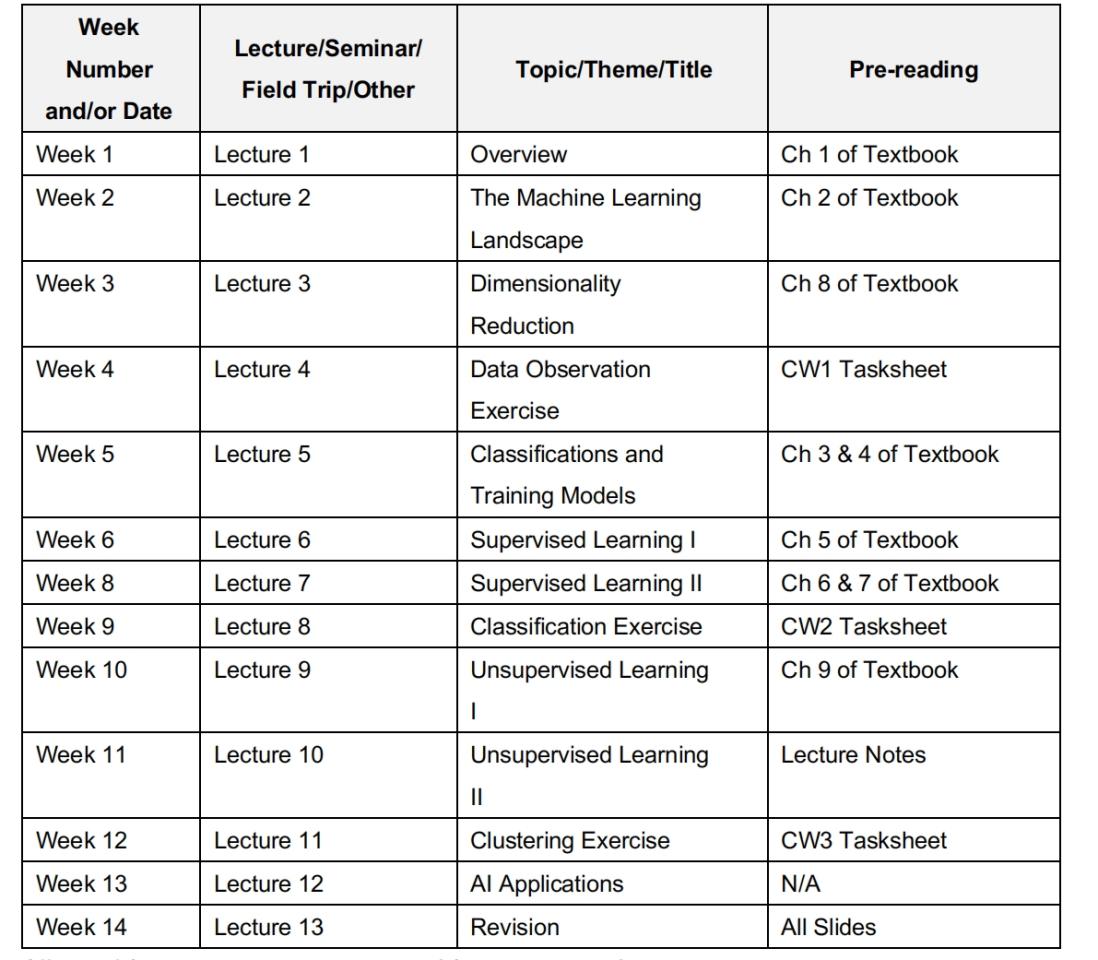
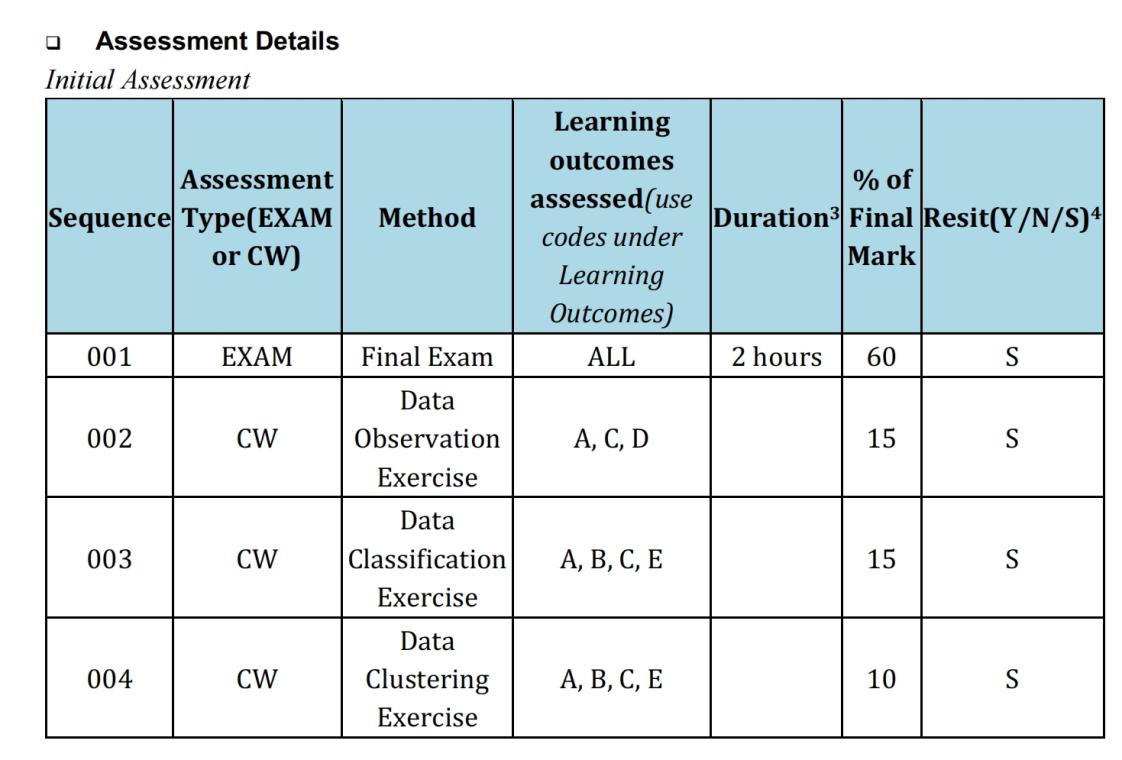
***INT104: Introduction to Artificial Intelligence***

*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 ?*

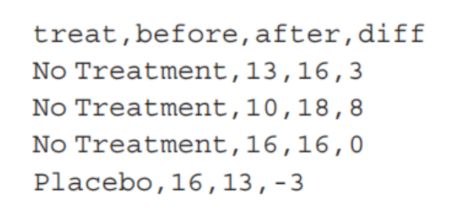
*Train: Training Datasets /Testing Datasets*

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

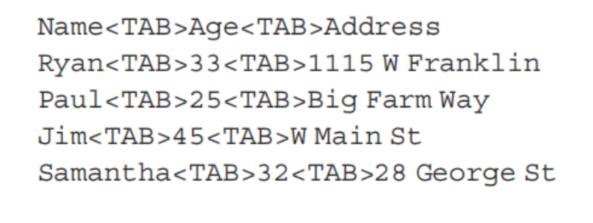
*2 Data preprocessing*

*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

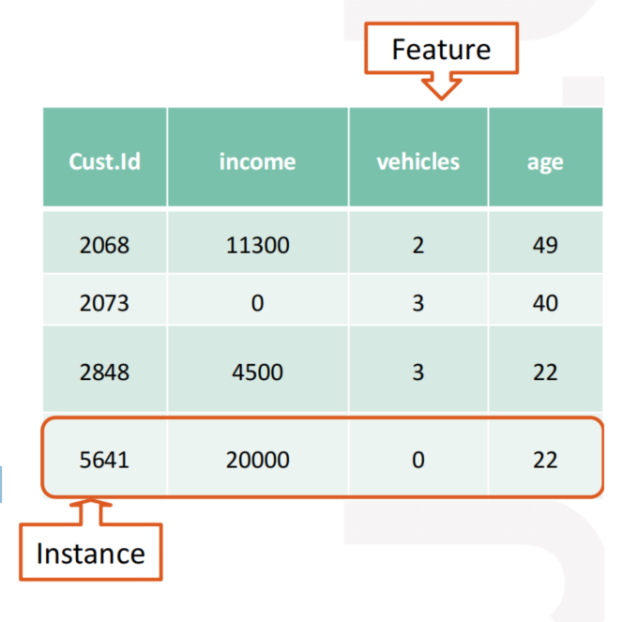
**

*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.).*

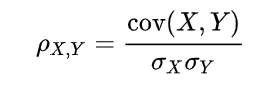
*• remove the outliers* 异常值

*• 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 :*

*=*

*=*

*=*

*=*

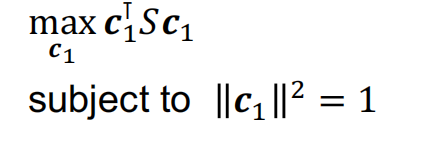
*= s.t. = 1 (M can be ignored as a scalar in optimize)*

