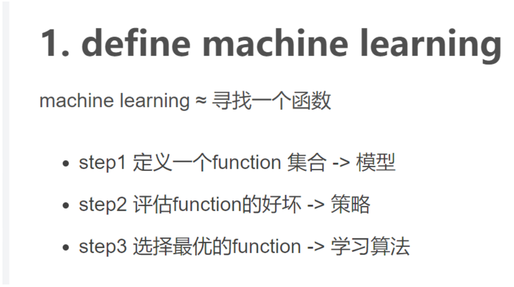
机器学习是一类算法的总称，这些算法企图从大量历史数据中挖掘出其中隐含的规律，并用于预测或者分类，更具体的说，机器学习可以看作是寻找一个函数，输入是样本数据，输出是期望的结果
More specifically, machine learning can be seen as finding a function, the input is sample data, and the output is the expected result.

Fundamentals:



1. define machine learning

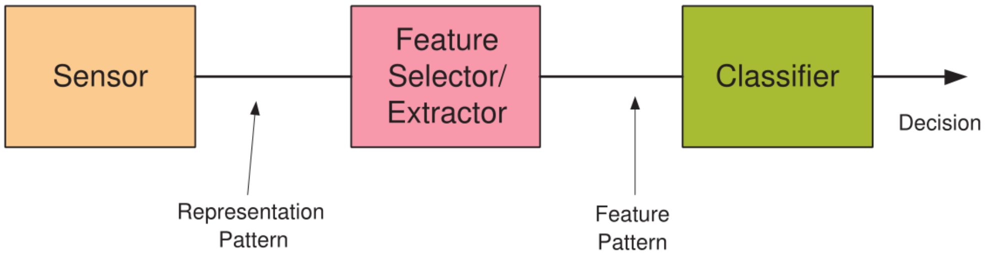
machine learning ≈ 寻找一个函数

- step1 定义一个function 集合 -> 模型
- step2 评估function的好坏 -> 策略
- step3 选择最优的function -> 学习算法

收集数据，输入数据，数据的探索与预处理，训练及测试算法，评估
Data collection，Input data，Data exploration and preprocessing，Training and testing algorithms， Evaluation.

- Basic models of Pattern Recognition (Machine Learning)

In pattern recognition we study how to design machines that can recognize and classify ”things”.



Representation pattern is the raw data we obtain from the sensor e.g. image or video pixels, price of stock, etc.
Feature pattern is a small set of variables obtained through some transformation-feature selection and/or extraction.
The trained classifier uses the feature pattern to make a decision regarding the pattern presented at its input.
表示模式是我们从传感器获得的原始数据，例如图像或视频像素、股票价格等。
特征模式是通过一些变换-特征选择和/或提取获得的一组小变量。
训练的分类器使用特征模式对输入的模式进行决策

Supervised Learning:
Supervised learning is a type of machine learning where the algorithm learns from labeled examples. Labeled data consists of input samples along with their corresponding output or target values. The goal of supervised learning is to train a model that can accurately predict the output for new, unseen inputs.

Unsupervised Learning:
Unsupervised learning, on the other hand, deals with unlabeled data, where the input samples do not have corresponding target values. In unsupervised learning, the objective is to uncover hidden patterns, structures, or relationships within the data without explicit guidance.

Overfitting:
model performs exceptionally well on the training data but fails to generalize to new, unseen data. It occurs when a model becomes too complex and starts to capture noise or irrelevant patterns present in the training set.

Generalization:
ability of a model to perform well on new, unseen data that it has not been trained on.

① SVM,只能用于二分类，

Support vector machine (SVM) is a kind of generalized linear classifier which classifies data by supervised learning. Its decision boundary is the maximum margin hyperplane [1-3].

② KNN,适合用于多分类问题，缺点：计算量大（改进：训练前对数据预处理，剔除关联性小的数据），不平衡时分类差（改进：加权值 weight，距离小的权值大）

The so-called k nearest neighbor means k nearest neighbors. It means that each sample can be represented by its nearest k nearest neighbors. Nearest neighbor algorithm is a method to classify every record in the data set.

③ k-means,将划分为 k 个聚类，缺点：k 难以估计和确定；人为确定，不同的 k 导致不同的结果（k-means++：初始中心相距尽量远）。

1) 先选取 k 个类的中心，
2) 然后求他们到 k 个中心的距离，将样本归类，
3) 利用均值方法更新中心。
4) 重复 2，3，直到所有的聚类中心不再更新。

K-means algorithm:
For a given sample set, the sample set is divided into K clusters according to the distance between samples.
Let the points in the cluster as close together as possible
Let the distance between clusters as large as possible.

- Bayesian methods
  - Bayes theorem and Bayes Formula.
  - Bayes minimum error probability classifier (Bayes decision rule).
  - Bayesian estimation.

$P(A|B)=P(B|A)*P(A)/P(B)$。

Theorem:
Bayes is that a method to calculate the possibilities of something happened, and we can give some conditions to the thing. We use the Prior and Likelihood, Posterior .

Bayesian decision theory is to estimate the subjective probability of some unknown states under incomplete information, then use Bayesian formula to modify the probability of occurrence, and finally use the expected value and modified probability to make the optimal decision.

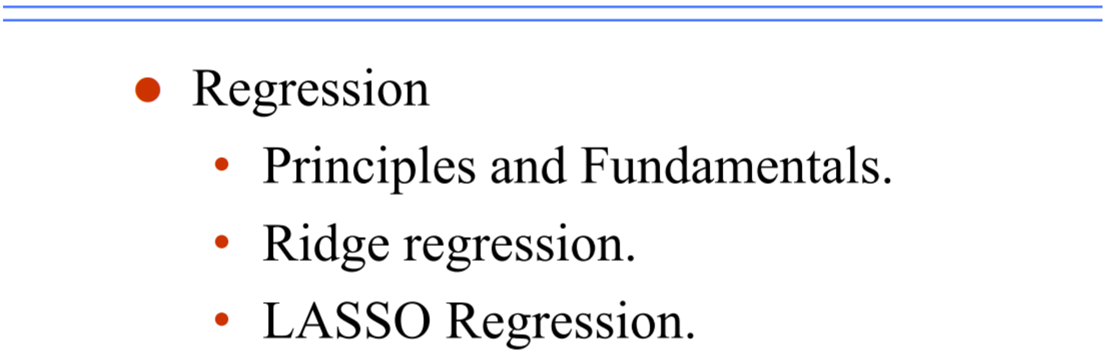
The essence of Bayesian inference is in the rule, known as Bayes' theorem, that tells us how to update priori probability, to find out Posteriori probability.
贝叶斯推理的本质在于规则，即贝叶斯定理，它告诉我们如何更新先验概率，以找出后验概率。

① 解释最大似然法和贝叶斯估计法之间的差异
Maximum likelihood estimation: the parameter of the distribution θ is fixed.
While Bayesian estimation assumes θ is a random variable.
Maximum likelihood estimation focuses on finding the parameter values that maximize the likelihood of the observed data.
Bayesian estimation incorporates prior knowledge and provides a posterior distribution that combines the prior and the likelihood to estimate parameters and quantify uncertainty.
最大似然估计假定分布参数固定，而贝叶斯估计则假定它是一个随机变量。最大似然估计侧重于寻找最大化观测数据可能性的参数值，而贝叶斯估计包含先验知识，并提供一个结合先验和似然的后验分布来估计参数和量化不确定性。

③解释朴素贝叶斯分类
Naive Bayes classification is a straightforward and effective algorithm that leverages Bayes' theorem and the assumption of conditional independence to classify instances into different classes based on their observed features.

