阿里云天池牛年读书会

机器学习原理、算法与应用
读书分享

分享嘉宾： SIGAI 雷明

SIGAI

天池读书会

TIANCHI天池

清華大學出版社
TSINGHUA UNIVERSITY PRESS

《机器学习的原理：算法与应用》

本书全面系统地讲述了深度学习、机器学习的主要算法。

直播嘉宾：SIGAI CEO 雷明

直播时间：2月24日20:00～21:00

扫码领取读书会相关
学习资源

目录

TIANCHI天池

个人简介

SIGAI创始人，《机器学习-算法、原理与应用》（清华大学出版社），《机器学习的数学》（人民邮电出版社）作者。有超过14年的学术研究与产品研发经验，发表论文数篇。曾就职于百度，任软件工程师/项目经理；zmodo/meshare，任CTO/技术合伙人（创业），估值100亿。目前致力于研发机器学习/深度强化学习框架，www.sigai.cn 。

图书介绍

TIANCHI天池

内容简介

内容全面深入。全书系统地讲解机器学习算法与理论，主要算法的理论讲解透彻、结构清晰，均有详细的推导和证明过程。

内容新。对于深度学习等重点算法，涵盖和反映了截至2017年学术界与工业界的新成果，确保本书的内容能够紧跟当前的学术和技术趋势。

理论与实践相结合。对于所有重点算法，除深入讲解算法的原理之外，还介绍了算法的工程实现细节，对各种算法的实际应用也进行了介绍。

对机器学习所需的数学知识进行全面系统的介绍，确保读者无须单独再看其他数学教材也能顺利学习。

推荐语

全面系统，理实结合，深入浅出，雅俗共赏。涵盖机器学习与深度学习的主流方法与理论，紧密结合工程实践与应用。自成体系，包括数学基础、核心概念、主要算法、开源代码和典型实战方案。适合广大技术人员入门提高。

纽约州立大学、哈佛大学数学科学与应用中心教授，国际微分几何大师丘成桐先生弟子，计算共形几何创始人 **顾险峰**

本书从机器学习相关的基础数学知识入手，到经典机器学习算法的理论推导实践，再到深度学习相关的卷积神经网络、生成式对抗网络和强化学习，帮助读者快速建立起系统性的机器学习知识体系，化繁为简全面掌握机器学习核心知识。相信本书会给各位读者的学习和研究工作带来帮助。

Yi+AI联合创始人、CTO，前阿里巴巴和百度IDL深度学习算法专家 **刘彬**

本书全面覆盖传统统计学习与深度学习的主要算法，使读者既能总览机器学习算法以知其全貌，又能观算法历史沿革而知其得失。书中内容从数学出发，经算法描述，到代码实现，完整呈现了这些方法的来龙去脉，是引领读者从学算法到用算法、写算法的一座桥梁。

清华大学博士，Xilinx工程师 **张振**

人工智能技术人才的培养，离不开高质量的机器学习书籍。本书从基础的数学理论出发，扩展到经典理论算法，以及非常前沿的深度学习相关算法，层层推进，形成系统化技术体系；同时，深入代码细节和理论细节，让读者知其然，同时知其所以然。因此，这是一本不可多得的适合人工智能技术相关人才学习的书籍！

清华大学博士，北京清微智能科技联合创始人、CTO **欧阳鹏**

本书作者依靠自己深厚的积累，使他得以同时具有学术界扎实的理论功底和工业界丰富的实践经验，引导大家深入浅出地从算法、开源代码实现，以及解决工业实际问题各个角度，进行了全面、系统、深入的讲解。本书是一本难得的适合各种不同层次读者，由浅入深全面掌握机器学习知识的优秀教材。