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

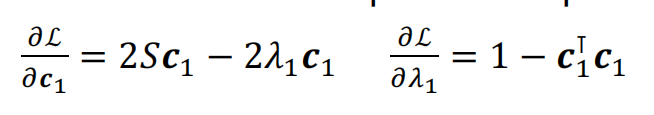
*,which is N\*M. And the S is called data covariance matrix.*

*---- 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

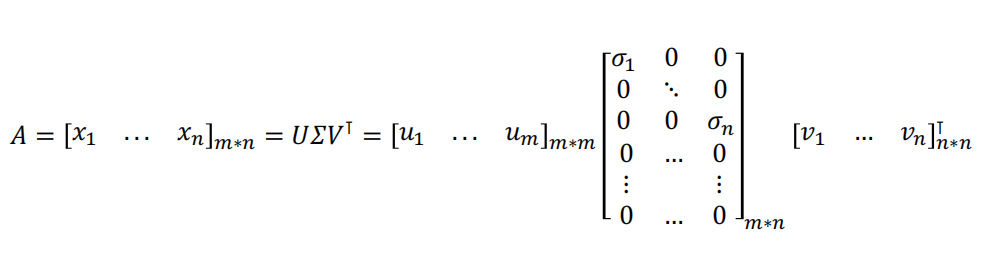
= λ ()

= 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 = [ ...... ]*

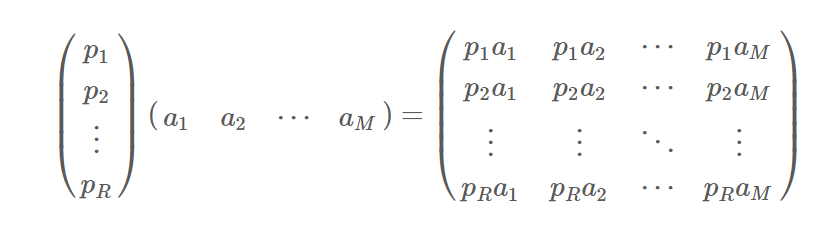
*V= [ ...... ]*

*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

= + / -

*----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。**

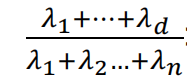
*This tell as why we need to write X as [ ..... ]*

*----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.

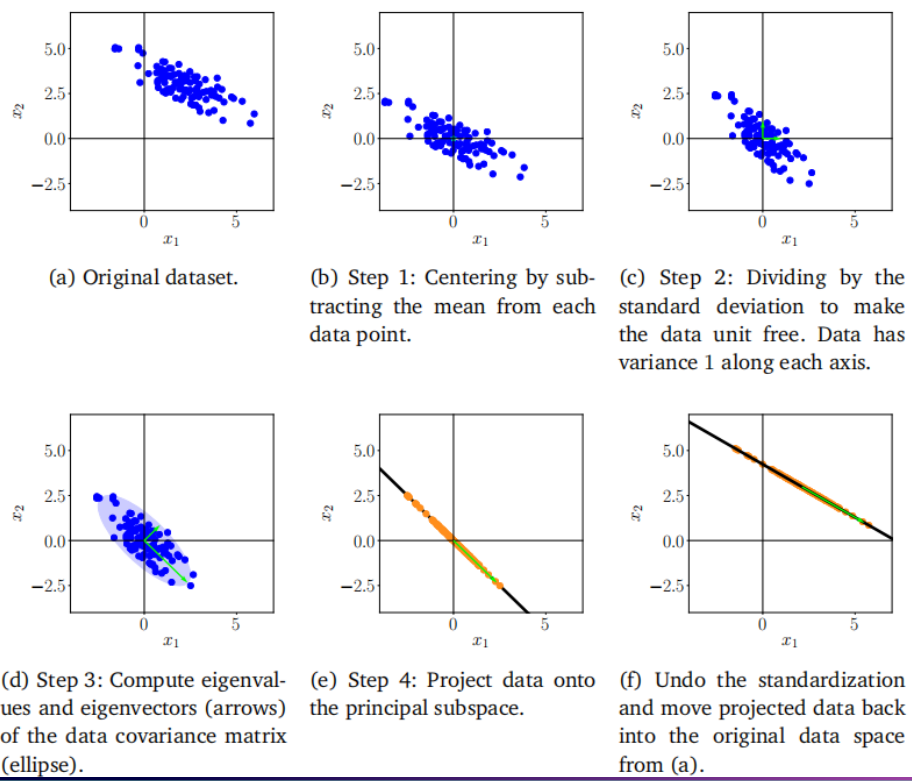
*1711566594382*

*---- 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*

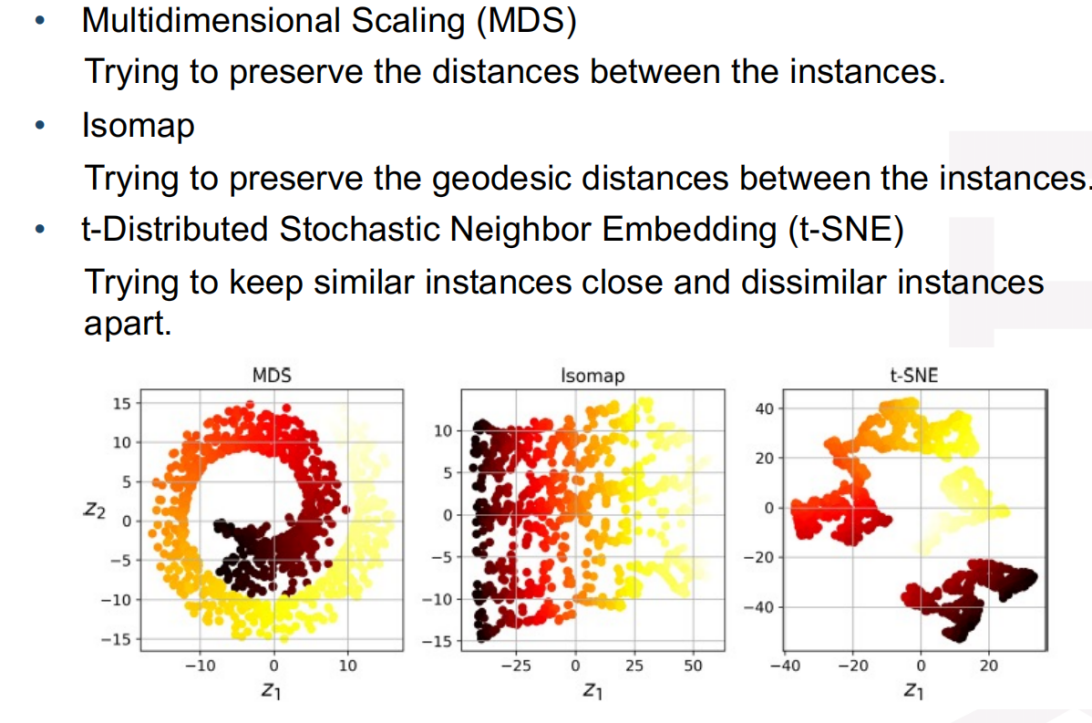
**

*• 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*



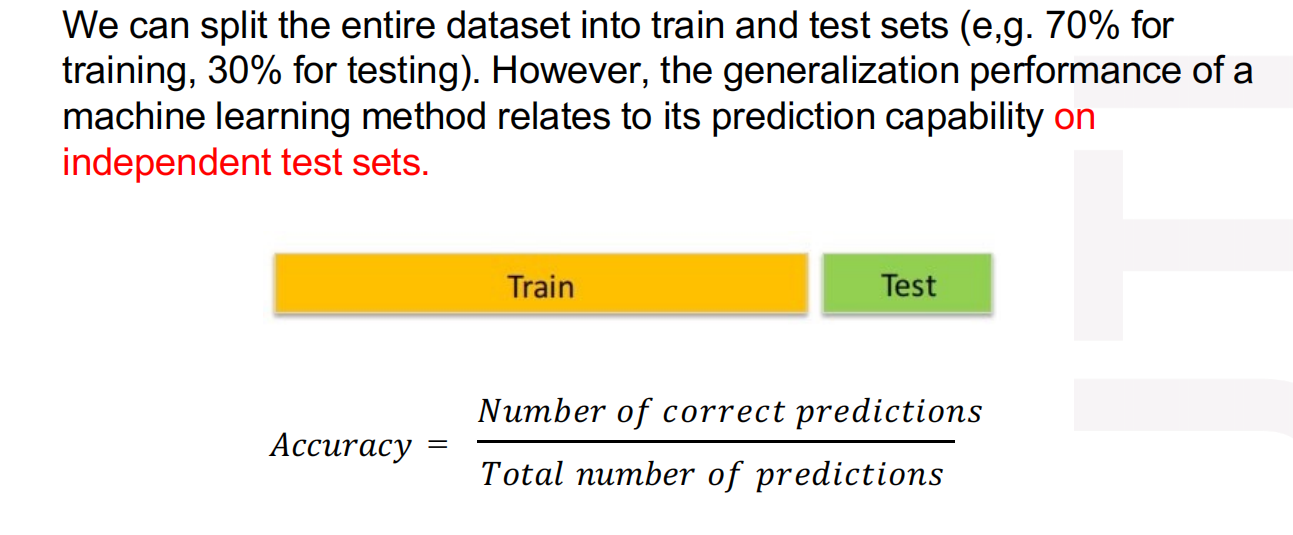
*3 Classification and Training models*

*3.1 Introduction to MINISET Datasets*

*It has 70k handwritten digits from 0-9, including 60k for Train and 10k for Test . Each digits are balanced samples. Each image has 28 x 28 (784) pixels , and each pixel ranges from 0(white) to 255 (black).*

