-------------------------------------------------------------1------------------------------------------------------------------

Data Mining is the extraction of: **implicit、previously unknown、potentially** useful information from data.

**Input**: typically, single table with data

Data rows represent independent from each other instances/examples of a particular concept.

Columns are attributes/features/variables which measure aspects of an instance.

**Output**: Compact structural patterns (knowledge) that describes the concept

Can be used for automatic decision-making, e.g., to predict the value of a particular [target] attribute of new instances.

Can be used to understand and explain data in a compact way and/or how automatic decision-making is done.

Numerical attributes: discrete or continuous

Categorical attributes: Values serve only as labels or names.

Nominal

Ordinal

Large datasets may require specific management: distributed storage and processing.

-------------------------------------------------------------2------------------------------------------------------------------

"At a high level, EDA is the practice of using visual and quantitative methods to understand and summarize a dataset without making any assumptions about its contents.

**What is the best-practice to make an effective EDA?**

Explore the data to identify columns with: missing data,

outliers, noise.

Understand the relationship, or lack of, between attributes/columns.

Identify useless attributes/columns, e.g. ID numbers, URLs, etc.

We leave any transformations for the data preparation step that follows.

-------------------------------------------------------------3------------------------------------------------------------------

**Data Preparation** Also known as: Data Preprocessing、Data Wrangling、Data Munging

：Treat missing values、Treat outliers、Feature engineering（Transform existing features、Add new features (extracted from existing features)）

MCAR – Missing Completely At Random

MAR – Missing At Random：data is missing randomly only within sub-samples of data

MNAR – Missing Not At Random：missing data has a structure to it

If more than **25-30% values are missing** in a column then **drop the column**.

**Impute missing values** ：：Simple method: use mean, median, most frequent, constant

Better method: KNN (K-Nearest Neighbour)

**（处理缺失值）**

**Drop rows with outliers、Add a new binary feature to mark outliers、Transform(Log,Cubic root)（处理异常值Handling Outliers）**

分类数据处理成数字：有序的用mapper 无序的用one-hot encoding

Rescale Features and Normalization：许多机器学习算法要求特征为相同scale

Scikit-learn Scalers: MinMaxScaler、RobustScaler、StandardScaler

Normalisation：Normalise data rows(L1绝对值之和，L2所有数的平方根，Lint该行最大绝对值)

Use RobustScaler if you want to reduce the effects of outliers(删除中位数,用IQR范围进行缩放)

StandardScaler results in a distribution with a standard deviation equal to 1 and a mean approximately 0.标准差等于1、平均数近似于0的分布

先进行对数转换，再进行缩放

创建新的features，encoding编码就是这样做的。

-------------------------------------------------------------4------------------------------------------------------------------

Styles of Machine Learning

1.Predictive/Supervised ：Classification/Numeric prediction techniques

2.Descriptive/Unsupervised：Association learning techniques：detecting associations between features

Clustering techniques：grouping similar instances into clusters

Clusters can be: disjoint vs. Overlapping 、deterministic vs. Probabilistic、flat vs. hierarchical

**K-means（disjoint，deterministic，flat）**

To cluster data into k groups (k is predefined):

1.Choose k cluster centers (e.g. at random)

2.Assign each instance (i.e. data point) to the closest center. All data points assigned to the same center form a cluster.

3.Compute the centroids of clusters.

centroid of a cluster: the mean of all data points in the cluster.

4.If one of the following is true then STOP. Otherwise go to step 5.

Centroids of newly formed clusters are the same as the centers used in step 2 OR

Data points remain in the same cluster OR

Maximum number of iterations are reached.

