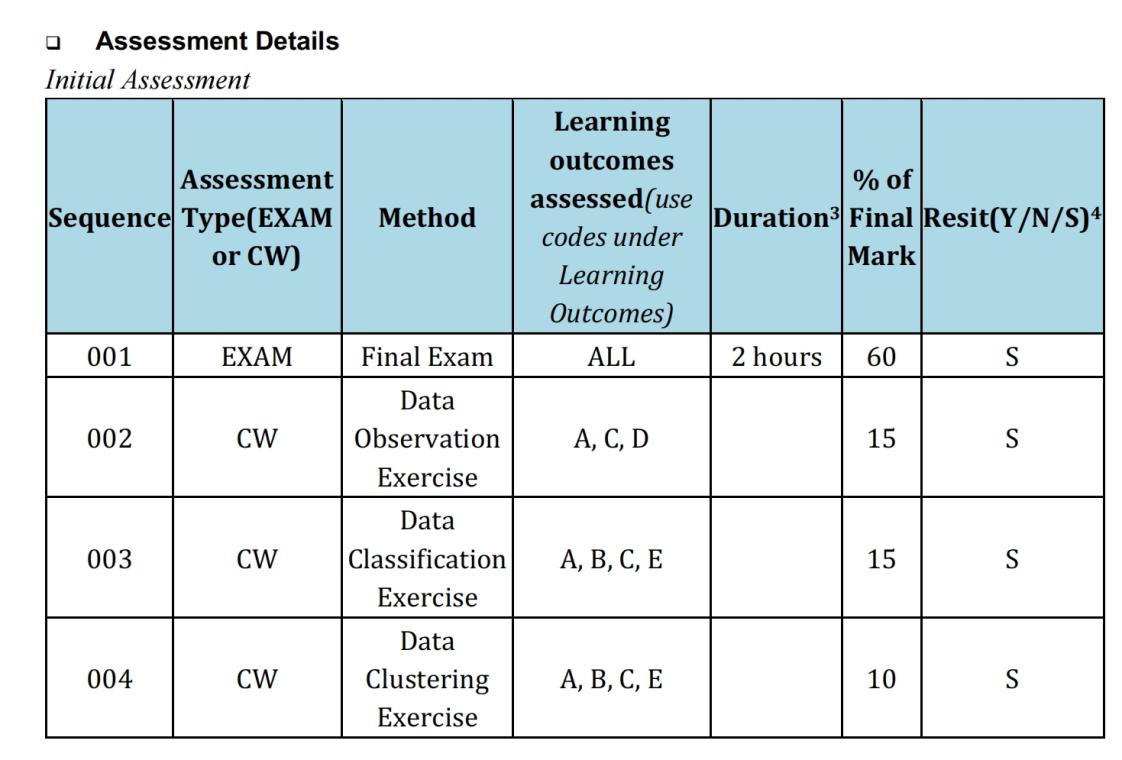
***INT104: Introduction to Artificial Intelligence***

*1 Introduction*

*1.1 Teaching Plan*

**

*1.2 Assessment*

**

*About Final exam*

*----Open book exam (2 hours)*

*---- MCQ: 54 Marks*

*---- Filling blanks: 24 Marks*

*---- Computation: 14 Marks*

*---- Essay Question: 8 Marks*

*1.3 Machine Learning : If a computer program improves its performance on task T measured by score function P , then we say its has learned some experience E.*

*1.3.1 Types of ML*

*Supervised Learning -- label available in training*

*Unsupervised Learning -- label unavailable in training*

*Semi-supervised Learning -- label partially available in training*

*Reinforcement Learning -- Use (s,a,s’,r ) from Bellman to maximize*

*1.3.2 How to select a model with good performance on task T ?*

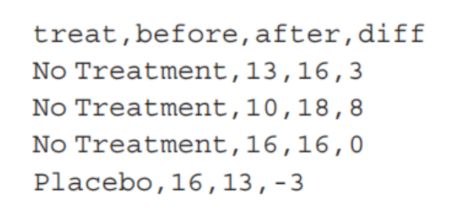
*Train: Training Datasets /Testing Datasets*