谷歌机器学习开发者专家 **李卓桓**

图书查询·扩展阅读

书圈

清华社官方微信号

扫我有惊喜

ISBN 978-7-302-53234-7
9 787302 532347 >
定价：88.00元

机器学习——原理、算法与应用

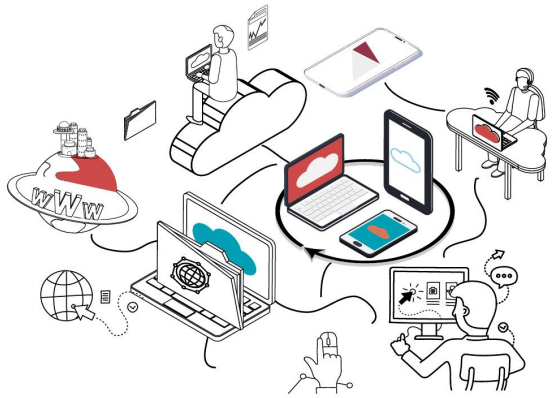
雷明 著

清华大学出版社

一书读懂机器学习算法的原理

学会训练自己的模型，编程实现自己的算法

微课版



机器学习

原理、算法与应用

雷明◎著

- 讲解54种算法，算法涵盖有监督学习、无监督学习和强化学习。
- 配有20个实验程序，包含17份源代码，帮助读者正确地掌握算法与开源库的使用。
- 配有35个讲解视频，对复杂、难以理解的知识点有清晰透彻的讲解。

清华大学出版社

■作者简介

雷明

致力于研发机器学习与深度学习、计算机视觉框架，SIGAI创始人。2009年毕业于清华大学计算机系，获硕士学位，研究方向为机器学习、计算机视觉，发表论文数篇。曾就职于百度公司，任高级软件工程师和项目经理；zmodo/meshare，任CTO与平台研发中心负责人。在机器学习、计算机视觉方向有丰富的学术研究与产品研发经验。

图书介绍

TIANCHI天池

- 机器学习简介
- 若干基本概念
- 有监督学习
- 无监督学习
- 数据生成问题

学习资料、直播回放交流

TIANCHI天池

大家可以使用**电脑**访问下方地址或者扫下方二维码进入天池读书会页面，获取读书会相关学习资料、项目实践代码和分享PPT等资源，还可以提前预约其他读书会直播，查看回放。

https://tianchi.aliyun.com/specials/promotion/activity/bookclub

扫码领取读书会相关学习资源

TIANCHI天池

第1部分-机器学习简介

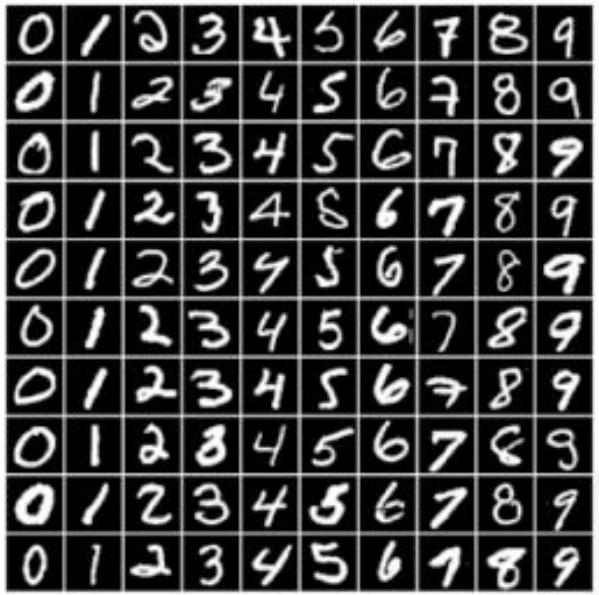
为什么需要机器学习

TIANCHI天池

对于我们习以为常的问题，计算机却非常难以处理。

如何判断一张图像是0-9这些阿拉伯数字中的哪一个?