朴素贝叶斯分类是一种直接而有效的算法，它利用贝叶斯定理和条件独立性假设，根据实例所观察到的特征将实例划分为不同的类

Linearly separable data refers to a dataset where instances of different classes can be perfectly separated by a linear decision boundary. SVM algorithms are designed to find such linear decision boundaries, but they can also handle non-linearly separable data by mapping it into a higher-dimensional feature space using kernel functions.
线性可分数据是指一个数据集,其中不同类的实例可以被一个线性决策边界完全分开。SVM 算法被设计用来寻找这种线性决策边界，但它们也可以通过将非线性可分数据映射到核函数来处理这些数据。

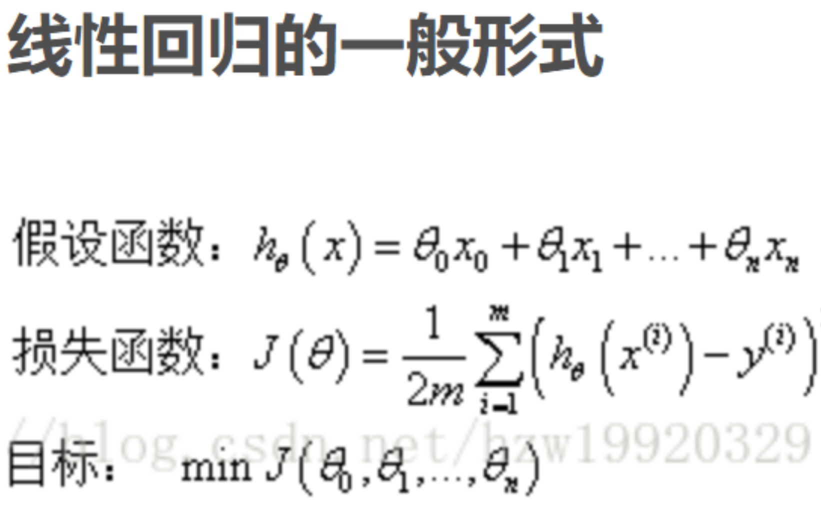


- Regression
  - Principles and Fundamentals.
  - Ridge regression.
  - LASSO Regression.

Regression is a supervised learning method often used in prediction tasks

Generally speaking, it is to find a suitable line or face according to the characteristics (x) in the dataset.. To fit our dataset labels

$$h_\theta(x) = \sum_{i=0}^{n} \theta_i x_i = \theta^T x$$

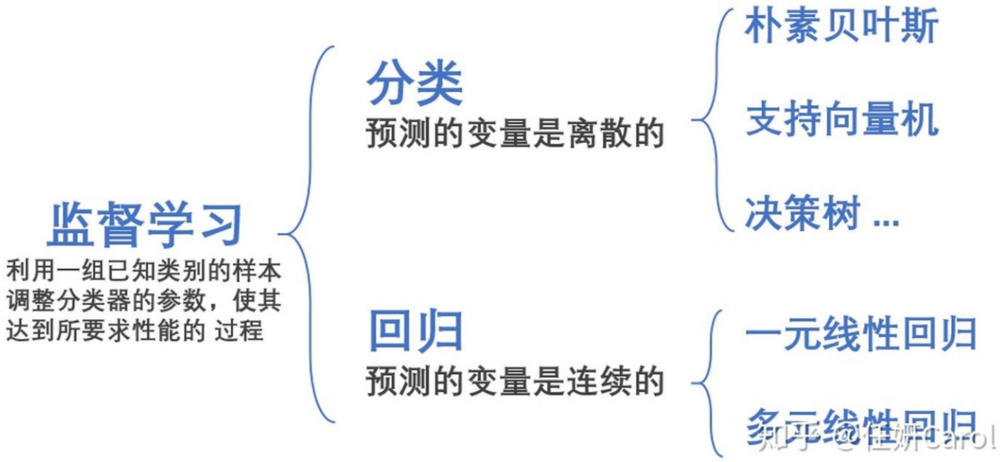


线性回归的一般形式

假设函数： $h_\theta(x)=\theta_0x_0+\theta_1x_1+\ldots+\theta_nx_n$

损失函数： $J(\theta)=\frac{1}{2m}\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right)-y^{(i)}\right)^2$

目标： $\min J(\theta_0,\theta_1,\ldots,\theta_n)$



Least square method and gradient descent method
线性回归用到的两个方法： 最小二乘法和梯度下降法
使拟合得到的预测数据与实际数据之间的误差的平方和是最小的
The square sum of the error between the predicted data and the actual data is minimized

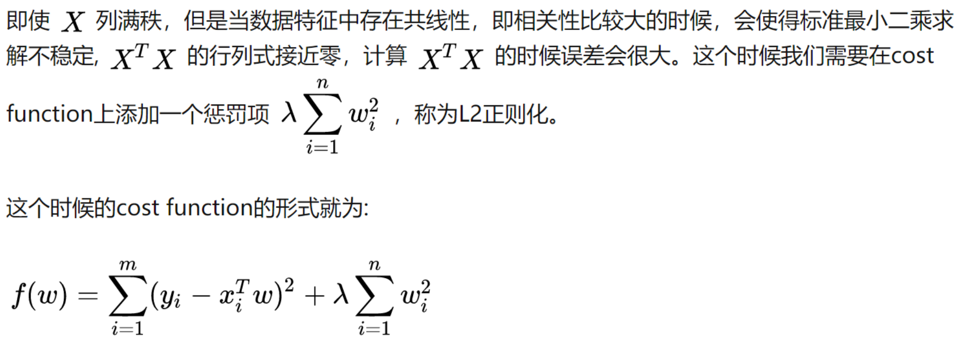
The so-called ridge regression is that for a linear model, the penalty term of L2 norm of the parameter is added to the original loss function, and the loss function is in the following form:

Ridge

- In essence, it is a model selection method in which the ridge parameter λ helps select/weight the variables appropriately.
- The choice of the ridge parameter is a tool to balance the “bias-variance” trade-off. The larger the value of λ the larger the bias and the smaller the variance. The parameter can be determined using cross validation technique.
- The ridge regression estimator is a shrinkage estimator that shrinks the least square

weights toward zero.

- It can be used with (positive definite symmetric PDS) kernels and hence can be extended to non-linear regression and more general feature spaces.



即使 X 列满秩，但是当数据特征中存在共线性，即相关性比较大的时候，会使得标准最小二乘求解不稳定，X^TX 的行列式接近零，计算 X^TX 的时候误差会很大。这个时候我们需要在cost function上添加一个惩罚项 $\lambda\sum_{i=1}^{n} w_i^2$ ，称为L2正则化。

这个时候的cost function的形式就为：

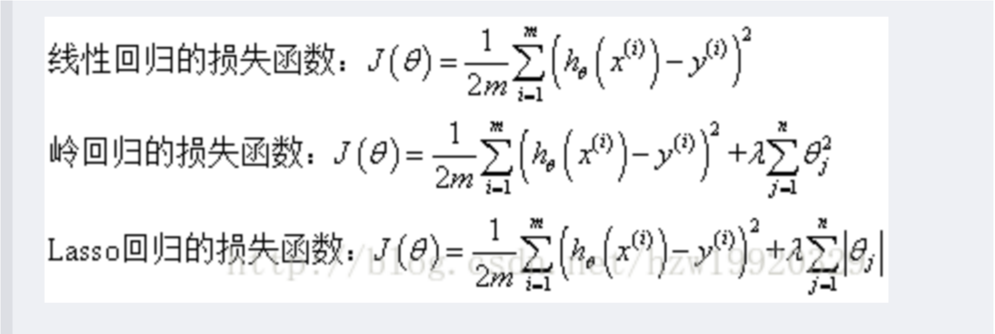
- $$f(w)=\sum_{i=1}^{m}(y_i-x_i^Tw)^2+\lambda\sum_{i=1}^{n}w_i^2$$

among λ Called the regularization parameter, if λ If the selection is too large, all parameters will be changed θ Both are minimized, resulting in under fitting, if λ If the selection is too small, the over fitting problem will not be solved properly, so it is difficult to solve λ It is a technical activity to choose the best way

(2) LASSO regression:

- Lasso uses L1 norm instead of L2 norm (ridge regression)
- Lasso is a short for Least absolute shrinkage and selection operator

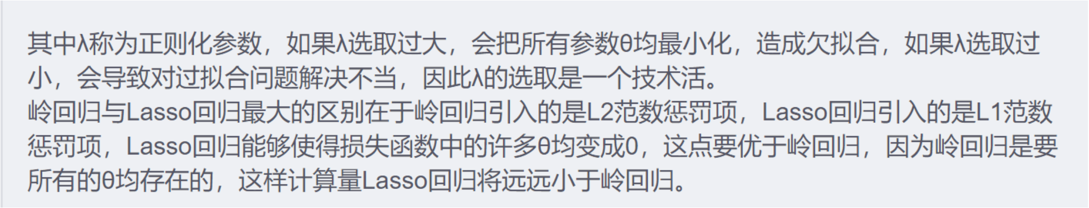
Essentially it combines variable subset selection and shrinkage to improve accuracy