*Score: Validation Datasets----Training datasets should not overlap with validation datasets*

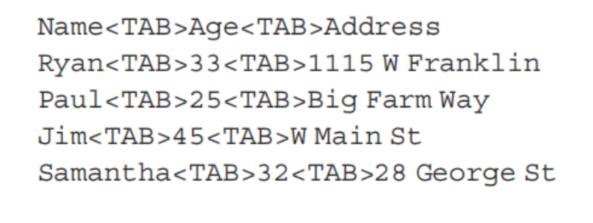
*2 Data preprocessing*

*2.1 Data Storage and Presentation*

*2.1.1 CSV (Comma Separated Values)* 逗号作为字段之间的分隔符

**

*2.1.2 TSV (Tab Separated Values)* Tab作为字段之间的分隔符

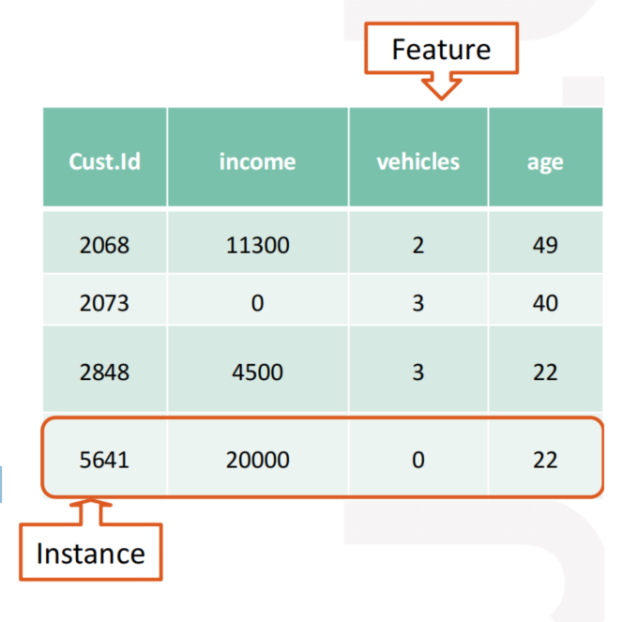
**

*2.1.3 XML (Extensible Markup Language)* 使用标签来标识数据，并且具有层级结构，可以表示复杂的数据关系。XML 可以通过定义自定义标签来适应不同的数据结构，但相对于 JSON，在表示数据时更为冗长。

*2.1.4 JSON (JavaScript Object Notation):* 使用键值对的方式来存储数据，并支持数组和嵌套对象。

*2.2 Data Preprocessing*

*Datasets: A dataset is a collection of instances, features, and target variables that are used to train and test machine learning models.*

**

*The following is the data preprocessing step in order.*

*2.2.1 Data Cleaning and Data Integration*

*• remove the corresponding instance.*

*• remove the whole column.*

*• set missing values to some value (zero, the mean, the median, etc.).*

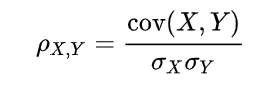
*• remove the outliers* 异常值

*• solve the consistency in data structure and format*

*2.2.2 Feature Selection*

*--- Filter Methods: features are selected and ranked according to their correlations with the target .*

*Pearson’s r correlation: ( no direct relationship with information entropy )*

**

*---- Wrapper Methods: use experiments to search for well-performing combinations of features*

*---- Embedded Methods: do feature selection when the model is training*

*2.2.3 Feature Extraction*

文本数据：对文本数据进行词袋模型、TF-IDF（词频-逆文档频率）计算等方式进行特征提取，也可以利用词嵌入（如Word2Vec等）将文本转换为密集向量表示。

图像数据：使用卷积神经网络（CNN）提取图像的特征，通常通过预训练的模型（如VGG、ResNet等）获取图像的特征向量。

时序数据：对时序数据进行滑动窗口统计、傅里叶变换等方式提取时序特征。

音频数据：对音频数据进行短时傅里叶变换（STFT）、梅尔频率倒谱系数（MFCC）提取声学特征。

*2.2.4 Data Transformation*

*• Handling text and categorical attributes Eg: [‘cat1’], [‘cat2’], [‘cat3’], [‘cat4’]*