每个人的书写习惯不同，字体，倾斜角度，笔画宽度也不同。



手写数字识别-MNIST数据集

为什么需要机器学习

如何找出一张图像中的所有行人？

人体是柔性物体，可以变形，从不同角度看，外观不同

衣服，帽子，包等附着物

光照变化

遮挡，部分出现

行人检测问题

为什么需要机器学习

TIANCHI天池

机器翻译的目标是给定一种语言的句子，
找出另一种语言对应的句子的最优方案，
二者有相同的语义

源语言的句子

$$\arg\max_{s_{english}} p\left(s_{english} \middle| \text{"我爱机器学习"}\right)$$

目标语言的句子

使条件概率最大化

机器翻译问题

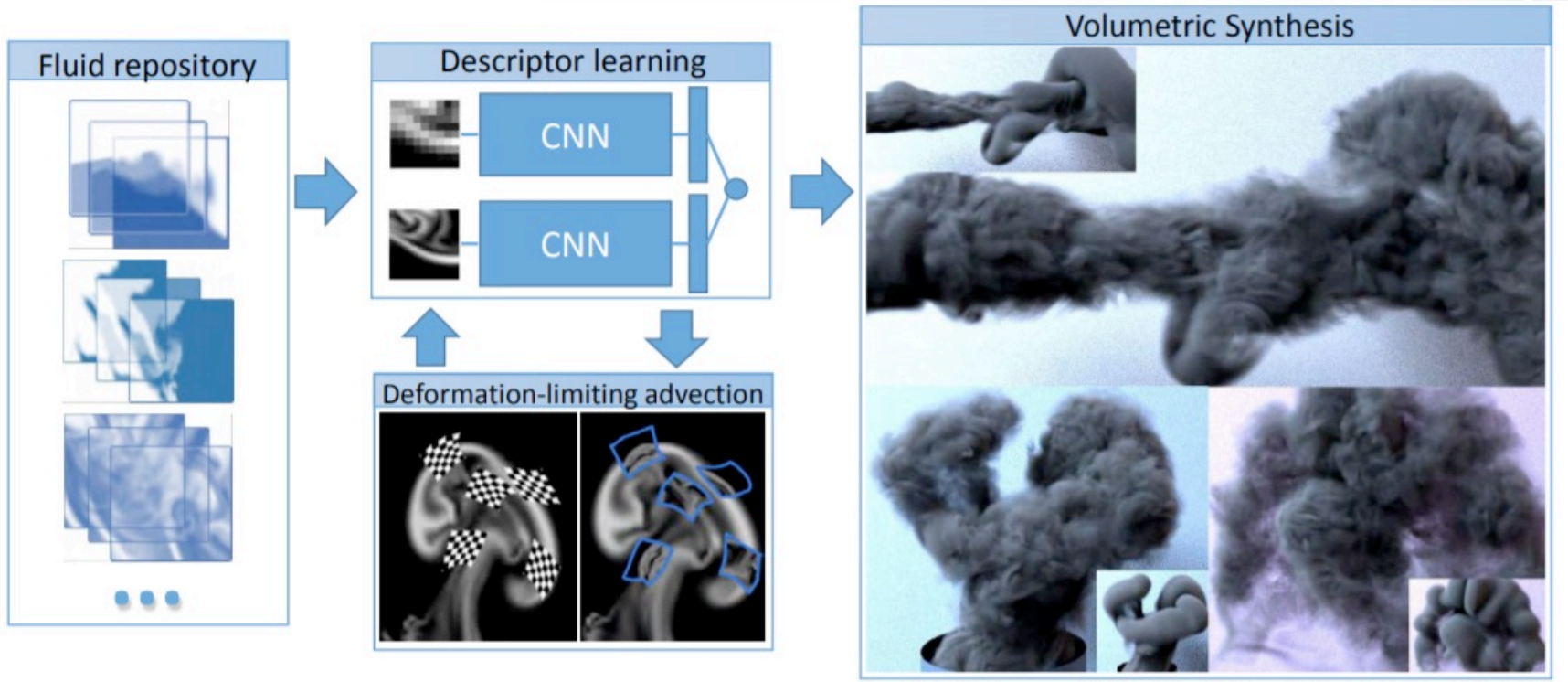
为什么需要机器学习

计算流体力学/CFD - 受力分析，仿真模拟

Navier-Stokes方程难以计算- 无解析解，数值求解的计算量太大

$$\nabla \cdot \mathbf{u} = 0$$

$$\frac{\partial \mathbf{u}}{\partial t} = -\mathbf{u} \cdot \nabla \mathbf{u} - \frac{1}{\rho} \nabla p + \mathbf{f}$$