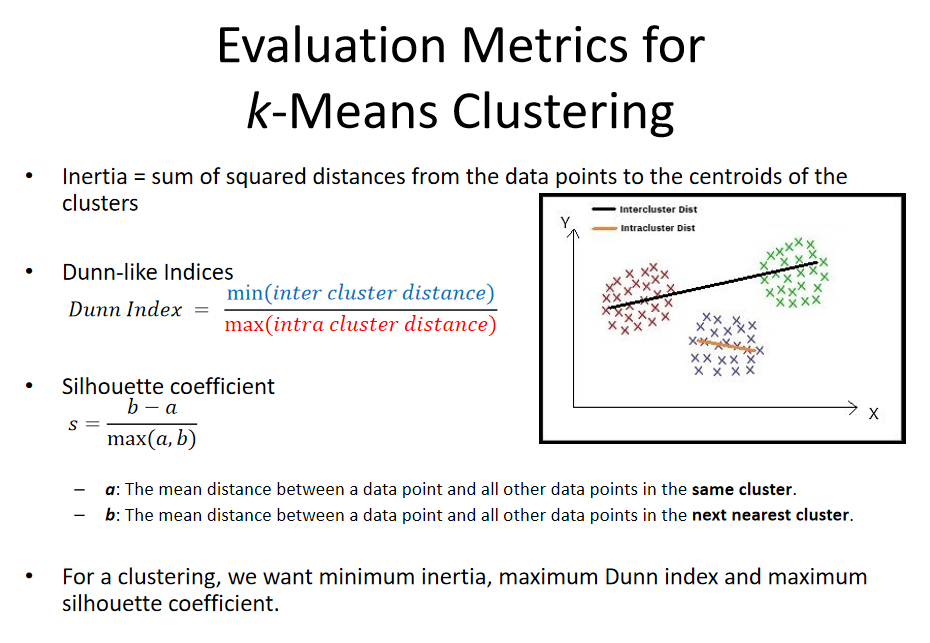
5.Use the centroids of the clusters as new centers and go to step 2.

**k-means minimizes sum of squared distances from the data points to the cluster centers**. This sum is also called **inertia**.

**To increase chance of finding a global minimum of the inertia**:

Restart with different **random** centers. Choose the clustering with the minimum inertia.

Smart choice of initial centers: **k-means++**.



这些指标（惯性、邓恩指数、剪影系数）的值可以在不同的k值下绘制出来，以便为k-means选择最佳的k值。

**Bisecting k-means**（二分法k-means）

在每个分叉步骤中把一个聚类分成两个子聚类（通过使用k-means），直到得到一个有意义的聚类集合。

-------------------------------------------------------------5------------------------------------------------------------------

分类的基础知识：......

分类算法

一个属性可以完成所有工作- OneR (one rule) algorithm

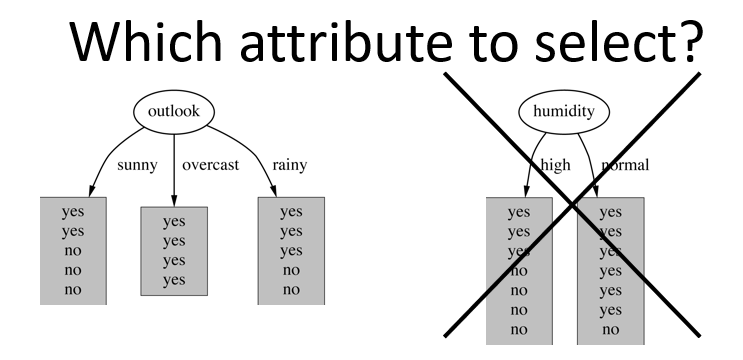
所有属性都是平等且独立的（Naïve Bayes朴素贝叶斯）

属性的加权线性组合可能可以做到（logistic regression逻辑回归）

**·Decisions Tree CLASSIFIER**

策略：自上而下

递归的分割和征服方式



首先：为根节点选择属性为每个可能的属性值创建分支

然后：将实例分成子集，从节点延伸出来的每个分支都有一个子集。

最后：对每个分支进行递归重复，只使用到达该分支的实例。

如果所有的实例都有相同的类别，则停止。

**Which is the best attribute**? Want to get the smallest tree & Heuristic: choose the attribute that produces the “purest” nodes

**Measure of purity: info[node]** - information value of a node measured in bits.

**策略：选择子节点信息量最低的属性**

**Use entropy to calculate information value: 计算子节点信息量**



**where p1+p2+…+pn = 1.**

Information gain = info[Before the split]-info[After split]

Pruning techniques to avoid overfitting: Use a validation dataset（ID3） 剪枝技术避免过拟合

Similar approach: CART

**·Random Forest Algorithm 是一种很牛逼的机器学习分类算法，它是一种bagging/bootstrap aggregation的集成机器学习算法.** 随机森林是一种分类算法，由许多决策树组成。在构建每个单独的决策树时，它使用bagging和特征随机性，以尝试创建一组不相关的树，其委员会的预测比任何单个树的预测更准确。