*---- Ordinal encoder : [0], [1], [2], [3]*

*---- One-hot encoder : [1,0,0,0], [0,1,0,0], [0,0,1,0], [0,0,0,1]* 消除了类别之间的大小顺序关系，每个类别都被表示为一个独立的维度，避免模型错误地认为类别之间存在顺序关系

*• Normalization*

*---- Min/max normalization*

*x’ =*

*---- Z-score normalization (Normalizing every value in a dataset such that the mean of*

*all of the new values is 0 and the standard deviation is 1 )*

*x’ =*

*---- Normalization by decimal scaling*

*X’ =*

*2.2.5 Data Reduction*

*• Data with high dimensions:*

*---- High computational complexity : In high-dimensional spaces, the distances between data points become sparser, which means more data is needed to represent the structure of the space. For example, in high-dimensional space, the calculation of Euclidean distance becomes more difficult and more samples are needed to keep the data representative.*

*---- May contain many irrelevant or redundant features*

*---- Difficulty in visualization*

*---- With high risk of getting an overfit model*

*• Projection: Data is not spread out uniformly across all dimensions. (All the data*

*lies within (or close to) a much lower-dimensional subspace of the high-dimensional space.*

*• Principal Component Analysis (PCA): it identifies the axis that accounts for the largest amount of variance in the training set. (As the larger the variance is , the classification is easier to identify.)*

*---- How to calculate the variance on Principal C1 ?*

*From the vector projection ,we know*

*where c and x are vectors. c is the direction vector on Principal C1 , N\*1 matrix, and is the features of X at i row, N\*1 matrix . and have been processed by = - .Then we can get :*

*=*

*=*

*=*

*=*

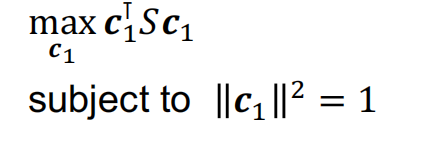
*= s.t. = 1 (M can be ignored as a scalar in optimize)*

*Let S = which is a N\*N matrix if is N\*1 matrix . X = [ ..... ]*

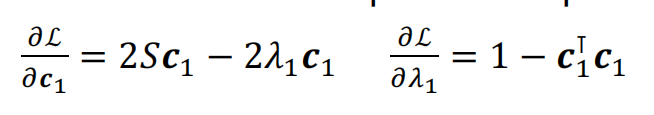
*,which is N\*M. And the S is called data covariance matrix.*

*---- Use Lagrange equation to optimize the problem.*

拉格朗日乘数:要求f(x,y) 在g(x,y)=0 存在极值,对L(x,y, λ) = f(x,y) - λ\*g(x,y)求梯度

**

*1711531375424*

**

*1711531403791*

*Then*

*=*)

*=* λ as λ is a 1\*1 scalar

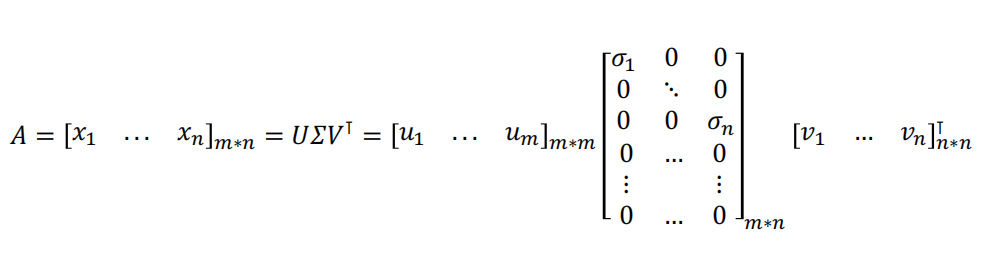
= λ ()