用深度学习进行流体模拟

机器学习所解决的问题

TIANCHI天池

- 目前尚无无精确的数学、逻辑模型的问题 - 只可意会，不可言传
- 虽然有精确的数学或逻辑模，但直接求解非常困难
- 一句话总结 - 很难直接编程求解

机器学习的思路

TIANCHI天池

与其把知识和经验总结好了告诉计算机，

还不如让计算机自己去学习

模仿人的学习能力，从样本数据中学习经验，将经验用于预测

1980年代开始逐渐成为人工智能的主流方法

TIANCHI天池

第2部分-若干基本概念

机器学习算法的分类

TIANCHI天池

有监督学习

分类问题 是什么？ 图像识别，语音识别

回归问题 是多少？ 预测寿命，预测房价

无监督学习

降维 抽象和压缩

聚类 怎么划分

半监督学习

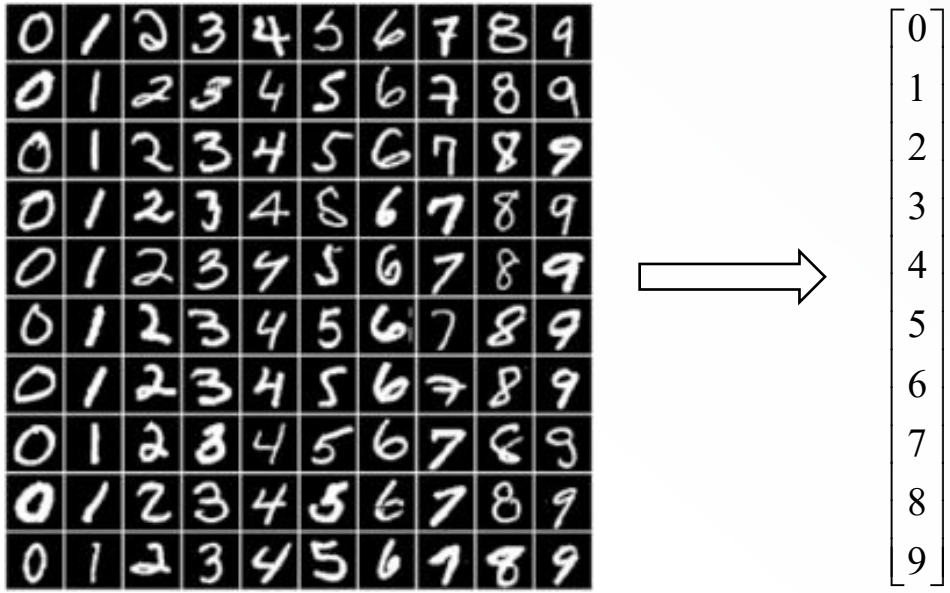
强化学习 怎么做？ 决策，如博弈与游戏，自动驾驶，机器人控制

数据生成问题 怎么造？ 画画，写诗，创作音乐

分类问题

实现从向量到整数的映射

$$\mathbb{R}^n \rightarrow \mathbb{Z}$$



$$\begin{bmatrix} 0 \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \\ 8 \\ 9 \end{bmatrix}$$