线性回归的损失函数：$J(\theta)=\frac{1}{2m}\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right)-y^{(i)}\right)^2$

岭回归的损失函数：$J(\theta)=\frac{1}{2m}\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right)-y^{(i)}\right)^2+\lambda\sum_{j=1}^{n}\theta_j^2$

Lasso回归的损失函数：$J(\theta)=\frac{1}{2m}\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right)-y^{(i)}\right)^2+\lambda\sum_{j=1}^{n}|\theta_j|$



其中λ称为正则化参数，如果λ选取过大，会把所有参数θ均最小化，造成欠拟合，如果λ选取过小，会导致对过拟合问题解决不当，因此λ的选取是一个技术活。
岭回归与Lasso回归最大的区别在于岭回归引入的是L2范数惩罚项，Lasso回归引入的是L1范数惩罚项，Lasso回归能够使得损失函数中的许多θ均变成0，这点要优于岭回归，因为岭回归是要所有的θ均存在的，这样计算量Lasso回归将远远小于岭回归。

总结：

(1)岭回归：L2 正则化范数

□本质上，它是一种模型选择方法，其中脊参数 λ 有助于适当地选择/加权变量。

□脊参数的选择是平衡“偏差-方差”权衡的工具。 λ 的值越大，偏差越大，方差越小。 可以使用交叉验证技术来确定该参数。

□ 岭回归估计量是一个收缩估计量，它将最小二乘权重缩小到零。

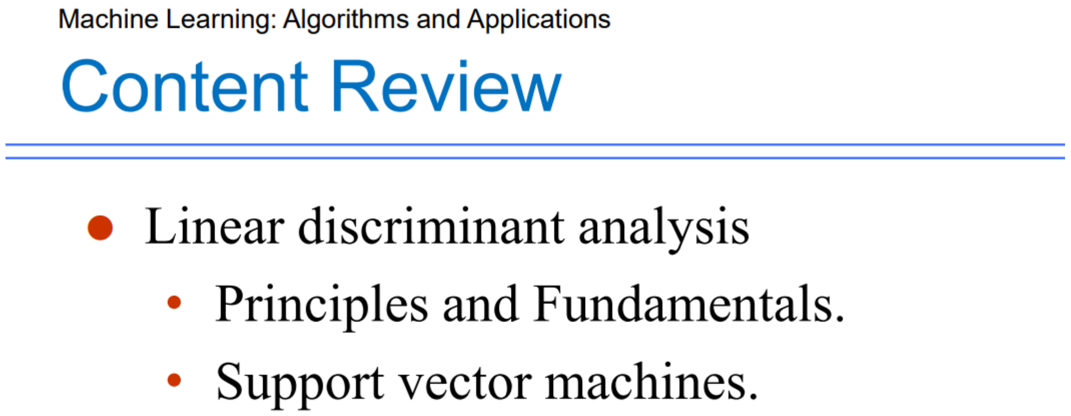
□它可以与（正定对称 PDS）内核一起使用，因此可以扩展到非线性回归和更通用的特征空间。

（2）LASSO 回归：L1 正则化范数，更少的 0

□Lasso 使用 L1 范数代替 L2 范数（岭回归）

□Lasso 是 Least absolute shrinkage and selection operator 的缩写

□本质上它结合了可变子集选择和收缩以提高准确性



Machine Learning: Algorithms and Applications

Content Review

- Linear discriminant analysis
  - Principles and Fundamentals.
  - Support vector machines.

、

The idea of linear discrimination is very simple: given the training sample set, try to project the sample to a straight line, so that the projection points of the same sample are as close as possible, and the projection points of different samples are as far away as possible;
线性判别的思想非常朴素：给定训练样例集，设法将样例投影到一条直线上，使得同类样例的投影点尽可能接近，异样样例的投影点尽可能远离；

线性分类感知机（Linear Classification Perceptron）是一种二元分类算法，用于将数据点划分到两个不同的类别中。

线性分类感知机基于线性分类模型，它假设不同类别的数据可以通过一个线性超平面来分割。该超平面可以将特征空间中的数据点划分为两个不同的类别。

具体而言，对于一个具有 d 个特征的数据点 x，线性分类感知机使用一个权重向量 w 和一个偏置项 b，定义一个线性函数：
f(x) = sign(∑(i=1)^(d) w_i * x_i + b)
其中：

w = [w_1, w_2, ..., w_d] 是权重向量。
x = [x_1, x_2, ..., x_d] 是特征向量。
b 是偏置项。
sign(·) 是符号函数，当其输入大于等于 0 时输出 1，否则输出-1。
线性分类感知机的训练过程是一个迭代的过程。在每次迭代中，算法根据当前的权重向量和偏置项对训练数据进行分类，如果分类错误，则更新权重向量和偏置项以调整分类边界。更新规则如下：
w ← w + η * y * x
b ← b + η * y
其中：
y 是训练样本的真实标签（1 或-1）。
x 是训练样本的特征向量。
η 是学习率（learning rate），控制每次更新的步长。
训练过程会一直进行，直到所有训练样本都被正确分类或达到预定的迭代次数。
线性分类感知机的目标是找到一个能够将不同类别的数据点完全分开的超平面。然而，如果数据不是线性可分的，线性分类感知机可能无法收敛。为了处理线性不可分的情况，可以使用扩展的感知机算法，如支持向量机（Support Vector Machine）等。

5. Linear discriminant analysis

① Support vector machines:

- Data set points lying on the canonical hyperplanes are called support vectors
- Distance between the canonical hyperplane and the separating hyperplane is the margin $\frac{1}{|w|}$
- Support vectors are equally close to the optimal hyperplane
- Support vectors are the training samples that define the optimal separating hyperplane and the most difficult to classify
- Support vectors are the most informative samples for classification task.

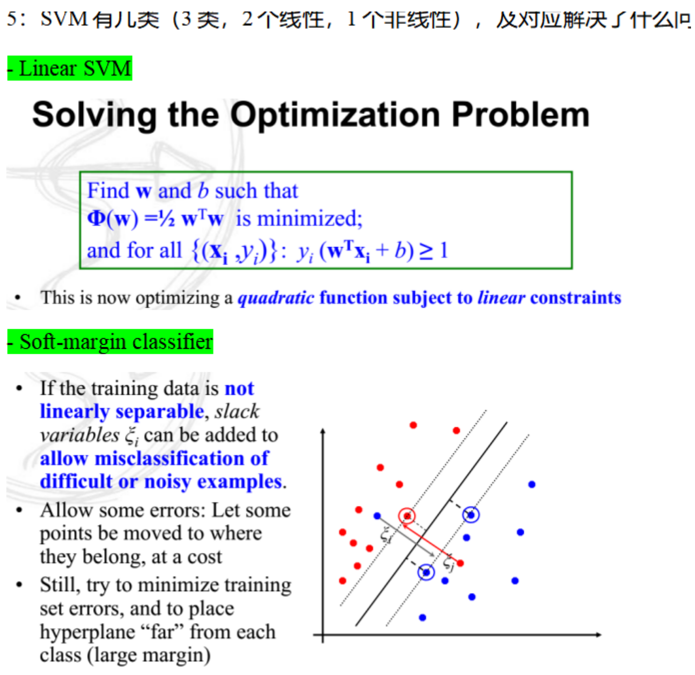
① 支持向量机：

□位于规范超平面上的数据集点称为支持向量
□典型超平面和分离超平面之间的距离是边距
□支持向量同样接近最优超平面
 支持向量是定义最佳分离超平面和最难分类的训练样本
□支持向量是分类任务中信息量最大的样本。

The purpose of SVM is to maximize all the data close to the boundary line, so that different points can be classified and close to the boundary line.