Ensemble learning集成学习指的是一组（或者集合）的基本机器学习算法，它们work collectively共同协作以达到更好的预测性能。

Bagging和Boosting是集成学习中的两种主要方法：

- Bagging：基本模型并行训练。

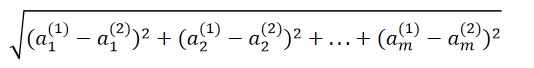
- Boosting：基本模型顺序训练。

常见的Boosting算法有AdaBoost、XGBoost、GradientBoost、BrownBoost等。

**·KNN CLASSIFIER(Instance Based Learning)---k-nearest-neighbor**

Distance Function：学习的内容一般取决于距离函数

Dataset with several numeric attributes: **Euclidean distance** （比较距离时不需要开根）



where a(1) and a(2) are two instances with m attributes.

**Another popular metric**: **city-block metric**: Adds differences without squaring them.

Discuss：

精准但是慢

Assumes all features are equally important。Remedy: feature selection or weights.

-------------------------------------------------------------6------------------------------------------------------------------

Evaluation of Classifiers

**Classifier/Classification algorithm**: OneR, Decision Tree, kNN, Naïve Bayes, SVM, Logistic Regression, Multilayer Perceptron

error rate = 错误的预测/总预测

Resubstitution error: error rate obtained from training data重构误差是乐观的，因为是训练集

先用训练集在分类器上训练多个模型，验证集找一个好的；再用训练加验证集在分类器上训

练多个模型，测试集评估找好的；最后整个数据集在分类器上训练多个模型找好的。

一般来说，训练数据越大，分类器就越好。

测试数据越大，误差估计就越准确。

训练数据不充分容易使模型欠拟合或过拟合

Insufficient test data: low confidence in the result.

如何划分数据集、如果数据量有限怎么办?

**Holdout** method搁置法：部分测试验证，剩余训练 ------Preferably use stratification分层法：保证每个集的lable比例相同

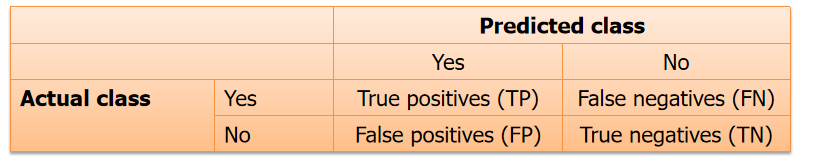
**Repeated Holdout**：通过用不同的子样本重复这个过程，可以使Holdout估计更加可靠，总错误率=每次迭代错误率/迭代次数

为了防止不同的test and validation集overlap重叠？

**k-fold cross-validation** avoids overlapping validation sets: 分成k组子样本

**ten-fold cross-validation广泛的实验表明，这是获得准确估计的最佳选择**

**repeated stratified cross-validation**：ten-fold cross-validation is repeated ten times and results are averaged

Stratification reduces the estimate’s variance分层能够减少估计的方差

The confusion matrix:

precision = TP/(TP+FP)

true positive rate (sensitivity or recall) = TP/(TP+FN)

true negative rate (specificity) = TN/(TN+FP)

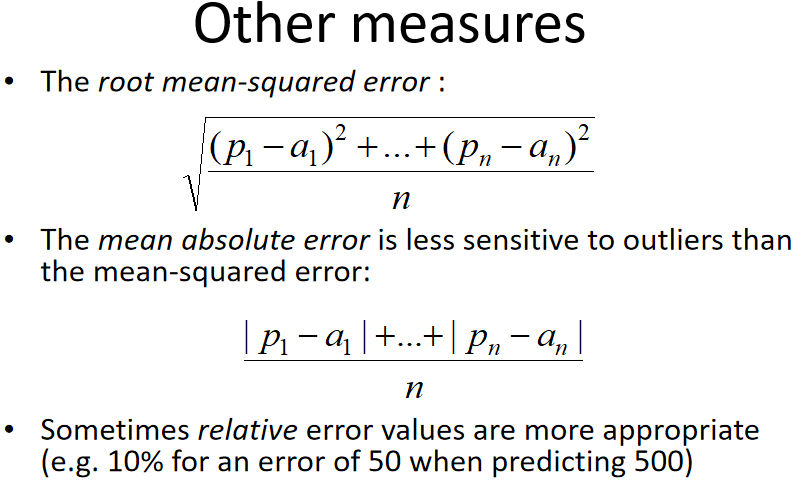
false positive rate = FP/(FP+TN)