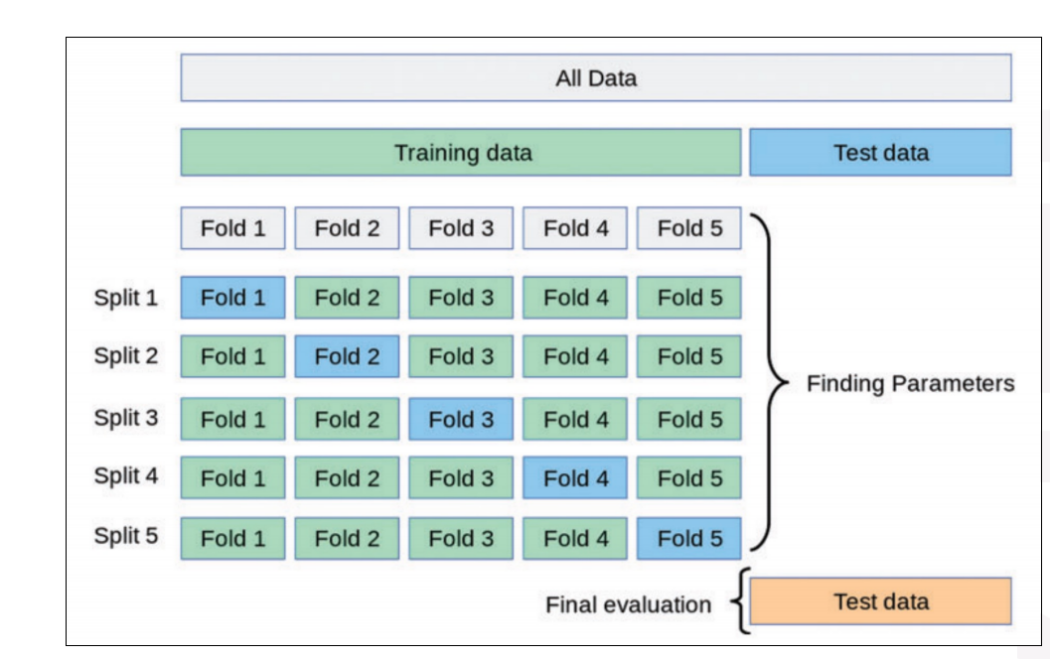
*3.2 Criteria to evaluate Classification models*

*3.2.1 Accuracy*

**

*3.2.2 K-fold Cross-Validation*

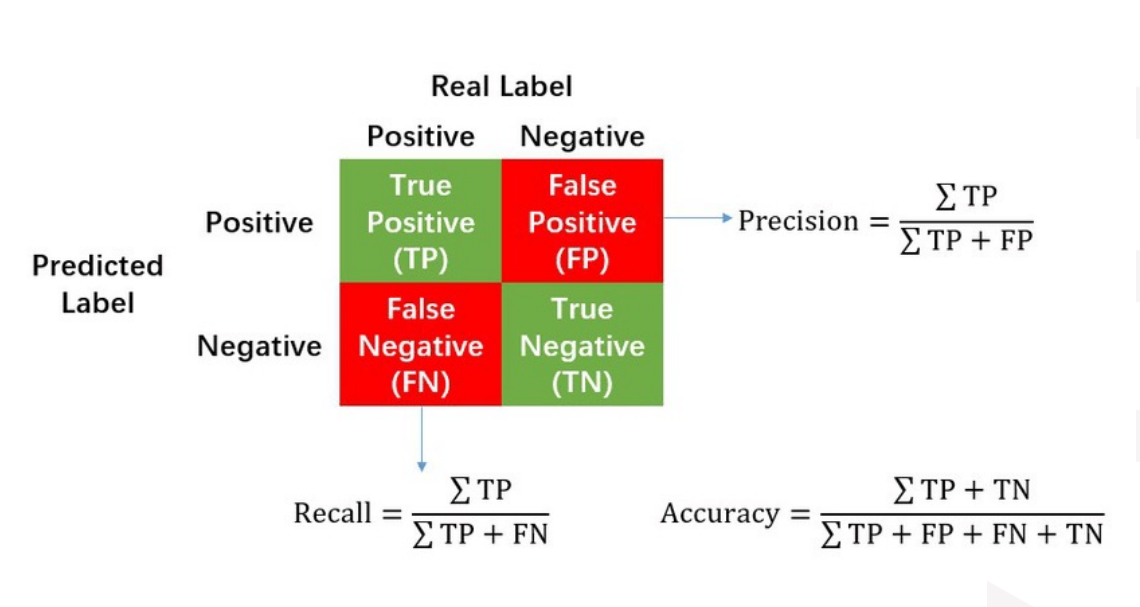
***To avoid selecting the parameters that perform best on the test data but maybe not the parameters that generalize best****, we can further split the training set into training fold and validation fold.*

**

先对数据拆分为训练集和测试集，然后通过再对训练数据的K折划分，其中（K-1） 折用于训练，剩下一折用于验证，这个过程重复K次，并通过获取训练的所有K个模型的平均值和标准差来计算一组特定超参数的模型性能分数，计算给出最优模型的超参数。

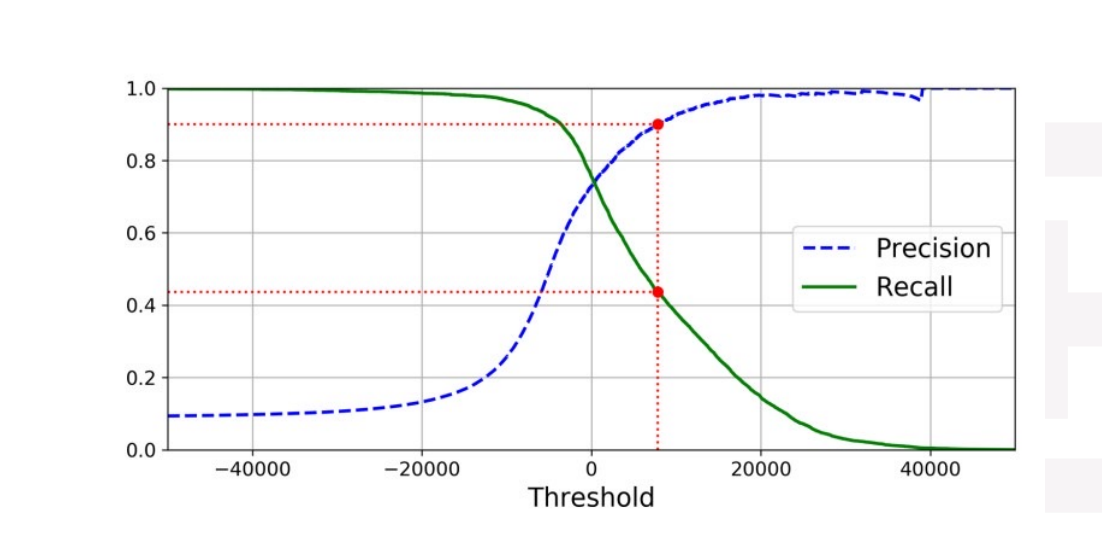
*If the classification task exists class imbalance , which means , for example in MINISET, we want to identify 5, where only10% of our samples were 5.If we set all the result of the classification, we will get a 90% accuracy, which will influence our actual model performance judgment . We use the following criteria to overcome:*

*---- Confusion Matrix (Binary Classification)*

**

在存在类不平衡的情况下，精确率（Precision）和召回率（Recall）特别重要，因为单纯的准确率可能会误导模型的真实性能c精确率越高，意味着模型生成的假正例越少。召回率越高，意味着模型遗漏的正类实例越少。