= λ

----为什么次优的就是通过 = λ 的第二最大值 ？

在主成分分析（PCA）中，"主成分"是指原始数据集在新坐标系统中的坐标轴，这些坐标轴按照数据变化的方向排序。第一主成分是数据变化最大的方向，也就是说，数据在这个方向上的投影有最大的方差。当我们计算协方差矩阵的特征值和特征向量时，最大的特征值对应的特征向量就是第一主成分的方向。这个特征向量指向的方向捕获了数据变化最大的部分。由于特征向量是协方差矩阵的正交基，这意味着其他的特征向量将会与第一主成分正交（即它们在数学上是垂直的）。所以，一旦我们确定了第一主成分，下一步就是选择第二主成分。**第二主成分也应该捕获数据中尽可能多的变化，但它不能是与第一主成分相同的方向。**因为我们已经从数据中去除了第一主成分的变化，**我们要找的第二主成分是在所有与第一主成分正交的方向中，方差最大的方向**。在PCA中，我们按照特征值的大小顺序（从大到小）来选择主成分，每个新选出的主成分都是在剩余方向上方差最大的方向，同时也与所有之前选出的主成分正交。

*---- Introduction to Singular Value Decomposition:*

**

*A = U \* Σ \* , A is a M\*N matrix*

*Where*

*( A \* ) \* = \* , is a M\*1 matrix*

*( \* A) \* = \* , is a N\*1 matrix*

*U = [ ...... ]*

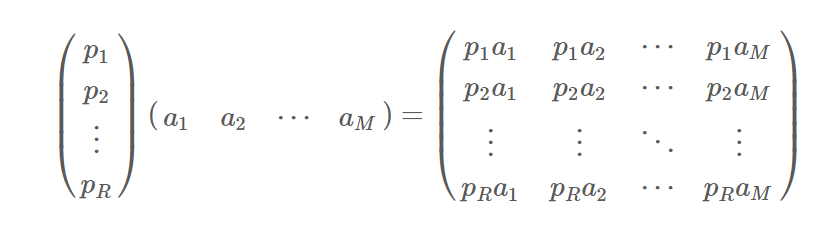
*V= [ ...... ]*

*So U \* = I , V \* = I*

Σ 是一个M\*N的矩阵，Σ除了对角线其它元素都为0，对角线上的元素称为奇异值

IMG_256

*----基变换的矩阵表示*

**

一般的，如果我们有M个N维向量，想将其变换为由R个N维向量表示的新空间中，那么首先将R个基按行组成矩阵A，然后将向量按列组成矩阵B，那么两矩阵的乘积AB就是变换结果，其中AB的第m列为A中第m列变换后的结果.**特别要注意的是，这里R可以小于N，而R决定了变换后数据的维数。也就是说，我们可以将一N维数据变换到更低维度的空间中去，变换后的维度取决于基的数量。因此这种矩阵相乘的表示也可以表示降维变换。**

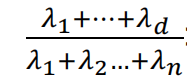
*This tell as why we need to write X as [ ..... ]*

*----How to reduce dimension in PCA ?*

*From Lagrange equation, we know we can get the principal c that have max variance by calculate*  = λ . as the difficulty to directly get all the thing by this formula, we can use S*ingular Value Decomposition instead , which can ignore* λ computation.