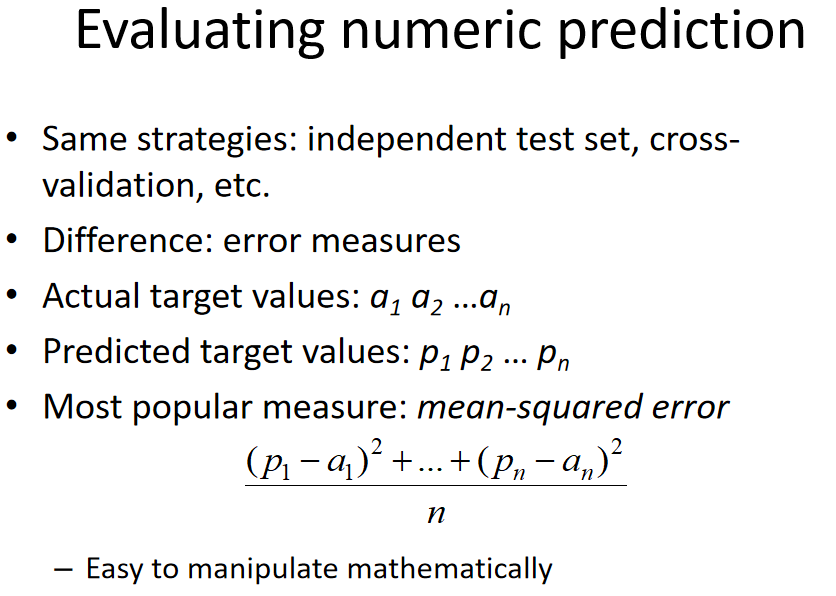
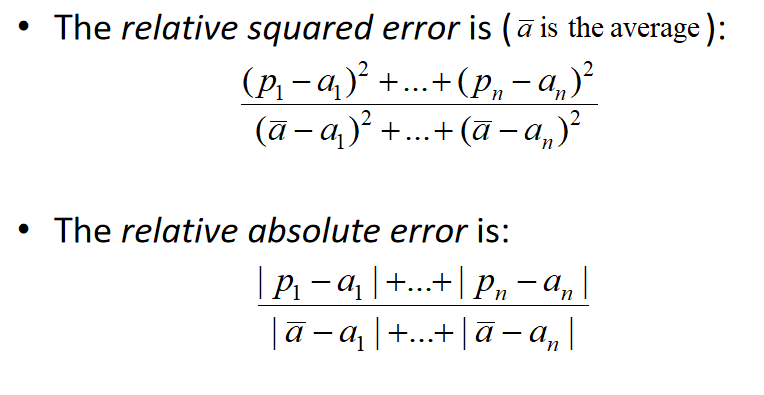
F1-score = (2recallprecision)/(recall+precision)，Also known as F-score

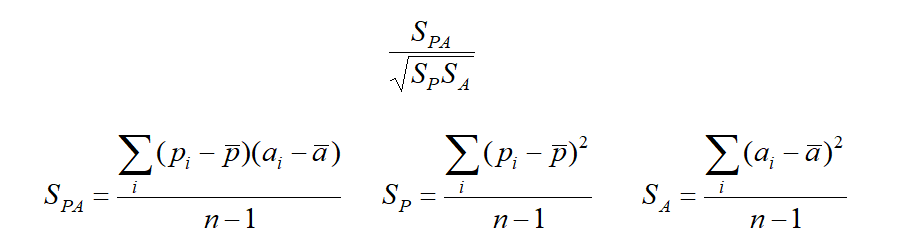
当观察结果在每个类别之间平衡时，ROC曲线是合适的，而当数据集不平衡时，

Precision-Recall曲线是合适的。

-------------------------------------------------------------7------------------------------------------------------------------

Linear Models：Numeric Prediction/Regression



Correlation coefficient相关系数：衡量预测值和实际值之间的统计相关性

范围在-1,1之间，性能越好数值越大

Classification: - Logistic Regression  
- Perceptron  
- Winnow  
- SVM

multi-response linear regression多响应线性回归：将回归方法用于分类问题

(不太理解并未整理，lec7 page13之后)

-------------------------------------------------------------8------------------------------------------------------------------

Dimensionality Reduction：

**休斯效应Hughes effect： 在训练样本数量固定的情况下，分类器的准确度首先会随着特征数量的增加而增加，但在某一时刻后，它开始急剧下降。**

**Feature selection**：Select some of the features and use only them for model training and testing

Dimensionality Reduction（**Feature Transformation是一种降维技巧**）：Transform features to a lower-dimensional space and use only the transformed features for model training and testing

特征选择同样降低了特征空间的维度，分别对训练集和测试集应用以上方法

Feature Selection Algorithms: Filter Methods、Wrapper Methods、Embedded Methods

1.**Filter feature selection methods** apply a **statistical measure** to **assign a scoring** to each feature.

statistical measure: Chi squared test 卡方检验

Information gain 信息增益

Correlation coefficient scores 相关系数得分

2.**Wrapper methods** consider **the selection of a set of features** as **a search problem**.