The linear separability is a property of a pair of sets of points. This is most easily visualized in two dimensions by thinking of one set of points as being colored blue

and the other set of points as being colored red. These two sets are linearly separable if there exists at least one line in the plane with all of the blue points on one side of the line and all the red points on the other side. This idea immediately generalizes to higher-dimensional Euclidean spaces if line is replaced by hyperplane.
线性可分离性是一对点集的属性。 通过将一组点视为蓝色，而将另一组点视为红色，可以最容易地在二维上看到这一点。 如果平面中至少存在一条线，而所有蓝点在该线的一侧，而所有红点在另一侧，则这两组线可线性分离。 如果用超平面代替线，则此思想立即推广到高维欧几里德空间。



5：SVM 有几类（3 类，2 个线性，1 个非线性），及对应解决了什么问

- Linear SVM

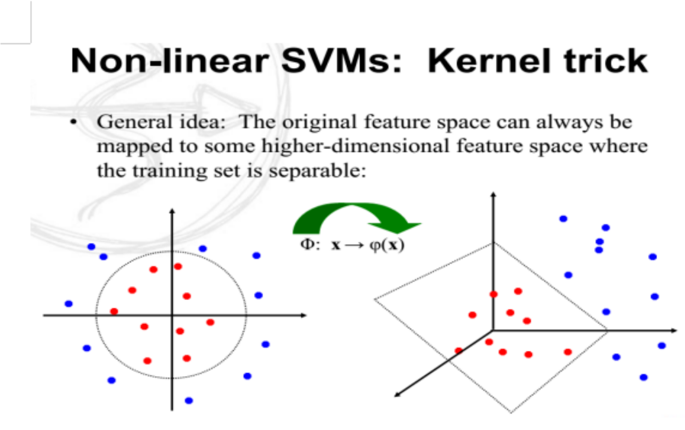
Solving the Optimization Problem

Find $\mathbf{w}$ and b such that
$\mathbf{\Phi}(\mathbf{w}) = ½\, \mathbf{w}^T\mathbf{w}$ is minimized;
and for all $\{(\mathbf{x_i}, y_i)\}: y_i(\mathbf{w}^T\mathbf{x_i} + b) \geq 1$

- This is now optimizing a ***quadratic* function subject to *linear* constraints**

- Soft-margin classifier

- If the training data is **not linearly separable**, *slack variables* ξ_i can be added to **allow misclassification of difficult or noisy examples.**
- Allow some errors: Let some points be moved to where they belong, at a cost
- Still, try to minimize training set errors, and to place hyperplane "far" from each class (large margin)



Non-linear SVMs: Kernel trick

- General idea: The original feature space can always be mapped to some higher-dimensional feature space where the training set is separable:

③What is the significance of support vectors in a support vector classification method?
支持向量在支持向量分类方法中的意义是什么？
They determine the position and orientation of the decision boundary, make the model

robust to outliers, contribute to memory efficiency 它们决定了决策边界的位置和方向，使模型对异常值具有鲁棒性，有助于提高内存效率

Two-class VS Multi-class
In the one-against-all approach, for a multi-classification problem with k classes, we take each class as a positive example and the other k-1 classes as negative examples. We then build k binary classifiers, each specialized to distinguish one class from the others.在 One-against-all 方法中，对于一个具有 k 个类别的多分类问题，我们将每个类别作为一个正例，将其他 k-1 个类别作为负例。然后，我们构建 k 个二分类分类器，每个分类器都专门用于将一个类别与其他类别进行区分。
One-against-one (one-to-one) is another common multi-classification strategy, which is also known as One-vs-One.
In the One-against-one method, for a multi-classification problem with k classes, we build k * (k-1) / 2 binary classifiers, each dedicated to distinguishing two classes . That is, we build a binary classifier for each pair of classes.One-against-one（一对一）是另一种常见的多分类策略，它也被称为 One-vs-One。
在 One-against-one 方法中，对于一个具有 k 个类别的多分类问题，我们构建 k * (k-1) / 2 个二分类器，每个分类器都专门用于将两个类别进行区分。也就是说，我们针对每对类别都构建一个二分类器。
Direct Multi-Class Classification:
In this approach, a multi-class classification algorithm is used directly, without transforming the problem into multiple binary classification tasks. These algorithms are specifically designed to handle multi-class classification problems, and they can assign instances to multiple classes simultaneously.

6.Kernel methods

ascending dimension

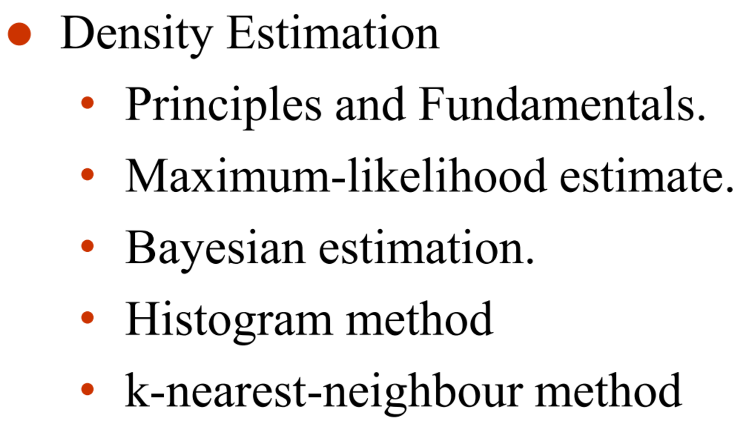
(1)Principles of kernels and kernelization

1. Kernels are a way to implicitly map data points from a lower-dimensional space to a higher-dimensional space without explicitly calculating the coordinates of the higher-dimensional representation. Kernels leverage the kernel trick, which allows us to compute the dot product between the transformed data points in the higher-dimensional space without actually performing the transformation.
2. Kernelization is the process of applying the kernel trick to algorithms that operate in the input space directly, allowing them to operate in the transformed feature space. By kernelizing an algorithm, we can take advantage of the benefits of working in higher-dimensional spaces without explicitly computing the transformation.

Kernel function or kernel method is an efficient tool to map low dimensional data into

high dimensional features. It not only helps us to deal with some linear inseparable problems, but also keeps the computational time complexity in low dimension
核函数或核方法是将低维度数据映射成高维度特征的高效工具，它不仅帮助我们完成部分线性不可分问题的处理，同时计算的时间复杂度依然保持在低维度

Support vector machines with kernels



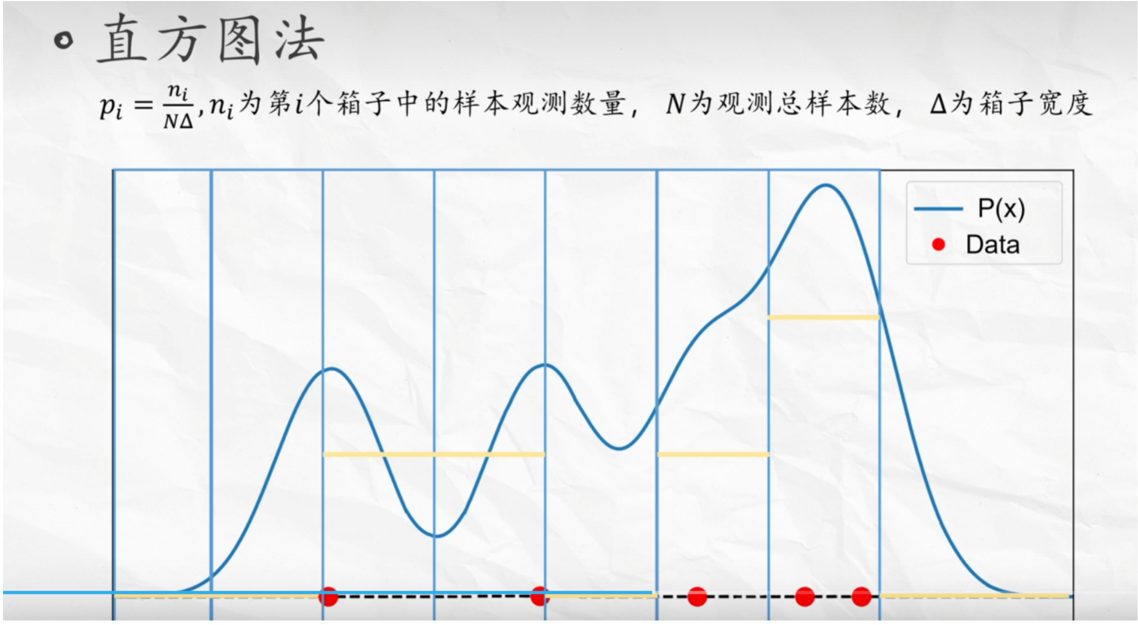
- Density Estimation
  - Principles and Fundamentals.
  - Maximum-likelihood estimate.
  - Bayesian estimation.
  - Histogram method
  - k-nearest-neighbour method

Use samples of the training data to estimate the unknown quantities and use the estimates in the design of the optimal classifier.Assumption about the form of the conditional density (e.g. normal density) transforms the problem from estimating $p(x|\omega i)$ to estimating the parameters, mean, μi and covariance Σi

.