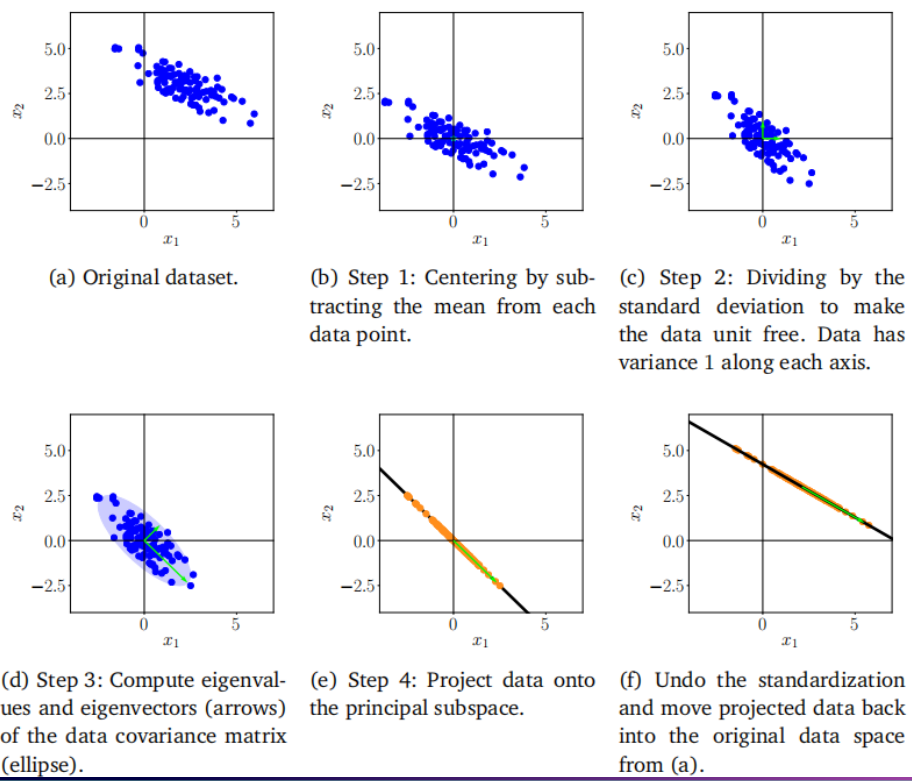
*1711566594382*

*---- Explained Variance Ratio*

**

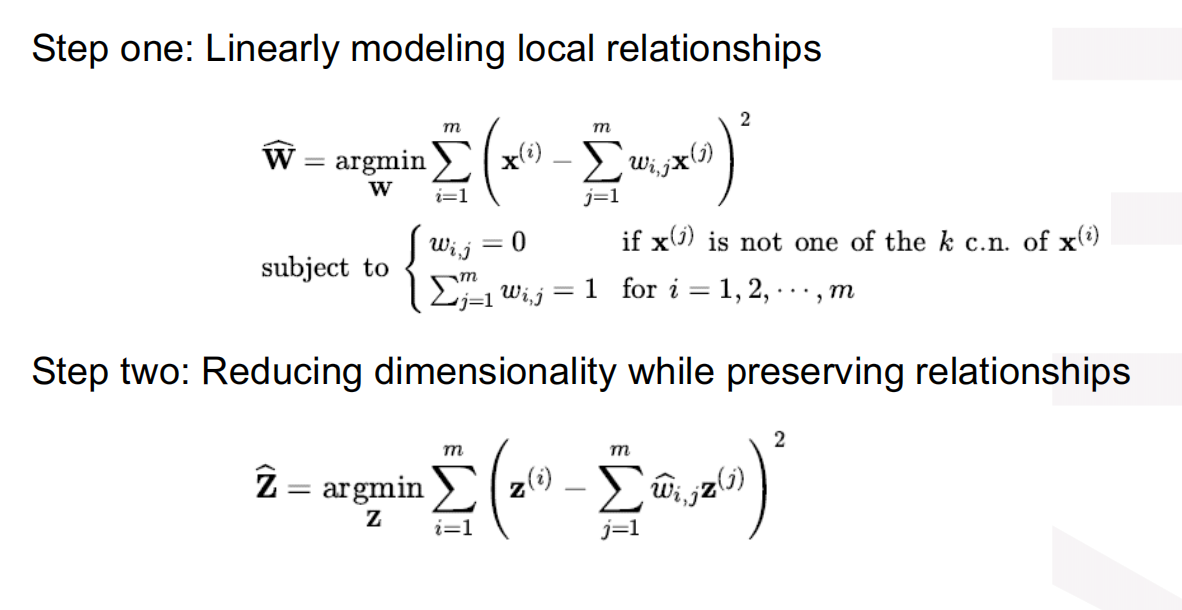
*通过前 d个主成分解释a%的原始数据方差。在PCA中，特征值 λ 表示与对应特征向量方向相关的方差量。特征值越大，表示数据在该特征向量（即主成分）方向上的分散程度越大。*