*---- Precision and Recall with Threshold*

**

分类问题中，模型通常会输出一个概率值，表示一个样本属于正类（通常是标签为1的类）的可能性。阈值是决定这个概率输出如何被解释为最终类别的一个界限。通常，如果模型输出的概率高于阈值，样本就会被划分为正类；如果低于阈值，就会被划分为负类。

如果提高阈值（即需要更高的概率才能将样本划为正类），模型变得更加保守，只有当模型非常确信样本为正类时才会这样预测。这通常会提高精确度，因为它减少了被错误分类为正类的负类样本的数量，但同时可能会降低召回率，因为一些真正的正类样本可能由于概率不足而未被识别。相反，如果降低阈值，模型变得更加宽松，更容易将样本预测为正类。这通常会增加召回率，因为更多的正类样本被正确识别，但同时可能会降低精确度，因为会有更多负类样本被错误地预测为正类。

因此，选择阈值是一个需要在模型预测精确性和覆盖率之间权衡的决策过程。在实际应用中，根据业务需求和成本考虑，你可能需要调整阈值以达到最佳的平衡。

*3.2.3 F1 Score:* ***the harmonic mean* 调和平均值** *of precision and recall. It gives more weight to low values and only get a high F1 score if both are high.*

*=*

*3.2.4 score: a more flexible F score that combines precision and recall*

*=*

*=*

*>1 , then the Precision polynomial can nearly goes to 0 , focuses more on Recall*

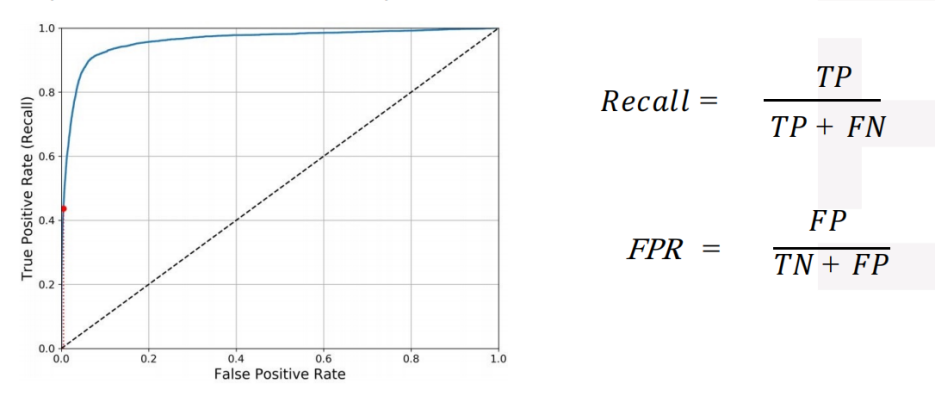
*<1 , then the Recall polynomial can nearly goes to 0 , focuses more on Precision*

*Set F1 , precision, or recall depends on the situation:*

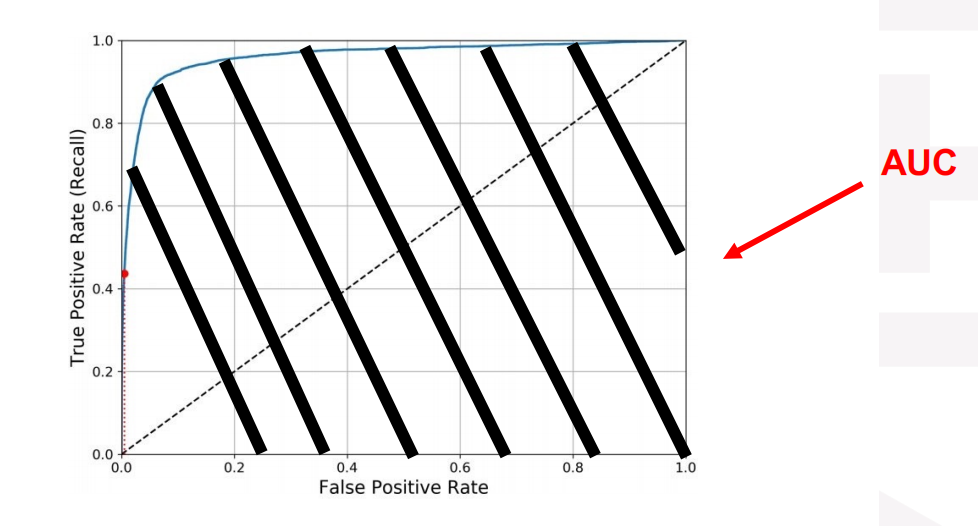
*---- Classifier to detect if videos are safe for kids: Reject many good videos but keep safe one (high threshold -->So we need precision)*

*---- Classifier to detect thieves : May give false positives but captures all thieves (low threshold -->So we need precision)*

*---- Receiver Operation Characteristics (ROC): A ROC curve is a graph showing the performance of a classification model at all classification thresholds. In another word, it presents Recall (True Positive Rate) VS FPR ( False Positive Rate ). The ROC graph summarizes all of the confusion matrices that each threshold produced.*

**

*---- Area Under the Curve(AUC): AUC ranges in value from 0 to 1. A model whose predictions are 100% wrong has an AUC of 0; one whose predictions are 100% correct has an AUC of 1.*

**

AUC的数学意义和统计解释与它作为一个度量的用途是相互关联的。AUC可以被解释为评估模型将正例排在负例前面的概率，数学上等价于在阈值的所有可能值上，模型预测一个随机选取的正例排在一个随机选取的负例之前的概率。这是因为AUC可以从一个叫做曼-惠特尼U统计量的角度来理解，该统计量用于比较两个独立样本的秩次。假设我们有一个正样本集合和一个负样本集合。对于每一对正负样本（一个正样本和一个负样本），如果正样本的预测概率大于负样本的预测概率，那么我们可以认为模型在这一对样本上做出了正确的预测。模型的AUC值是所有可能的正负样本对中，模型预测正确的比例。

*3.3 Multi-class Classification and Multi-label Classification*

*Multi-class classification refers to classification tasks that can distinguish between more than two classes:*

*---- One-versus-the-rest(O v R) strategy: train multiple binary classifiers for each class, select the class whose classifier outputs the highest score.( train N times ,where N means the number of the class)* 例如，如果有四个类别A、B、C和D，那么会有四个分类器，每个分类器分别负责识别“A与非A”、“B与非B”、“C与非C”和“D与非D”.

*---- One-versus-one (O v O) strategy: train a binary classifier for every pair of classes*

*( train N(N-1)/2 times ,where N means the number of the class )* 比如对于类别A、B和C，就会训练三个分类器：一个区分A和B，一个区分A和C，还有一个区分B和C。在分类新样本时，每个分类器对样本进行分类，并给出其决策。最终决策通常是通过投票的方式进行，即每个分类器的决策相当于一票，样本将被归类到得票最多的类别。

*Multi-label classification refers to classification system that outputs multiple binary tags.*

多类分类（Multi-class Classification）

定义：每个实例只能分配到一个类别中。例如，一个给定的水果可以被分类为苹果、香蕉或橙子之一，但不能同时是这几个类别。

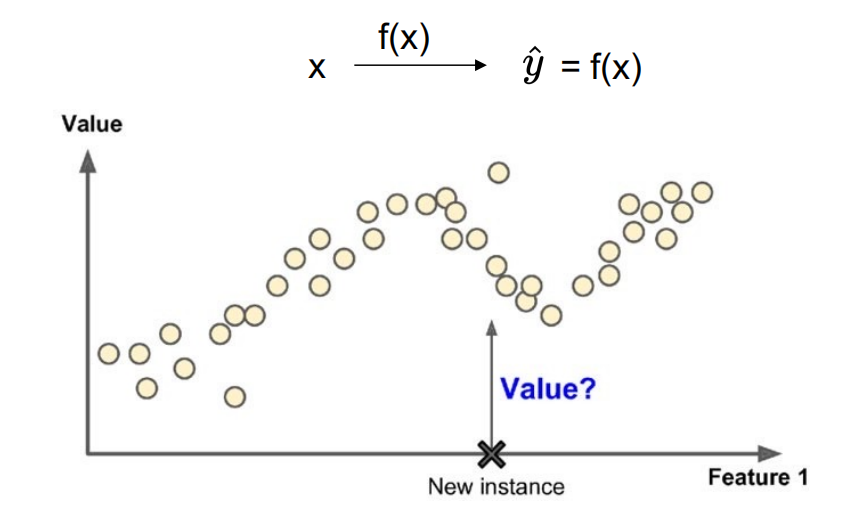
目标：预测单一输出变量的多个类别之一。

多标签分类（Multi-label Classification）

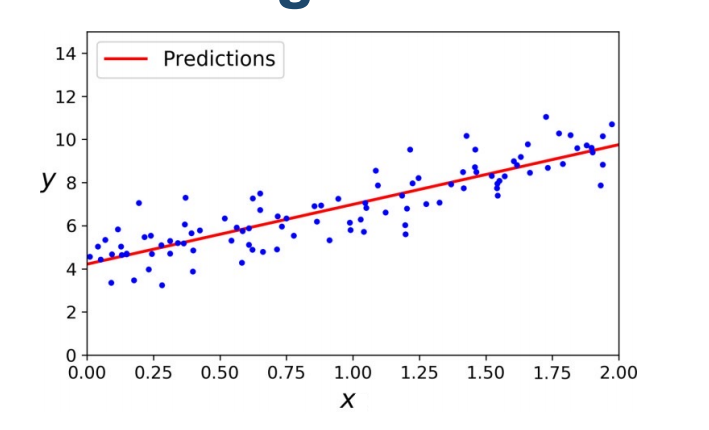
定义：每个实例可以同时拥有多个类别标签。这意味着，例如，一部电影可以同时被标记为动作、喜剧和冒险。