Maximum-likelihood assumes that the parameter values are fixed but unknown. Best estimate is the value that maximizes the probability of obtaining the observed samples.

Bayesian estimation assumes that the parameters are random variables with some prior distribution. Use observed samples to convert this to a posterior density. Observed samples allow good estimation of parameters - Bayesian learning.



- 直方图法

$p_i = \frac{n_i}{N\Delta}$, n_i为第i个箱子中的样本观测数量， N为观测总样本数， Δ为箱子宽度

k-nearest neighbour approaches the estimation by fixing the probability k/n and determining the volume V which contains k samples centred around point x.

④ TP——将正类预测为正类数
⑤ FN——将正类预测为负类数
⑥ FP——将负类预测为正类数
⑦ TN——将负类预测为负类数
⑧ 由此：
⑨ 精准率定义为：P = TP / (TP + FP)
⑩ 召回率定义为：R = TP / (TP + FN)
11 F1 值定义为： F1 = 2 P*R / (P + R)
12 精准率和召回率和 F1 取值都在 0 和 1 之间，精准率和召回率高，F1 值也会高。

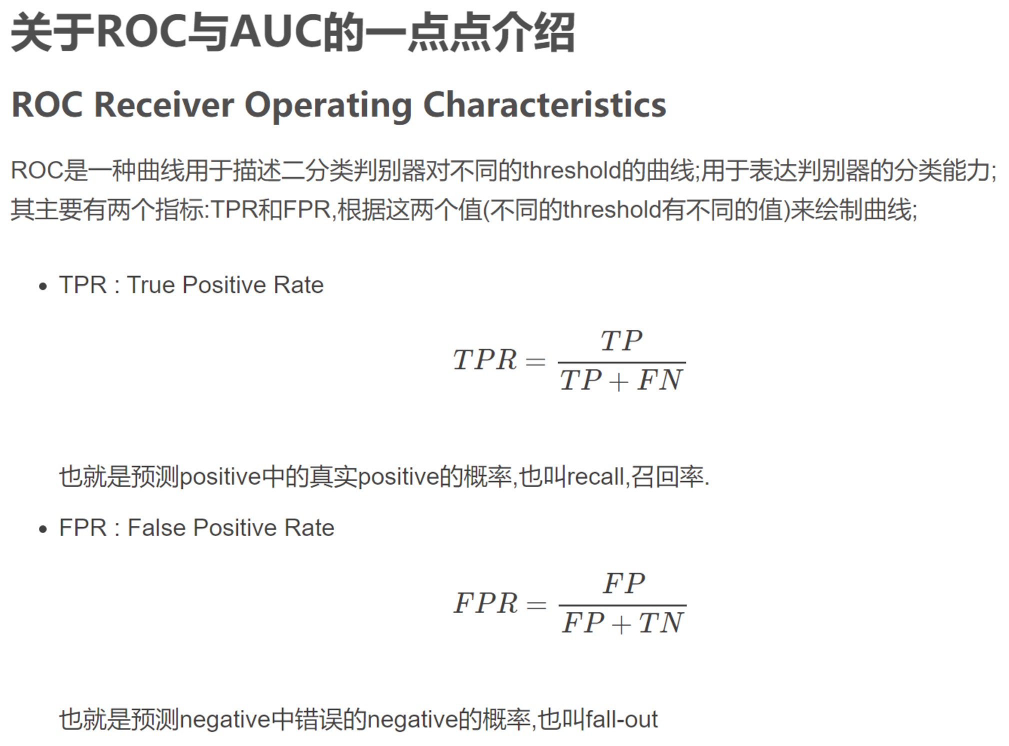
（2） ROC、AUC 评定模型好坏
① ROC 由两个变量 TPR（y 轴），FPR（x 轴）画图
② TPR = TP / (TP + FN); FPR = FP / (FP + TN)
③ AUC 是 ROC 的下方的面积。

7. Performance assessment and evaluation

(1) General ideas of assessment:

The area under the ROC curve (AUC) is a significant performance measure for classifiers. It quantifies the classifier's discrimination ability, robustness to class imbalance, and threshold independence. A higher AUC indicates better classifier performance, making it a valuable tool for evaluating and comparing different classifiers.

ROC 曲线下面积（AUC）是分类器的一个重要性能指标。它量化了分类器的识别能力、对类别不平衡的鲁棒性和阈值独立性。AUC 越高，AUC 表明分类器性能越好，成为评价和比较不同分类器的有价值的工具。



关于ROC与AUC的一点点介绍

ROC Receiver Operating Characteristics

ROC是一种曲线用于描述二分类判别器对不同的threshold的曲线;用于表达判别器的分类能力;
其主要有两个指标:TPR和FPR,根据这两个值(不同的threshold有不同的值)来绘制曲线;

- TPR : True Positive Rate

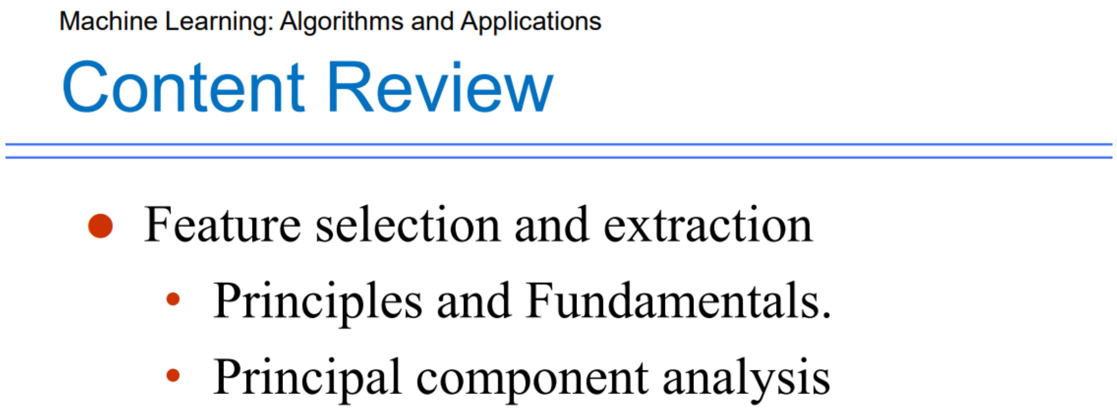
$$TPR = \frac{TP}{TP + FN}$$

也就是预测positive中的真实positive的概率,也叫recall,召回率.

- FPR : False Positive Rate

$$FPR = \frac{FP}{FP + TN}$$

也就是预测negative中错误的negative的概率,也叫fall-out



Machine Learning: Algorithms and Applications

Content Review

- Feature selection and extraction
  - Principles and Fundamentals.
  - Principal component analysis

(1) General considerations of feature selection and feature extraction:

Feature Selection: Identify those variables or features that do not contribute to the classification task. In essence select d features out of available p features.

Feature Extraction: Transform from the p-dimensional feature space to a lower dimensional space.

（1）特征选择和特征提取的一般注意事项：

特征选择：识别那些对分类任务没有贡献的变量或特征。 本质上从可用的 p 个特征中选择 d 个特征。

特征提取：从 p 维特征空间变换到低维空间。

The effect of the two is the same, that is, trying to reduce the number of attributes (or features) in the feature data set; But the two methods are different.
Feature extraction method is mainly through the relationship between attributes, such as the combination of different attributes to get new attributes, which changes the original feature space.