*---- Summary*

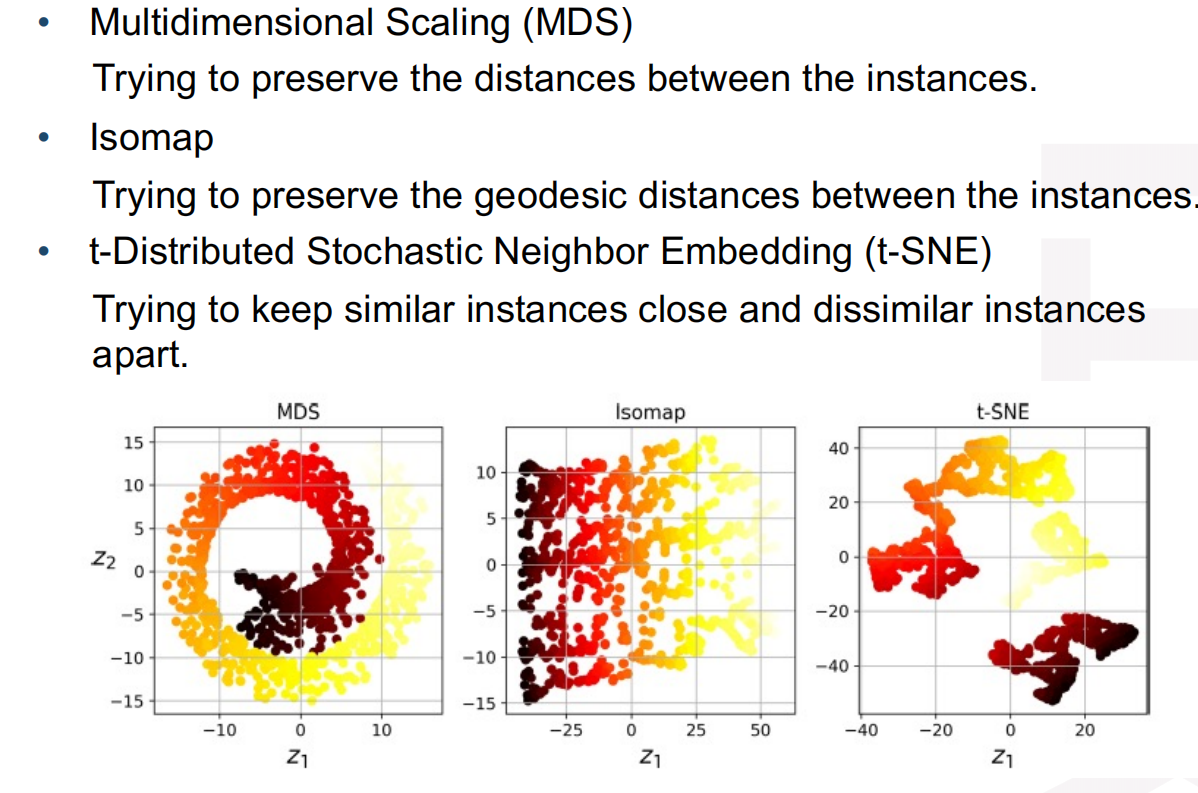
**

*• Manifold Learning :一类旨在发现数据在高维空间中的低维结构的算法。在机器学习和数据科学中，Manifold learning 是一种非线性降维技术.“Manifold”这个术语来源于数学，它指的是一个局部像欧几里得空间的空间。在机器学习中，这意味着尽管数据可能存在于高维空间中（例如有很多特征的空间），但数据的实际变化可以通过一个低维空间来表示，即数据点可以被嵌入到一个低维的流形中。例如，你可以想象一张二维的纸，它在三维空间中可以被弯曲和扭曲，但不管怎样变化，纸的表面本身仍然是二维的。类似地，在高维数据分析中，即使数据存在于一个很大的特征空间中，它们也可能沿着某种较低维度的表面或曲线分布。Manifold learning 的关键假设是数据是均匀分布在这样一个流形上的，而这个流形在更高维的空间中可能会有弯曲或者扭曲。算法的目标就是要揭示和利用这个低维流形的结构来实现数据降维或其他任务。常见的Manifold learning算法包括LLE, t-SNE, Isomap*