目标：预测可能相关的多个类别标签。

*3.4 Regression*

**

*3.4.1 Simple Linear Regression*

**

*In this term ,we assume that each sample has only one feature and the label of the sample is ,so we can get a list , where the i means one sample from the all and our numbers of data is m.*

*Usually the predicted value (fitted value) is not perfect. The difference between the fitted value and real value y is known as residuals .*

*= k + b*

*= -*

*The regressed value usually pursues a minimum of* ***Residual Sum of Square (RSS):***

*argmin*

*3.4.2 Linear Regression*

*Not like the previous one ,we can have multiple features in one sample .*

*= + + + ..... +*

*=*

*----* ***=***  *the vector of sample i , where are features*

*----* ***=***  *the vector of parameter*

*---- is the predicted value i*

*---- n is the number of the features*

*Loss function:* ***Mean Squared error (MSE)*** *for a Linear Regression model*

*L(θ) =*

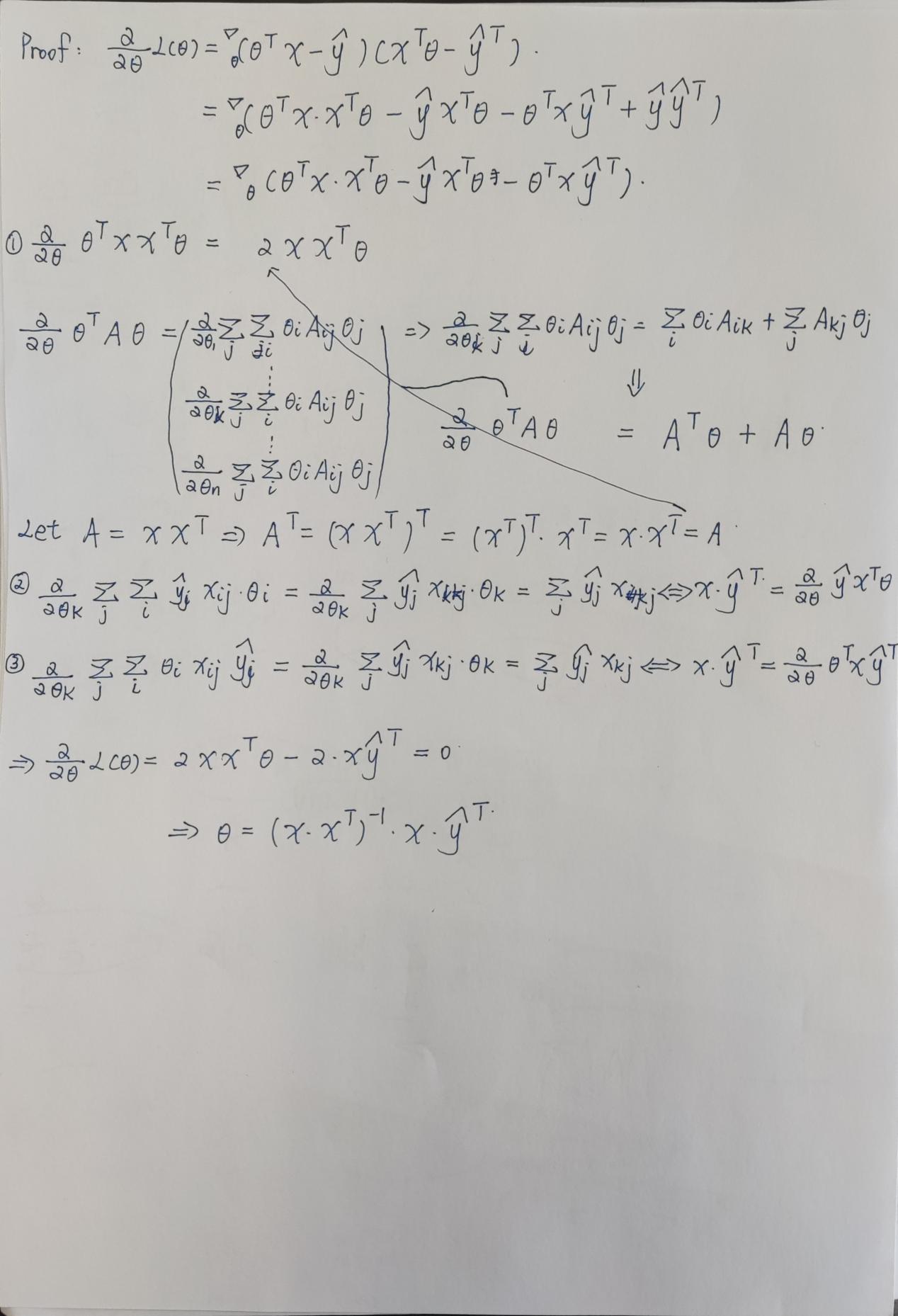
*=*

*---- = [ ..... ]*

*---- = [ ..... ]*

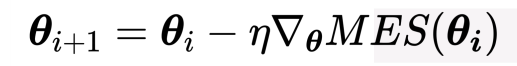
*In order to get the*  ***argmin L(θ)*** *, we can calculate :*