The method of feature selection is to select a subset from the original feature data set, which is an inclusive relationship and does not change the original feature space
两者的效果是相同的，即尽量减少特征数据集中属性（或特征）的个数；但这两种方法是不同的。
特征提取方法主要是通过属性之间的关系，如不同属性的组合得到新的属性，从而改变原有的特征空间。
特征选择的方法是从原始特征数据集中选择一个子集，该子集是一个包含关系，不改变原始特征空间

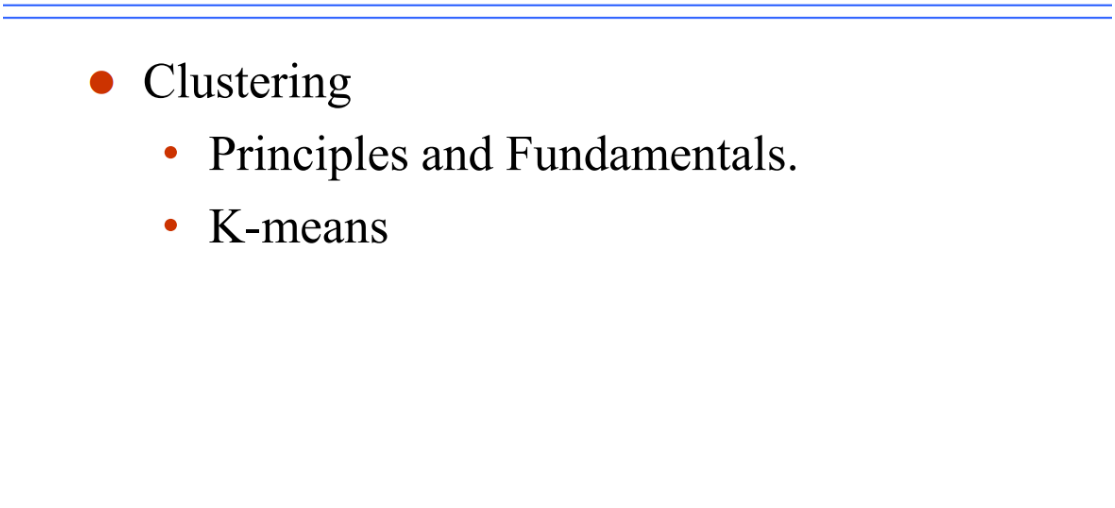
装袋(bagging)是一种根据均匀分布从数据集中重复抽样（有放回）的技术。通过有放回的抽样构建出个自助样本集，每个自助样本集与原始训练集一样大。然后分别在个自助样本集上训练出个基分类器，最后对所有检验样本进行投票判决分类，选取票数最多的一类作为预测输出.
Bagging is a technique of repeated sampling from data sets according to uniform distribution. A self-help sample set is constructed by sampling with put back, and each self-help sample set is as large as the original training set. Then, base classifiers are trained on each self-help sample set. Finally, all test samples are classified by voting decision, and the class with the largest number of votes is selected as the prediction output

多专家组合（multiexpert combination）
全局方法（Global）：并行架构，给定一个输入，所有的基学习器产生一个输出，

然后使用所有的输出获得一个判断，如投票和层叠；
局部方法（Local）：与全局方法不同的是，存在一个模型考察输入，并选择一个或几个学习器来产生输出，比如混合专家；
Multi expert combination
Global: parallel architecture, given an input, all base learners produce an output, and then use all the outputs to obtain a judgment, such as voting and cascading;
Local method: different from global method, there is a model to examine the input and select one or several learners to produce the output, such as mixed experts;

投票

组合多个学习器最常用的方法就是投票，原理就是多个学习器并行处理相同的输入，然后对每个输出“求和”，这里的“求和”只是代表一种运算，并不限于加法。基本的框架如图：（W1、W2、W3 代表学习器的权重）
The most common way to combine multiple learners is voting. The principle is that multiple learners process the same input in parallel, and then "sum" each output. The "sum" here only represents an operation, not limited to addition. The basic framework is as follows: (W1, W2, W3 represent the weight of the learner)



- Clustering
  - Principles and Fundamentals.
  - K-means

Clustering：
The process of dividing a collection of physical or abstract objects into multiple classes composed of similar objects is called clustering. The cluster generated by clustering is a set of data objects, which are similar to the objects in the same cluster and different from the objects in other clusters.
It is an unsupervised learning in Machine learning. (no labels)
将物理或抽象对象的集合分成由类似的对象组成的多个类的过程被称为聚类。由聚类所生成的簇是一组数据对象的集合，这些对象与同一个簇中的对象彼此相似，与其他簇中的对象相异

K-means is a super simple clustering method. The main reason for its simplicity is that it only needs to set a k value when using it (setting needs to aggregate data into

several categories)
1) Choose k points as the center of each category.
2) Calculate the distance between all other points and the center k and classify the related points into the same category.
3) According to the points of each category, recalculate the center of each category.
4) Repeat the step 2 and 3 until the center stop changing.

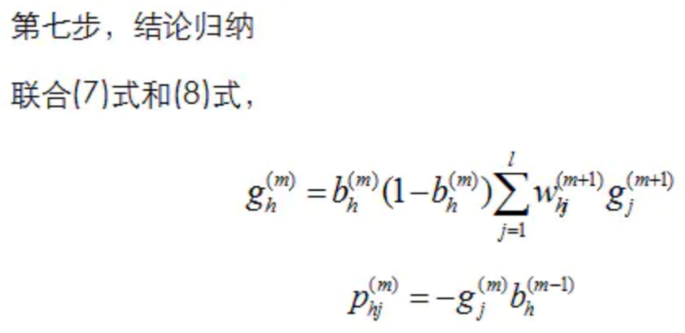
Ways:
1. quick choose:

$$K \approx \sqrt{n/2}$$

1. Elbow method（肘部法）： 肘部法通过绘制不同 K 值下的聚类误差（或称为 SSE，Sum of Squared Errors）与 K 值的关系图像来选择 K 值。聚类误差是指每个样本点到其所属聚类中心的距离的平方和。在图像中，随着 K 值的增加，聚类误差会逐渐减小，但当 K 值增加到一定程度后，聚类误差的下降速度会变得平缓。选择肘部法通常是在图像上找到一个“肘部”点，即在该点之后，聚类误差的下降速度不再显著。这个“肘部”对应的 K 值可以作为合适的聚类数量。
2. Gap statistic（间隔统计量）： 间隔统计量是一种与随机数据进行比较的方法，用于选择最佳的 K 值。该方法计算真实数据和随机数据之间的差异。对于每个 K 值，算法会生成多个随机数据集，并计算它们的聚类误差。然后，通过比较真实数据的聚类误差与随机数据的聚类误差，计算出一个“间隔统计量”。选择最佳的 K 值是通过找到间隔统计量达到最大值的 K 值。
3. Silhouette Coefficient（轮廓系数）： 轮廓系数是一种衡量聚类质量和样本点之间的紧密度与分离度的指标。对于每个样本点，轮廓系数计算其与同簇其他样本点的平均距离（a）和与最近邻不同簇的样本点的平均距离（b），然后计算轮廓系数为(b - a) / max(a, b)。轮廓系数的取值范围在[-1, 1]之间，数值越接近 1 表示样本点越好地被分配到正确的簇中。通过计算不同 K 值下的平均轮廓系数，并选择最大的平均轮廓系数所对应的 K 值。

Why we need to calculate many times of the k
Answer:Because the k is decided by ourselves, Repeated computation prevents local optimization.

BP 算法



第七步，结论归纳

联合(7)式和(8)式，

$$g_h^{(m)} = b_h^{(m)}(1-b_h^{(m)})\sum_{j=1}^{l} w_{hj}^{(m+1)} g_j^{(m+1)}$$

$$p_{hj}^{(m)} = -g_j^{(m)} b_h^{(m-1)}$$

生成式 Generative Model 模型和判别式 Discriminant Model 模型

④ 对于判别式模型，只需要学习二者差异即可。比如说猫的体型会比狗小一点。

⑤ 而生成式模型则不一样，需要学习猫张什么样，狗张什么样。有了二者的长相以后，再根据长相去区分。

⑥ 区别：对于输入 x 和类别标签 y，生成式模型估计它们的联合概率分布 P(x,y)，判别式模型估计条件概率分布 P(y|x)。生成式模型可以根据贝叶斯公式得到判别式模型，但反过来不行。

⑦ 生成式模型：朴素贝叶斯 naive Bayes，KNN，HMM，贝叶斯网络，混合高斯模型，马尔可夫随机场。

⑧ 判别式模型:判别式分析 discriminant analysis，线性回归 Linear Regression，逻辑回归 Logic Regression，神经网络 NN，支持向量机 SVM，高斯过程 Gaussian Process，条件随机场 CRF，CART Classification and Regression Tree

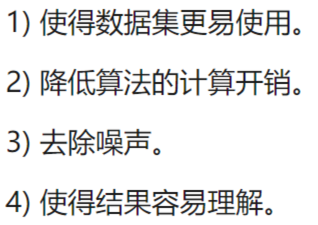
降维方法 dimension reduction

⑨ LASSO：通过参数缩减达到降维的目的

⑩ 主成分分析法 PCA principal component analysis：又称 K-L 变换，通过线性变换将原始数据变换为一组各维度线性无关的表示，用于提取数据的主要特征分量，常用于高维度数据的降维。

11 线性判别式分析 Linear discriminant analysis：LDA，模式识别经典算法。

降维的优点：



1) 使得数据集更易使用。

2) 降低算法的计算开销。

3) 去除噪声。

4) 使得结果容易理解。

Reduction of dimensionality: can simplify the model, potentially improving its interpretability and computational efficiency.