手写数字识别

回归问题

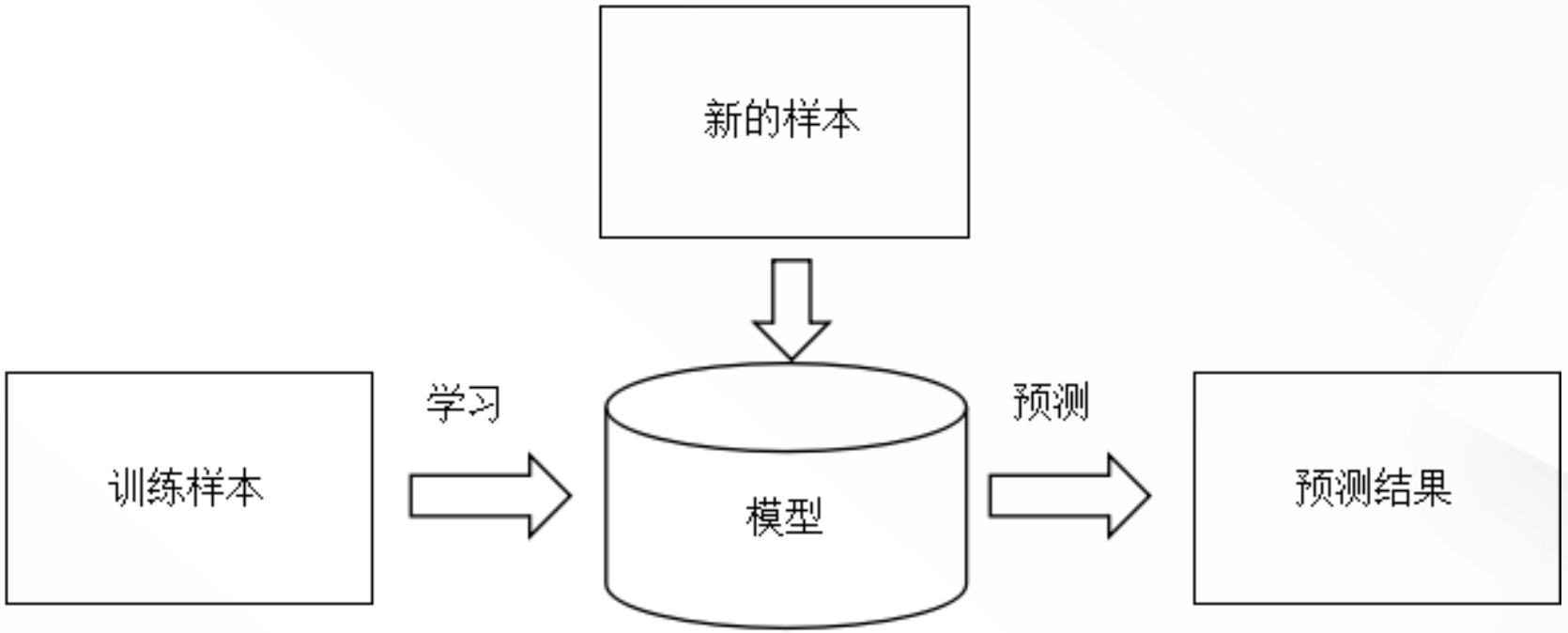
实现从向量到实数的映射

$\mathbb{R}^n \rightarrow \mathbb{R}$

性别	年龄	学历	工作年限	所在城市	行业	收入（万）
男	31	本科	9	北京	金融	80
男	24	本科	2	深圳	金融	20
男	45	博士	18	深圳	互联网	230
女	25	本科	3	深圳	互联网	35
女	27	硕士	2	北京	财务	18
女	35	博士	8	上海	教育	30

根据个人信息预测其收入

有监督学习的一般流程



$$f(\mathbf{x}):\mathbf{x}\rightarrow y$$

iris/鸢尾花数据集

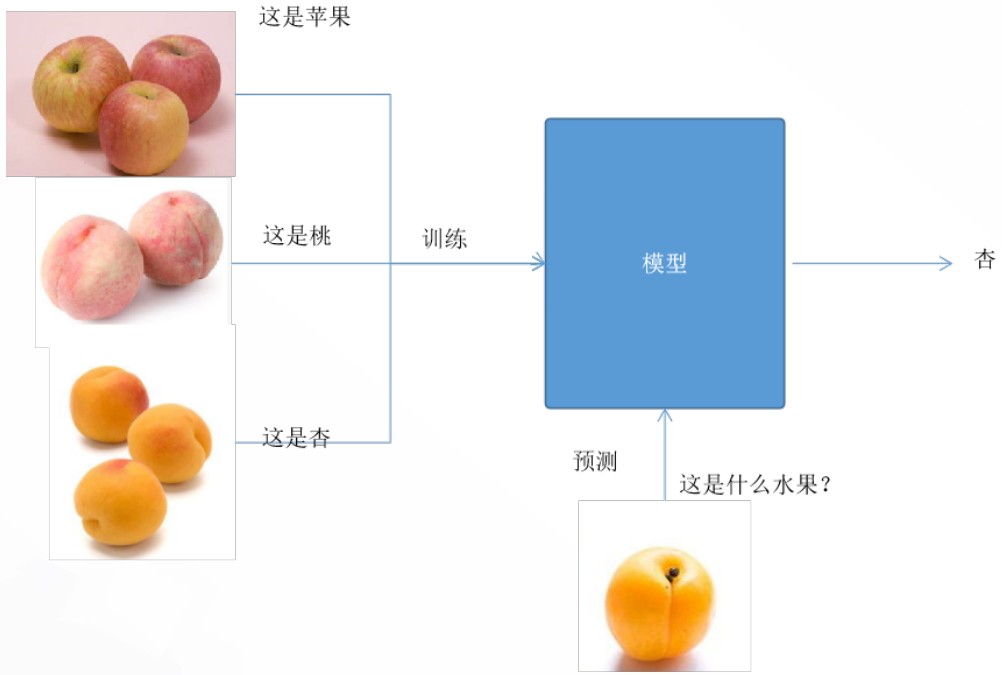
TIANCHI天池

sepal length	sepal width	petal length	petal width	类型
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
7.0	3.2	4.7	1.4	versicolor
6.4	3.2	4.5	1.5	versicolor
6.9	3.1	4.9	1.5	versicolor
6.3	3.3	6.0	2.5	virginica