***= 0***

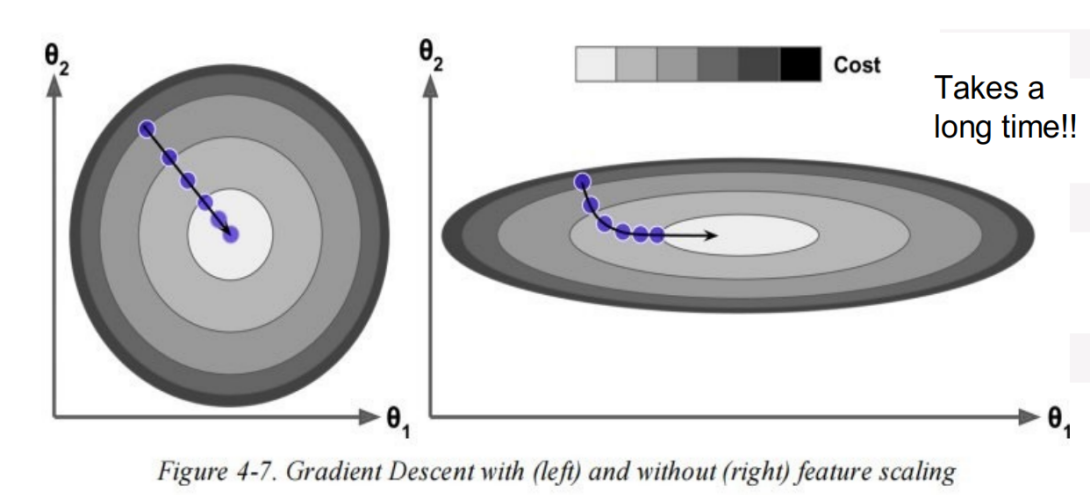
******

*Therefore , we get*  ***=***

*---- Gradient Descent :*

******

*---- Feature scaling before Gradient Descent :*

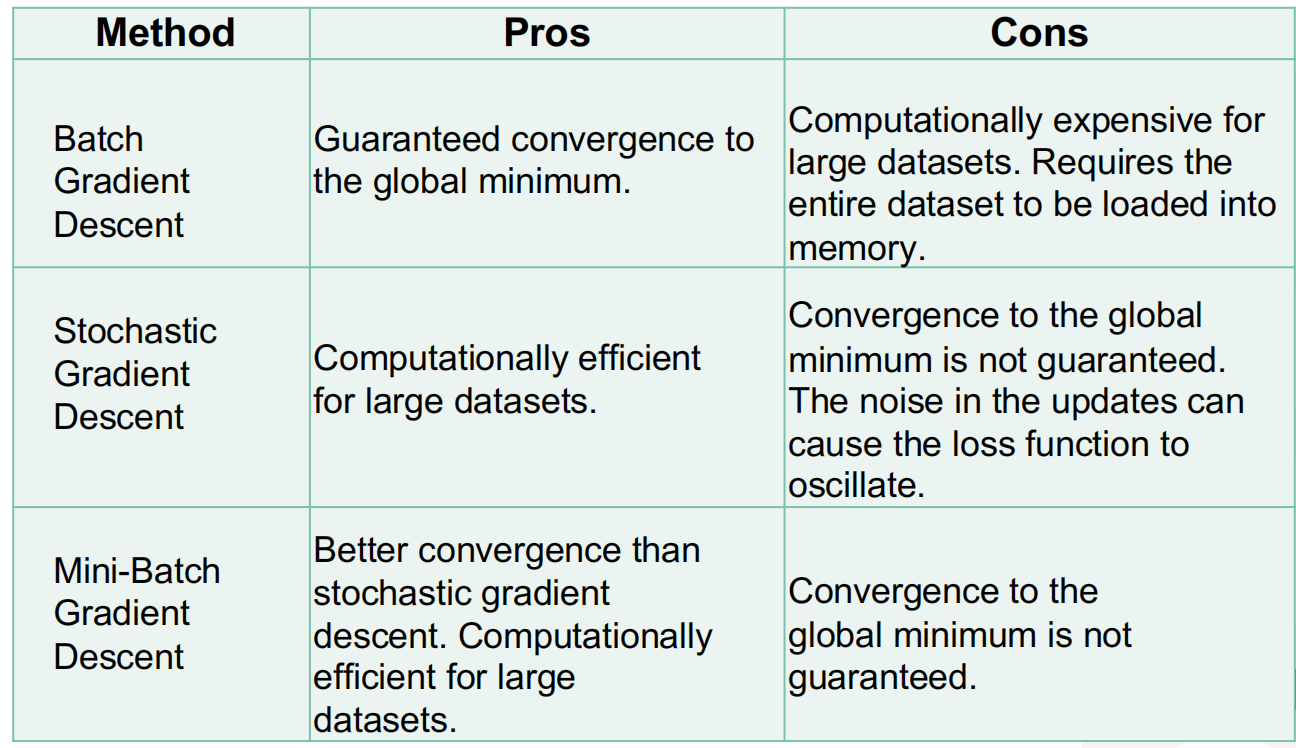
******

*---- Type of Gradient Descent :*

***Batch Gradient Descent (Full Gradient Descent) :*** *Use the whole training set to compute the gradients at every step.*

***Stochastic Gradient Descent*** *picks a random instance from the training set at every step and computes the gradients based only on that single instance.*

***Mini-batch Gradient Descent*** *computes the gradients on small random sets of instances called mini-batches.*

**

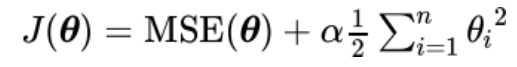
*3.4.3 Polynomial Regression*

**

*随着模型复杂（这里主要指多项式最高项的次数），模型越复杂，能够更好拟合训练的数据，误差自然越来越小，测试集测试的时候小小的波动不能保证模型反应过度导致测试的结果误差反而可能越来越大。*

*3.5 Regularized Linear Models*

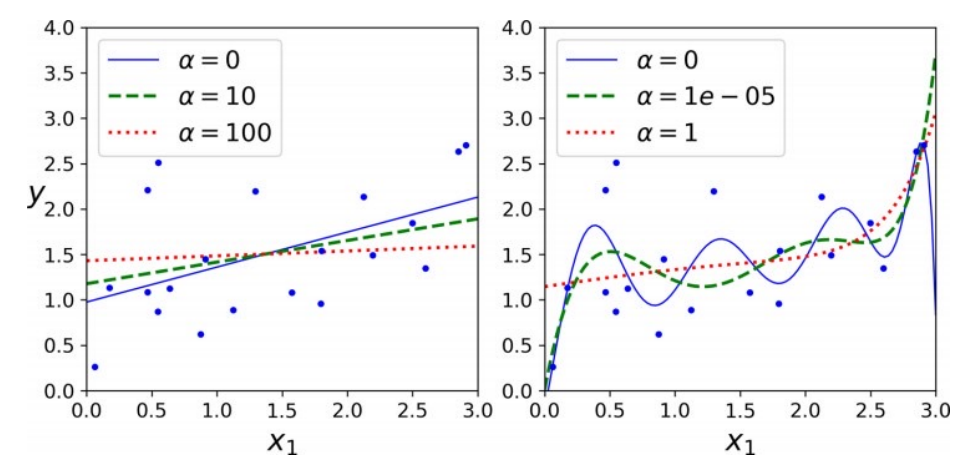
*3.5.1 Ridge Regression(L2):*

**

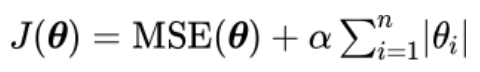
较大的参数值会导致模型在训练数据中非常敏感，即模型会对训练数据的小波动做出较大的反应，这通常是过拟合的一个标志。

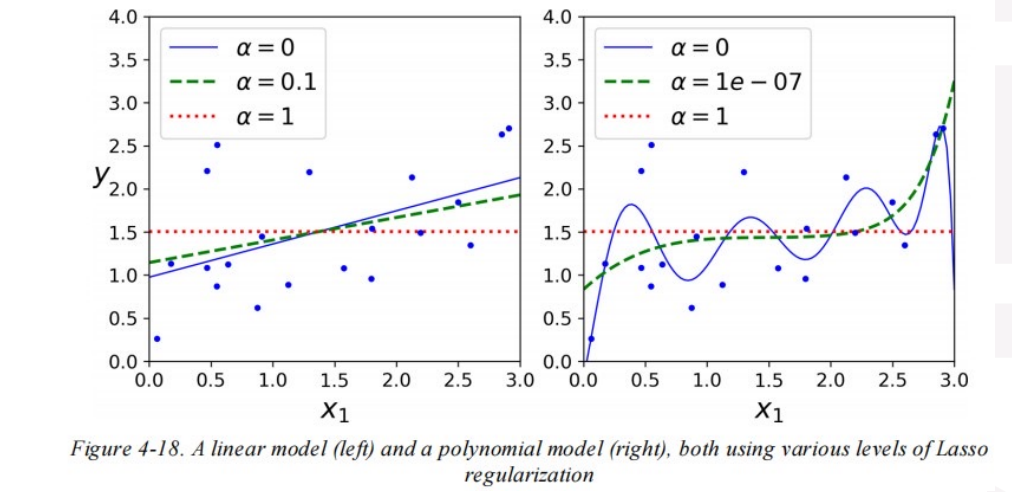
通过减小模型参数的值，正则化倾向于使模型的预测行为更加平滑，不易受到训练数据中随机噪声的影响。这样，即使模型在训练集上的性能略有下降，其在新未见数据上的泛化能力却可能因为减少了过拟合而得到提高。

但是过大的 α 值强化了对参数大小的约束，可能导致欠拟合；较小的 α 值则对参数的约束较弱，可能导致过拟合。



*3.5.2 Lasso Regression(L1):*

**

**

**它具有自动进行特征选择（由于Lasso模型在优化过程中将许多参数压缩至零，因此它实际上对那些对模型预测贡献不大或者是冗余的特征进行了自动筛选）和输出稀疏模型（L1正则化倾向于产生稀疏的参数向量）的能力。**

**比较L1正则化和L2正则化：**L1正则化因其能产生稀疏解而常用于特征选择，特别是在特征数量远多于样本数量的情况下（如某些高维统计场景）。稀疏模型不仅易于解释，还可以减少模型的存储需求和计算成本。L2正则化则因其提供的解是唯一的（L2正则化项是严格凸的）和稳定的而受到青睐，特别是在解决数据中存在多重共线性的问题时。

**L1具有非光滑性**：L1正则化项是参数的绝对值之和，这是一个非光滑函数。在数学优化中，非光滑函数的极值（最小化问题的解）往往出现在不可导的点。对于L1正则化，这些不可导点正是参数为零的地方。这种非光滑性导致在优化过程中，一些参数会直接被设置为零。

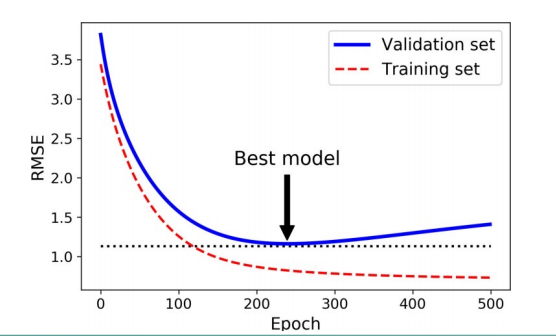
**L2具有光滑性：**L2惩罚项是参数的平方和，这是一个光滑函数。在数学优化中，光滑函数的极值通常出现在可导的点。L2正则化不会导致参数精确为零，因为对于任何非零的参数，其平方都是正的，从而导致非零的梯度。