准备、评估不同的特征组合，并与其他组合进行比较。我们使用预测模型来评估特征组合并根据模型准确性分配分数。包装器方法的一个例子是递归特征消除recursive feature elimination (RFE)算法。

3.**Embedded methods** learn **which features best contribute to the accuracy of the model** while the **model is being created**.

regularization methods是最常见的嵌入式特征选择方法

它们在预测算法（如回归算法）的优化中引入额外的约束，使模型偏向于较低的复杂度（较少的系数）。

正则化算法有：LASSO 、Ridge Regression、Elastic Net

Feature Transformation： Linear：Principal Component Analysis (PCA)

Non-linear: Multidimensional scaling (MDS)、T-SNE

-------------------------------------------------------------9------------------------------------------------------------------

Introduction to Recommender Systems：

Recommender systems can be loosely broken down into three categories: **content-based filtering systems**, **collaborative filtering systems**, and **hybrid systems** (combination of the other two).

（除此之外，还有两种Social and demographic recommenders、Contextual recommenders）

1. recommend similar items **based on item content**（text documents）

2. you are given a **matrix of preferences** (explicit or implicit) by users for items

（前提是要有user and item IDs and a notion of preference by users for items ）

1. Social and demographic recommenders suggest items that are liked by friends, friends of friends, and demographically-similar people.
2. Contextual recommendation algorithms recommend items that match the user’s current context\situation.

挑战：冷启动（对于新用户）、Sparse data的处理

**Content-based Filtering (CFB)**

将物品内容和用户档案表示为a set of descriptors or terms. +一种算法从而给用户提出推荐

可选择算法：**KNN**，Relevance feedback、Genetic algorithms、Neural networks、Bayesian classifiers

标注二者速度最好，其次KNN实例越多越慢，最后那俩小垃圾因为需要多次迭代。贝叶斯需要大量实例，相关性反馈和KNN一个instance就够了。

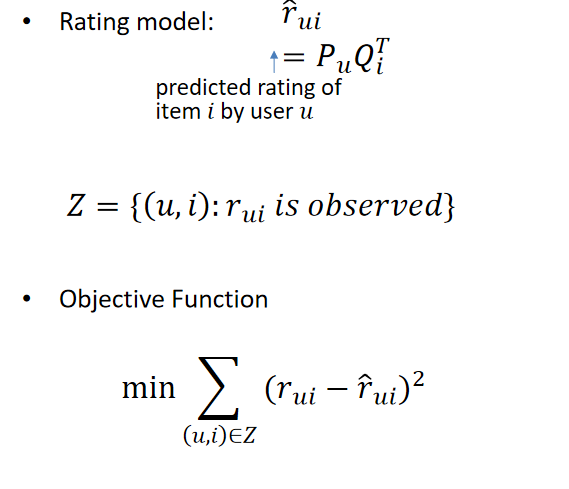
**Collaborative Filtering (CF)**

Types of CF: neighbourhood-based (memory-based)、model-based

**Neighbourhood-based** algorithms approach the CF problem by using the entire dataset.

和active用户相似的用户的物品偏好（user based），或和active用户物品的相似物品基于物品的共同购买（item based）（个别情况冷启动困难，无法避免过拟合，速度慢）

**model-based**：Extract some information from the dataset, and use that as a "model" to make recommendations without having to use the complete dataset every time.

---> Standard model CF: **Matrix Factorization , MF-based** methods for CF, 用两个小矩阵去近似评级矩阵，从而用评级矩阵中的潜在模式进行推荐

---> Model-Based CF: Matrix Factorization (MF)

**Optimisation methods:**

Stochastic Gradient Descent (SGD)随机梯度下降

Alternating Least Squares method (ALS)交替最小二乘法

-------------------------------------------------------------10----------------------------------------------------------------

Introduction to Text Mining：非结构化文本数据

-Classify/rate text documents

-Discover previously unknown patterns

**Task：**

Identify/detect: Sentiment、Spam、Offensive language/cyberbullying、Authorship、Language

Descriptive analysis: Topic Modeling

Document Clustering

**Pre-Processing：**

Tokenization:

Break the document into words or combination of words (n-grams).