直播相关资料获取及回放查看地址： https://tianchi.aliyun.com/specials/promotion/activity/bookclub

无监督学习 - 聚类

将样本集合划分成几个子集，保证每个子集内的样本相似，不同子集之间的样本差异很大



有监督的水果分类

无监督的水果聚类

直播相关资料获取及回放查看地址：https://tianchi.aliyun.com/specials/promotion/activity/bookclub

TIANCHI天池

第3部分-有监督学习

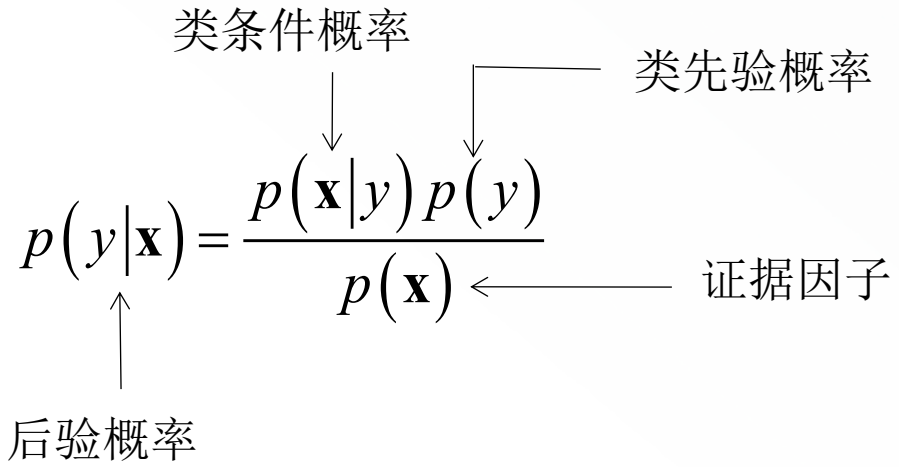
直播相关资料获取及回放查看地址：https://tianchi.aliyun.com/specials/promotion/activity/bookclub