**L2具有全局惩罚：**L2正则化对所有参数施加全面的惩罚，即使参数值很小，也会因为其平方项存在而被进一步推向零，但永远不会完全达到零。这导致模型参数在所有维度上都被均匀压缩，而不会产生稀疏性。

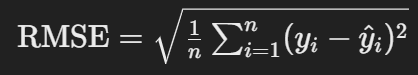
*3.5.3 Elastic Net*

1714577106425

*3.6 Early stopping*



*We stop training as soon as the validation error reaches a minimum. ( RMSE is Root Mean Square Error )*

**

*3.7 Logistic Regression*

*---- Logistic Regression is commonly used to estimate the probability that an instance belongs to a particular class.*

*Sigmoid function :*

*---- Softmax Regression (Multinomial Logistic Regression)*

*=*

*4 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)*

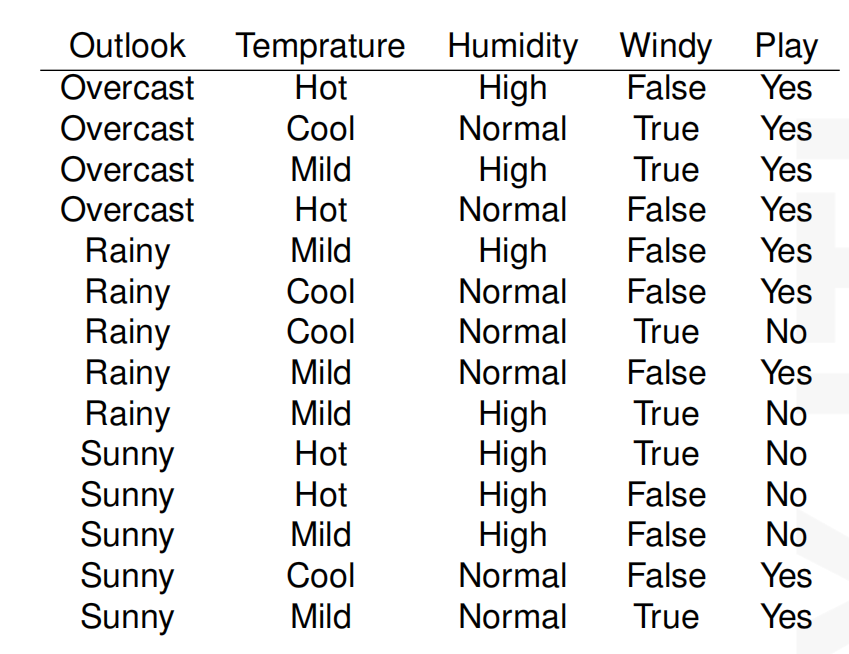
*How can we make use of Bayes’ Rule for Classification? Maximize the posterior probability of observations . This method is named MAP estimation (Maximum a posteriori) .* 注意这里只介绍传统贝叶斯学习，不涉及神经网络，没有参数需要更新，就是对数据集做计算而已。

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：

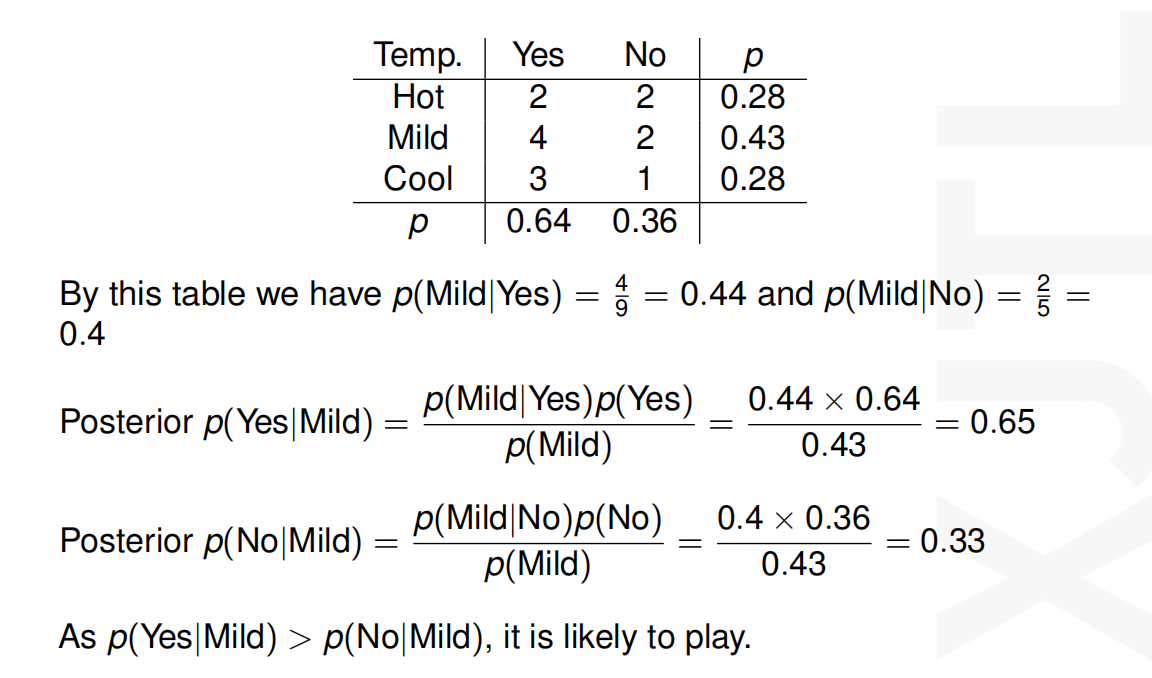
*P (c |x ) =*

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 ) .