Noise reduction: some of the noise present in the original measurements may be eliminated.

PCA 的过程

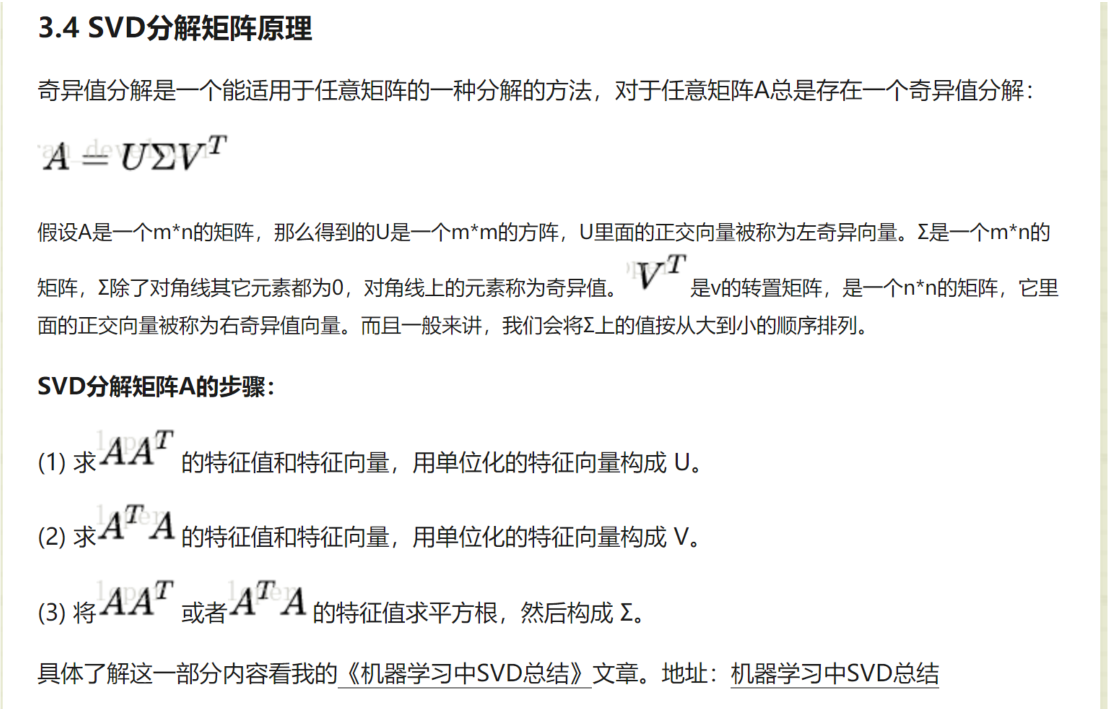
Definention:

The main idea of PCA is to map the n-dimensional feature onto K dimension. The K dimension is a new orthogonal feature, also known as the principal component. It is a K dimension feature reconstructed on the basis of the original n-dimensional feature.

基于特征值分解协方差矩阵实现 PCA 算法、基于 SVD 分解协方差矩阵实现 PCA 算法

PCA algorithm based on eigenvalue decomposition covariance matrix, PCA algorithm based on SVD decomposition covariance matrix

1. 算出均值 2.算出方差 3.算出协方差



3.4 SVD分解矩阵原理

奇异值分解是一个能适用于任意矩阵的一种分解的方法，对于任意矩阵A总是存在一个奇异值分解：

$$A = U\Sigma V^T$$

假设A是一个m*n的矩阵，那么得到的U是一个m*m的方阵，U里面的正交向量被称为左奇异向量。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。V^T 是v的转置矩阵，是一个n*n的矩阵，它里面的正交向量被称为右奇异值向量。而且一般来讲，我们会将Σ上的值按从大到小的顺序排列。

SVD分解矩阵A的步骤：

(1) 求 AA^T 的特征值和特征向量，用单位化的特征向量构成 U。

(2) 求 A^TA 的特征值和特征向量，用单位化的特征向量构成 V。

(3) 将 AA^T 或者 A^TA 的特征值求平方根，然后构成 Σ。

具体了解这一部分内容看我的《机器学习中SVD总结》文章。地址：机器学习中SVD总结

SVD process

1:To remove the average value, that is, each bit feature subtracts its own average value.

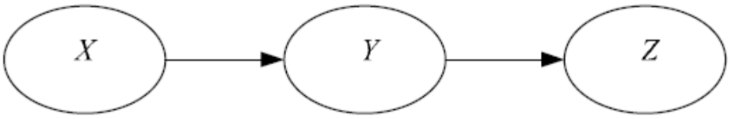
2. The covariance matrix was calculated.
3. The eigenvalues and eigenvectors of covariance matrix are calculated by SVD
4. The data is transformed into a new space constructed by K feature vectors

- Graphical Models
  - Principles and Fundamentals.
  - Connection types of nodes.
  - Probability Derivations based on graph.
  - Principles of HMMs

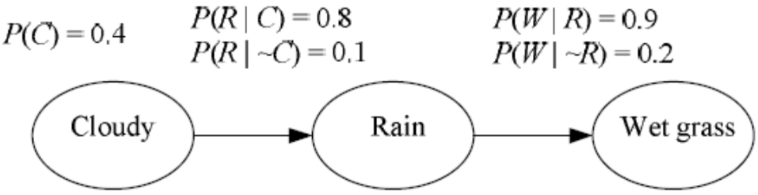
Graphical models:Aka Bayesian networks, probabilistic networks