贝叶斯分类器

利用贝叶斯公式进行分类，实现因果推理

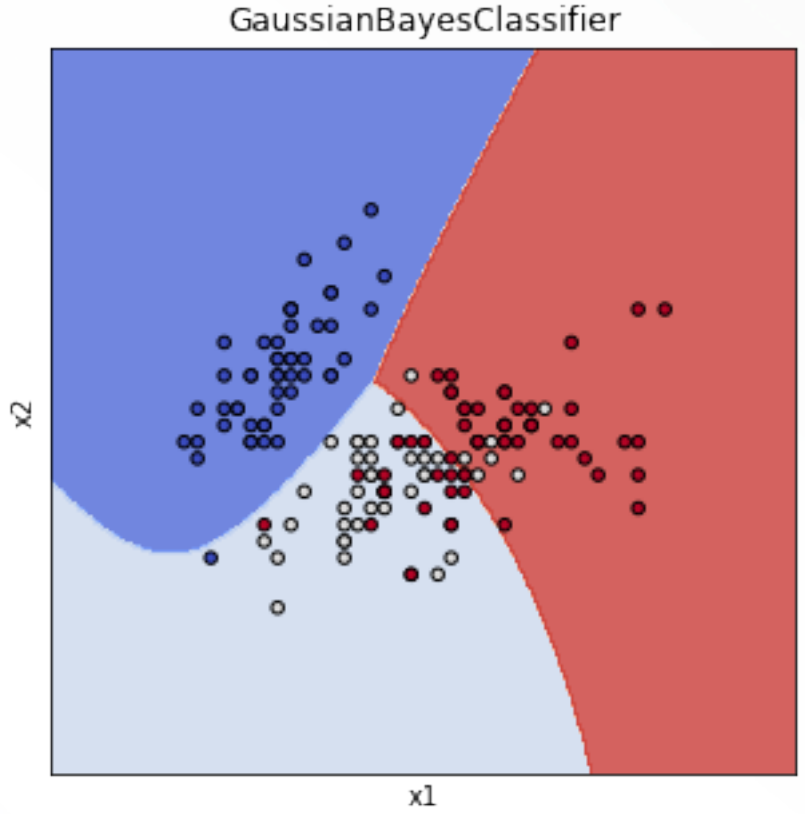
将样本判定为后验概率最大的那个类

贝叶斯公式



最大化后验概率 $\arg\max_y p(\mathbf{x}|y)p(y)$

贝叶斯分类器



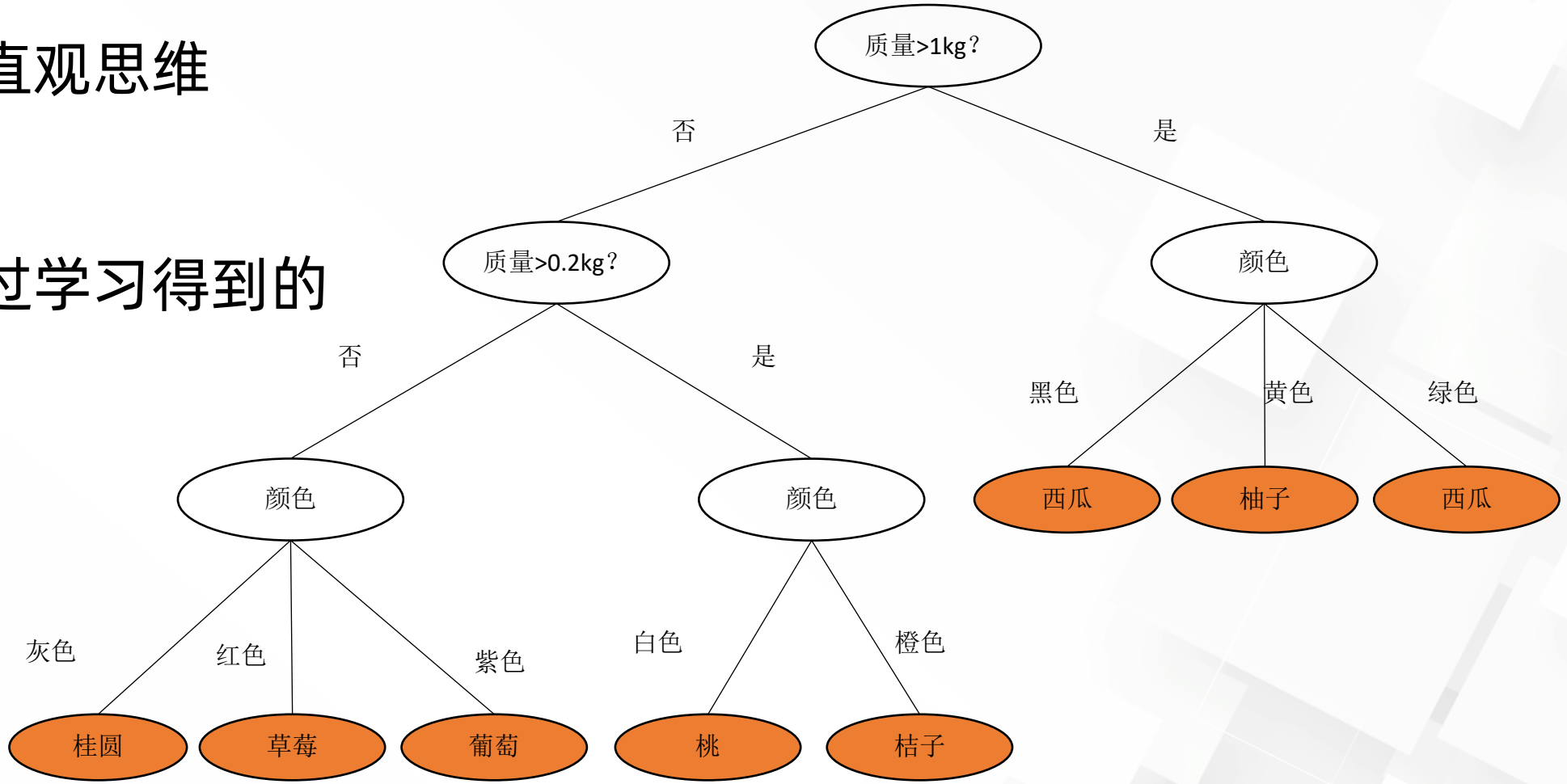
贝叶斯分类器对iris数据集的分类结果

决策树

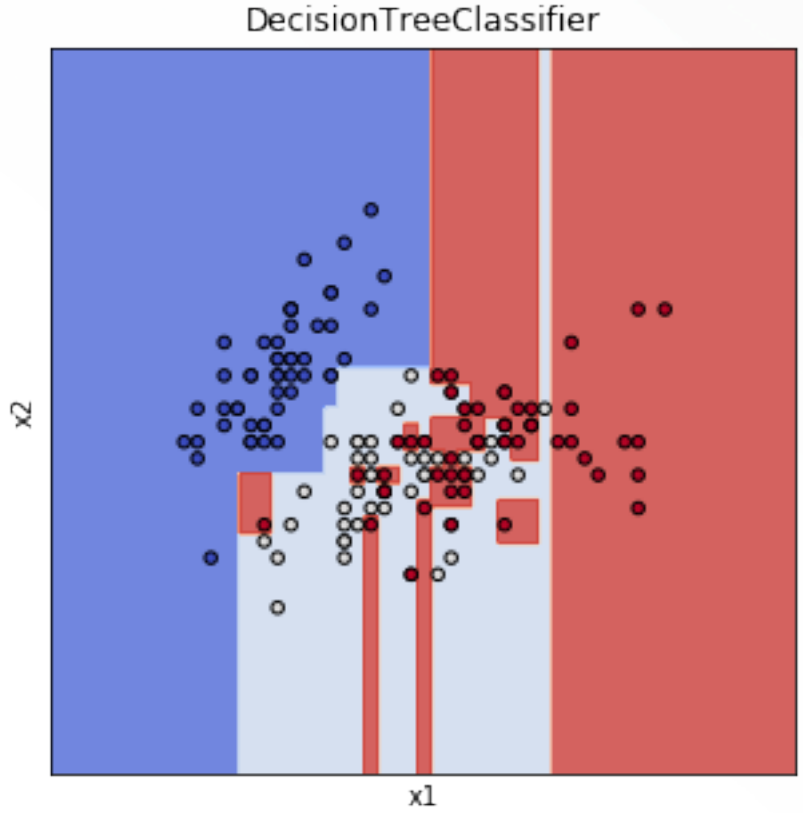