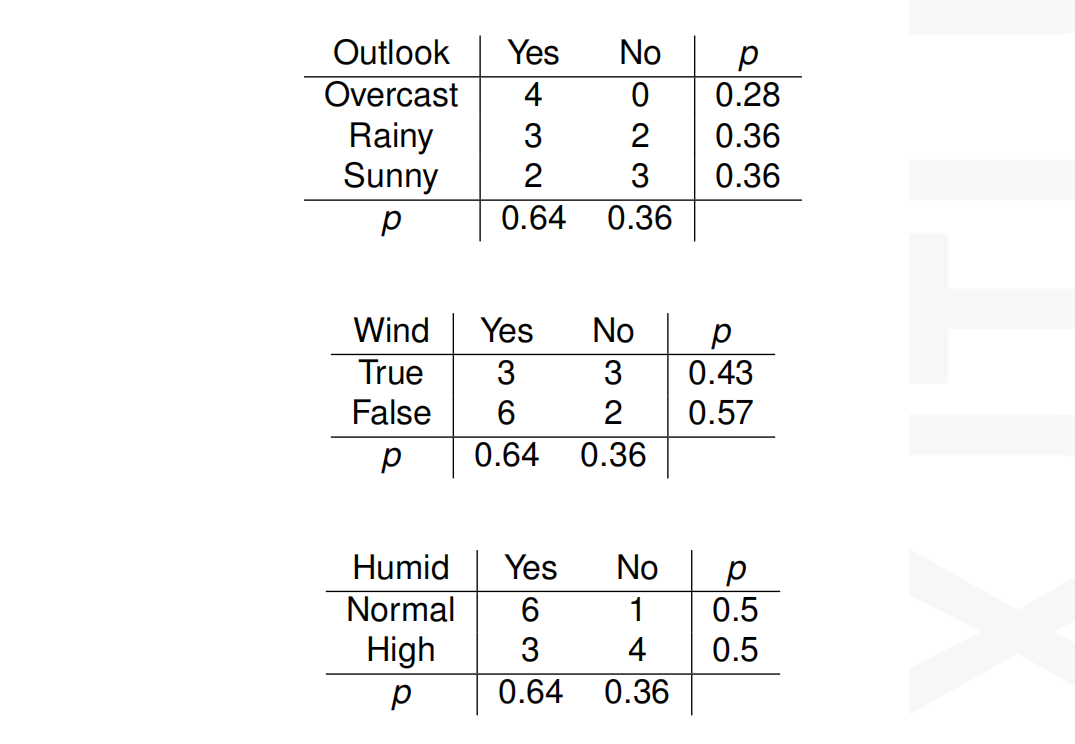
*Exercise :*

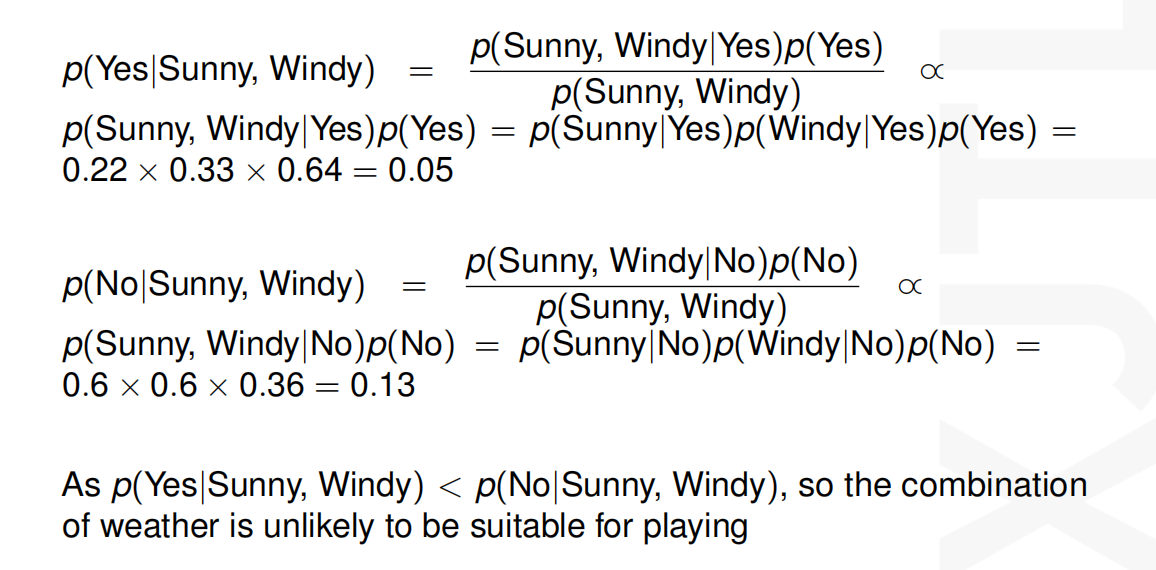
**

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

**

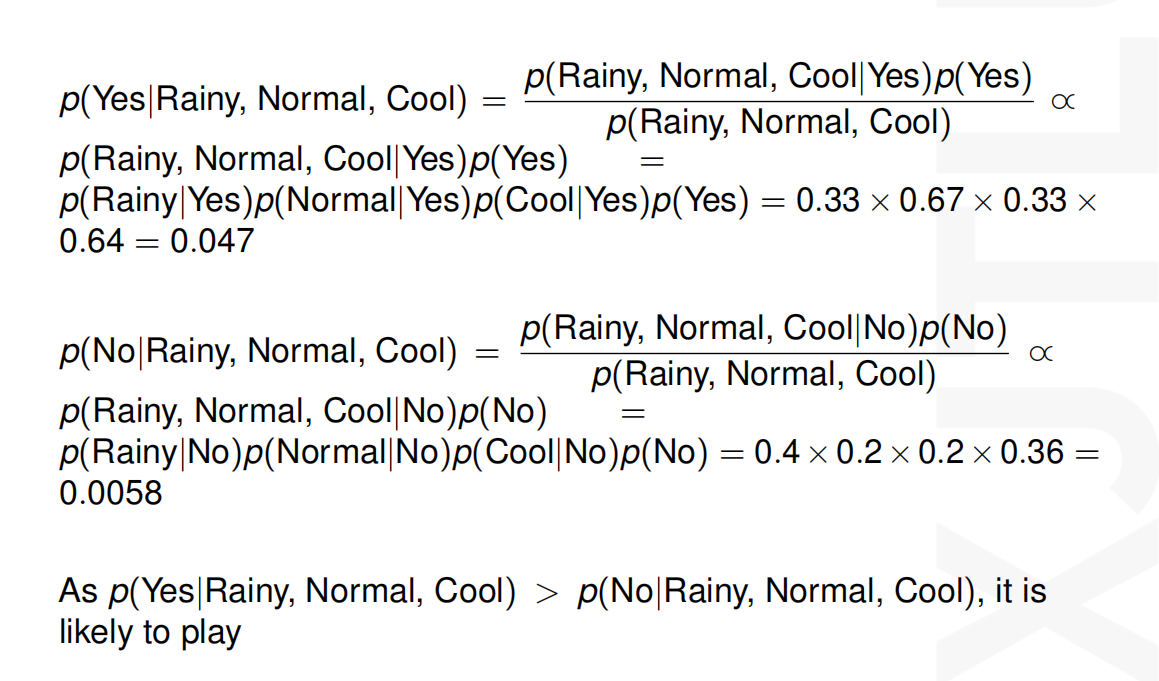
*2 Will you play on the day of (Sunny,Windy) ?*

**

**

值得注意的是，我们实际在进行这些计算时，我们不会只是看到图表中这么点数据，无法直接从图表得到 P(Yes | Sunny, Windy) 的数值(这里这个问题二者实际都是50%，但是我们现在问题是基于假设独立同分布计算，虽然下面也能展开但对于双方都是常数，直接计算会导致和不为1(原因在于实际这个表中数据不是独立同分布导致，所以需要使用上面方法，只计算分子，不考虑相加为1比较）

*3 Will you play on the day of (Overcast,Normal Humidity, Cool) ?*

**

*5 Support Vector Machine*

*SVMs are learning systems that:*

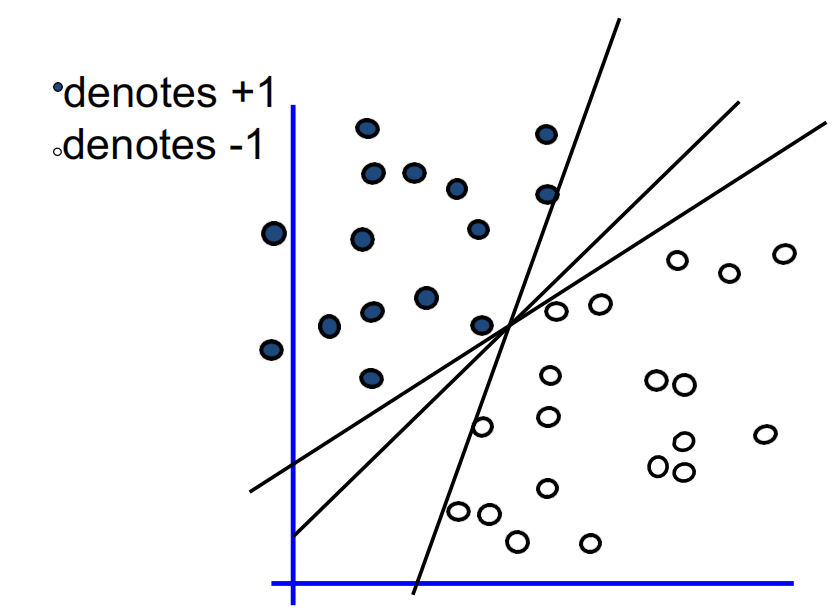
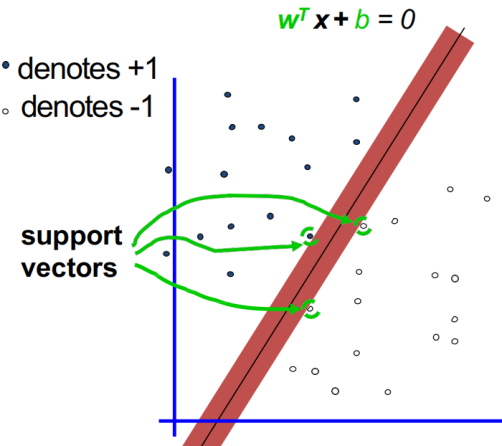
*---- use a hyperplane of linear functions*

*---- in a high dimensional feature space* ***( Kernel function )***

*---- trained with a learning algorithm from optimization theory* ***( Lagrangian duality )***

*---- Implements a learning bias derived from statistical learning theory* ***( Generalization )***

*5.1 Linear SVM Classification*

* *

*We can see that there are many lines that can classify the data,which one we want to choose ?* ***Classifier Margin*** *is defined as the width of the decision boundary, where the width of the decision boundary grows to exactly touch the instances.*

*Firstly ,Let explain why x + b =0 represents a hyperplane . Think of “ax+by+c=0” represents a line in feature 2 and “ ax+by+cz+d=0” represents a plane in feature 3. 在数学和几何学中，超平面是一个维度比整体空间少一的子空间。对于 n 维空间，一个超平面实际上是一个 𝑛−1维的子空间, 其中 w为超平面的法向量，b用来滑动平移。我们可以通过计算 x + b >0 和 x + b <= 0 进行分类。*

*If y = 1 ,then we set x + b > 1* ***( 1 is the unit marginal )*** *If y = - 1 , then we set x + b <= - 1 . Therefore , we can get a margin:*

***Maximum margin linear classifier*** *is a linear classifier with maximum boundaries. It’s the simplest form of SVM, called Linear SVM(LSVM) .*

*We want to maximize this one, which means we want to :*

*minimize*

*subject to ( + b ) >= 1*