Case 1: Head-to-Tail

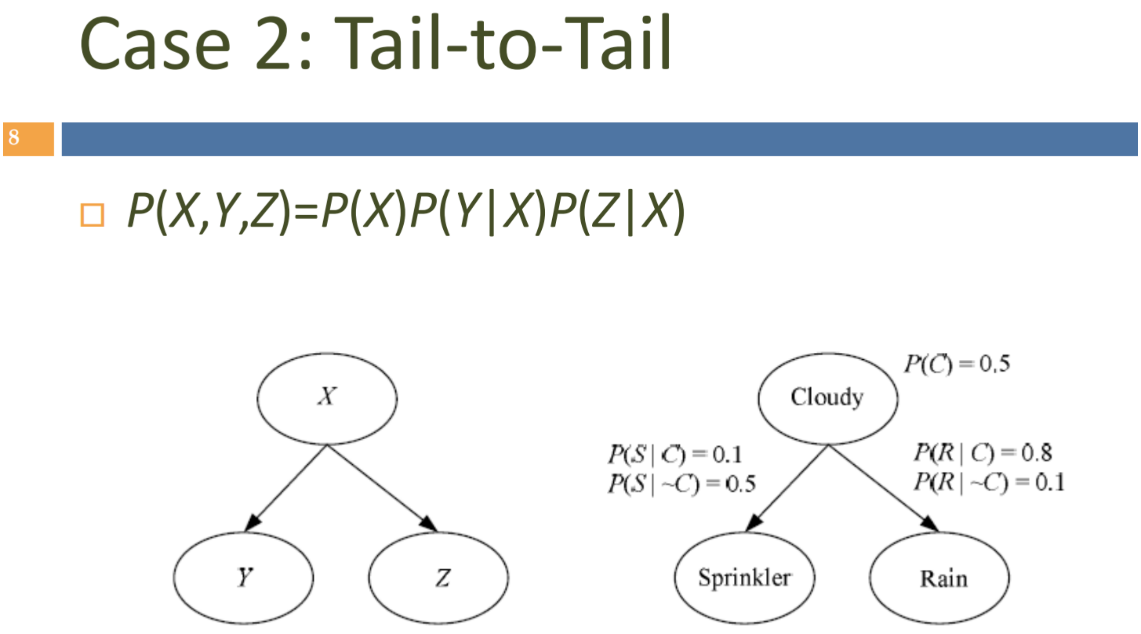
- $P(X,Y,Z)=P(X)P(Y|X)P(Z|Y)$



(a) Model



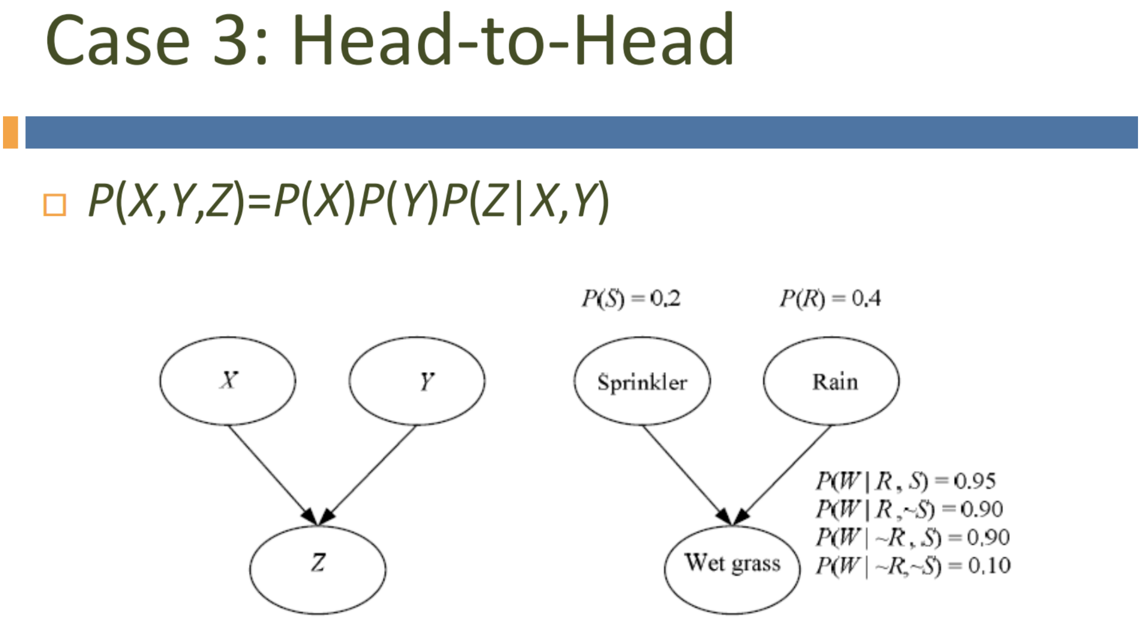
- $P(W|C)=P(W|R)P(R|C)+P(W|\sim R)P(\sim R|C)$



Case 2: Tail-to-Tail

8

□ $P(X,Y,Z)=P(X)P(Y|X)P(Z|X)$



Case 3: Head-to-Head

□ $P(X,Y,Z)=P(X)P(Y)P(Z|X,Y)$

隐马尔科夫模型(Hidden Markov Model,HMM)

HMM 的结构

HMM 是结构最简单的动态贝叶斯网。

如下图所示为 HMM 的结构，HMM 的变量可分为两组：

x 为观测变量

y 为状态变量(隐变量-hidden variable)，y 的取值范围通常是有 N 个可能取值的离散空间

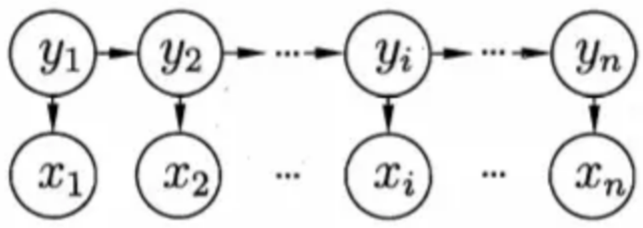


图 14.1 隐马尔可夫模型的图结构

HMM-graph.png

The next state is determined only by the current state and does not depend on any previous state. Y (T) is determined by Y (t-1) and has nothing to do with other states.

CNN（卷积神经网络) Convolutional Neural Networks,
Fully connected deep neural network, as the name suggests, each neuron is connected with the adjacent layer of neurons.
全连接深度神经网络，顾名思义，每个神经元都与相邻层的神经*元连接*
Convolutional Neural Networks (CNNs) are primarily designed for processing grid-like data, such as images or audio spectrograms. Key components of CNNs include convolutional layers, pooling layers, and fully connected layers.
Recurrent Neural Networks (RNNs) are designed to process sequential and time-series data. Unlike feedforward neural networks, RNNs have feedback connections, allowing them to have memory or context.

The Gradient Descent algorithm is an optimization algorithm commonly used in machine learning and deep learning to iteratively minimize the loss function and find the optimal values of the model's parameters.
梯度下降算法是机器学习和深度学习中常用的一种优化算法，用以迭代地最小化损失函数，并寻找模型参数的最优值。

local receptive fields

Our window refers to sliding a pixel, usually called a stride, or sliding multiple steps.
我们的窗口指滑动了一个像素，通常说成一步(stride)，也可以滑动多步

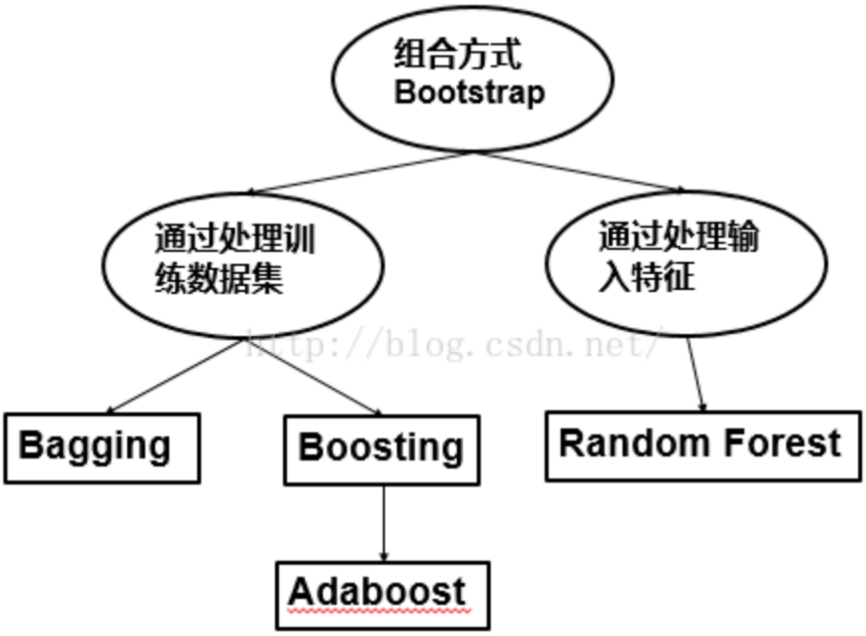
权值共享(Shared weights and biases)

$$\sigma\left(b+\sum_{l=0}^{4}\sum_{m=0}^{4} w_{l,m}a_{j+l,k+m}\right)$$

a1 表示隐藏层的输出，a0 表示隐藏层的输入，而*就表示卷积操作(convolution operation) 这也正是卷积神经网络名字的由来

池化(Pooling)

CNN 还有一个重要思想就是池化，池化层通常接在卷积层后面。池化这个词听着就很有学问，其实引入它的目的就是为了简化卷积层的输出。
从左往右依次是输入层，卷积层，池化层，输出层。输入层到卷积层，卷积层到池化层已经详细介绍过了。池化层到输出层是全连接，这和 DNN 是一样的。



Shallow unsupervised learning, auto edncode, is the compression of the graph, the input data is very large. After training and decoding, error is obtained
Spass gets features between pixels. Select each point, called the base, with different weights.