Remove prefixes and suffixes from words, i.e. represent each word by its root/stem: stemming.

从单词中删除前缀和后缀，即用词根/词干表示每个单词：词干提取

Normalisation:

Remove stop words: articles, conjunctions and prepositions.删除停用词：冠词、连词和介词

**Text Document Modelling：**

**·**Bag of Words model: A piece of text/document is represented as bag of words, disregarding grammar and word order.词袋模型，忽略语法和词序

**·**N-gram model: represent spatial relationships between words. Generalisation of bag of words.

表示单词之间的空间关系。词汇包的概括。

**·**Example:

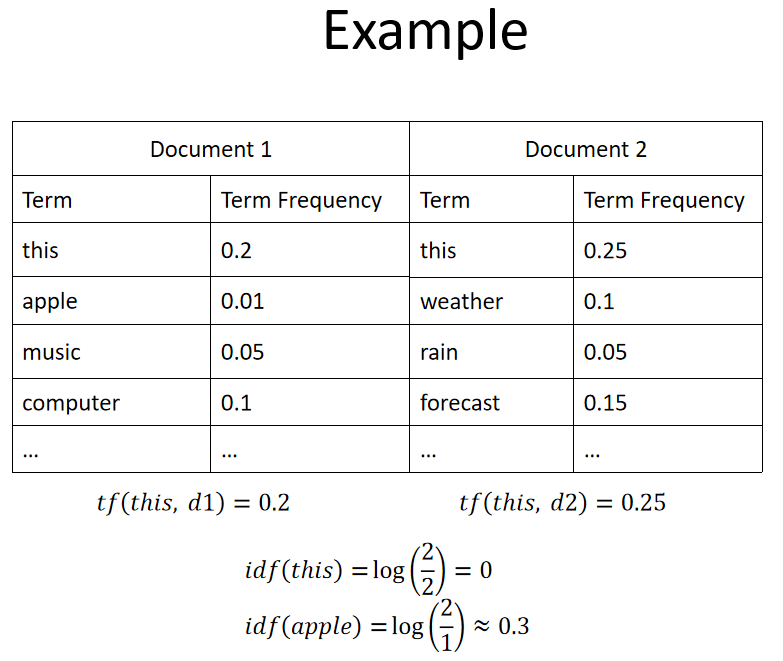
Document: “Three midterm exams in a week, and I love exams! Can we have another midterm exam this week please?“

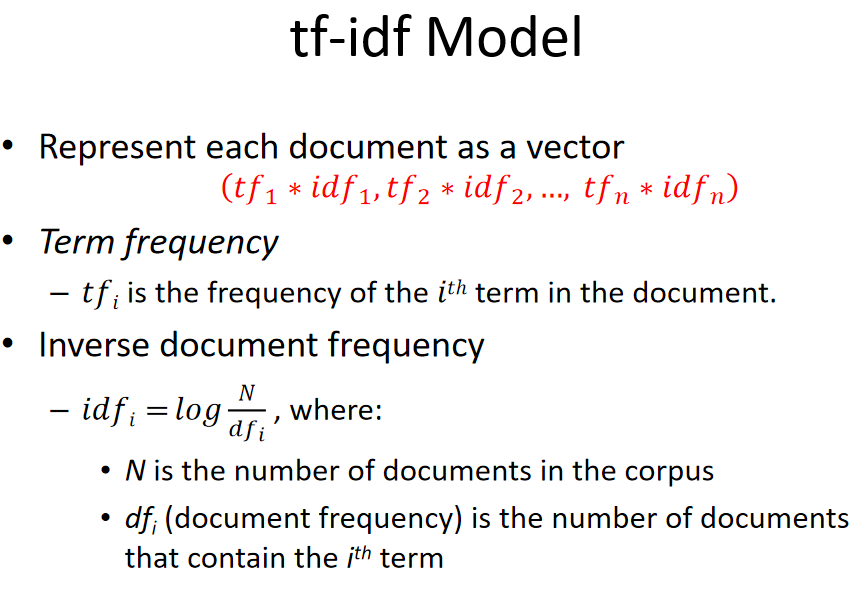
Bag of words: {three:1, midterm:2, exam:3, week:2, I:1, lov:1, can:1, we:1, hav:1, another:1, this:1}