*• Locally Linear Embedding: LLE is a powerful nonlinear dimension reduction (NLDR) technique. It is a Manifold Learning technique that does not rely on projections.*

**

*• Other Techniques*



*3 Naive Bayes classifier*

*Bayes’ Rule：The famous Bayes’ Rule states the relationship between prior probability distribution and posterior probability distribution.*

*P (c |x ) =*

*where c is considered as the class, x is considered as a feature*

*---- P(c) is named as prior probability*

*---- P(x|c) is named as class-conditional probability (CCP, also known as likelihood)*

*---- P(c|x) is named as posterior probability*

*---- P(x) is considered as evidence factor (observation)*

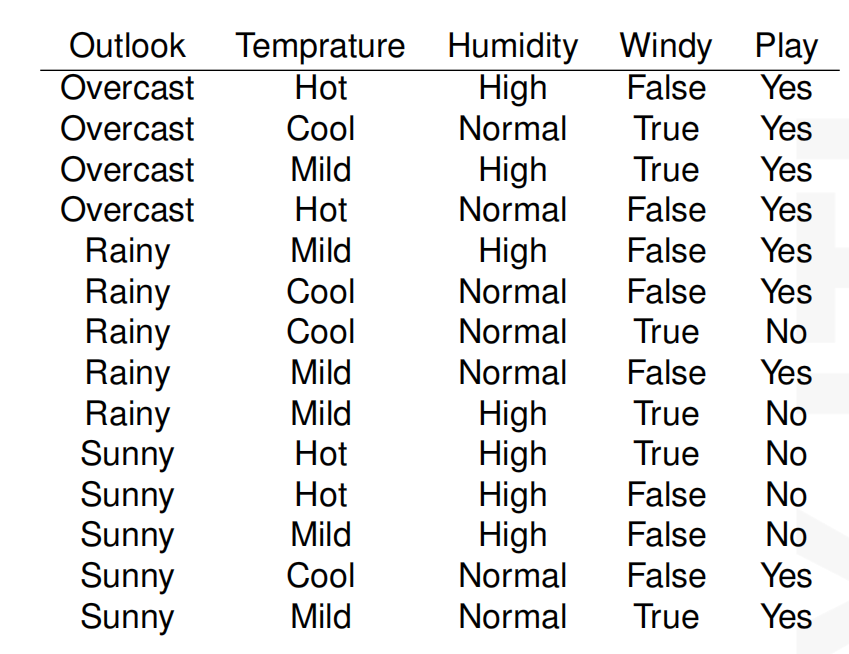
*How can we make use of Bayes’ Rule for Classification? Maximize the posterior probability of observations . This method is named MAP estimation (Maximum a posteriori) .* 注意这里只介绍传统贝叶斯学习，不涉及神经网络，没有参数需要更新，就是对数据集做计算而已。

朴素贝叶斯和一般贝叶斯区别在于假设各个特征间条件独立，如下解释：

*P (c |x ) =*

这个公式中实际上c 是分类，x 是特征 ，但是一些物质特征不只有一个，实际写成 P(c | x1 , x2, ....,xN) . 根据贝叶斯公式，我们需要计算 P(x1 , x2 ,... ,xN|c) 我们一无所知不知道哪些特征是否依赖于其他特征，朴素贝叶斯假设各个特征间条件独立，导致这里直接简化为 P( x1| c) P(x2 |c ) ....P(xN |c ) .

*Exercise :*

**

*1 Will you play on the day of Mild ?*

*2*