Bigrams: {three midterm:1, midterm exam:1, exam week:2, …}

**Document-Term Matrix：**

Document-term matrix文档术语矩阵是一个矩阵，其元素表示文档集合（语料库）中出现的术语（即单词或 n-gram）的数量或频率，行对应文档，列对应术语。

A term frequency–inverse document frequency (tf-idf) is a weighting scheme to reduce the significance of terms occurring frequently in the corpus.术语频率-逆文档频率 (tf-idf) 是一种加权方案，用于降低语料库中频繁出现的术语的重要性。



**Stop Word Elimination：**

**Stop words** are words used frequently in the language, but **not providing significant information**.

Any group of words can be chosen as the stop words for a given purpose

Removing such words could **improve the performance but also could impair the result.**

**Stemming：**

**Reduce terms to their roots/stems: prefixes, suffixes,…**将术语缩减为词根/词干：前缀、后缀……

Example，**compressed** and **compression** are both accepted as equivalent to **compress**

**Advanced Representation Methods**

Word Embeddings (Word2Vec, Google’s BERT)词嵌入

->Low-dimensional vector for each token/word每个词的低维向量

->相似的词有相似的向量

Example, the vector of dog will be most similar to the vector of cat & king-men+women would be similar to the vector of queen

-------------------------------------------------------------11----------------------------------------------------------------

Lecture 11 Overview of DL

* A critical ingredient is the use of much larger quantities of data than has heretofore been possible. 现在有大量的数据
* Recent successes have arisen in settings involving high capacity models—ones with many parameters. 现在有很多很大的，有很多参数的模型

Deep Learning vs Machine Learning

* Deep Learning is Machine Learning.
* Works within the same data mining workflow:

- Data collection

- Data pre-processing

- Train a model = deep network

Feature engineering: features are learned in the process of training a model

- Evaluate with a test set

Cross validation typically not used as deep learning is used with huge datasets

* Deep Learning can be:

- Supervised (classification, numeric prediction)

- Unsupervised (clustering, e.g., clustering of images)

* Problem: a trained deep neural network typically does not give an insight into the data (unlike, say, a decision tree or a list of rules)经过训练的深度神经网络通常不能对数据进行深入了解（与决策树或规则列表等不同）。

Deep Network Architectures

* Feed Forward Neural Network 前馈神经网络 Multilayer Perceptron network
* Recurrent Neural Network 递归神经网络
* Symmetrically Connected Networks 对称连接的网络

The neural network renaissance and deep learning revolution

人们对神经网络和深度学习重新感兴趣，在语音识别和计算机视觉领域有很大的发展

GPUs, graphs and tensors

* The easy availability of high-speed computation in the form of graphics processing units has been critical to the success of deep learning techniques以图形处理单元形式出现的高速计算的便利性对于深度学习技术的成功至关重要。
* When formulated in matrix-vector form, computation can be accelerated using optimized graphics libraries and hardware 当以矩阵-向量形式制定时，可以使用优化的图形库和硬件来加速计算
* As network models become more complex, some quantities can only be represented using multidimensional arrays of numbers 随着网络模型变得越来越复杂，一些数量只能用多维数组来表示

Such arrays are sometimes referred to as tensors, a generalization of matrices that permit an arbitrary number of indices

这样的数组有时被称为张量（tensor），是矩阵的一种泛化，允许有任意数量的索引

* Software for deep learning supporting computation graphs and tensors is therefore invaluable for accelerating the creation of complex network structures and making it easier to learn them

因此，支持计算图和张量的深度学习软件对于加速创建复杂的网络结构并使其更容易学习是非常宝贵的。

Key developments

The following developments have played a crucial role in the resurgence of neural network methods:

* the proper evaluation of machine learning methods; 对机器学习方法进行适当评估；
* vastly increased amounts of data; 数据量大增
* deeper and larger network architectures; 更深、更大的网络架构
* accelerated training using GPU techniques  使用GPU技术加速训练

Deep Learning Development Libraries/Frameworks

* Tensorflow
* PyTorch
* Keras
* MXNet
* The Microsoft Cognitive Toolkit
* Caffe
* Deeplearning4j (Java)
* Chainer

-----------------------------------------------------------END----------------------------------------------------------------