最符合人类的直观思维

一组判定规则

这组规则是通过学习得到的



决策树



决策树对iris数据集的分类结果

直播相关资料获取及回放查看地址：https://tianchi.aliyun.com/specials/promotion/activity/bookclub

logistic回归

直接预测出一个样本是正样本的概率

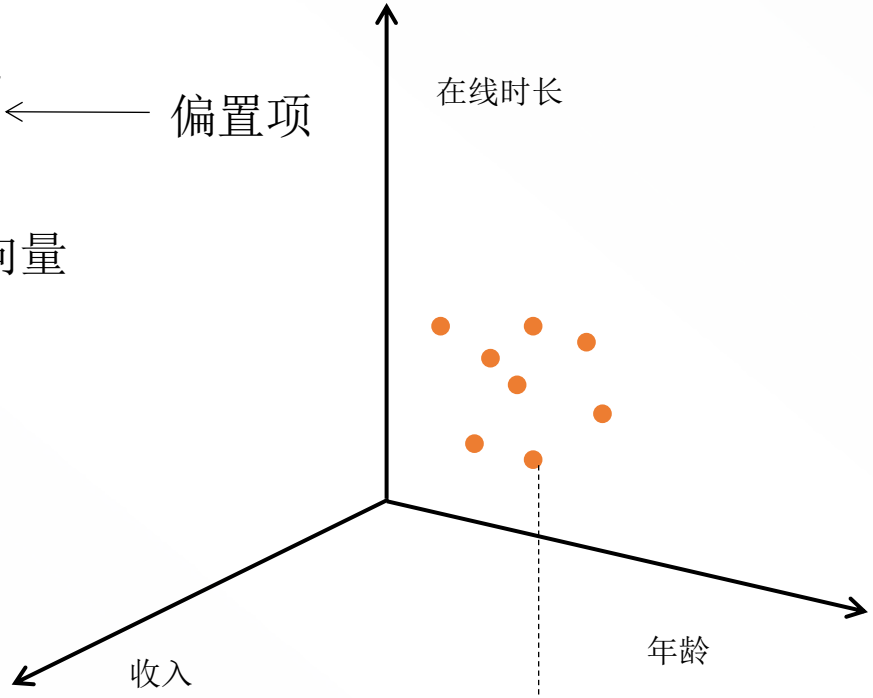
样本是正样本的概率

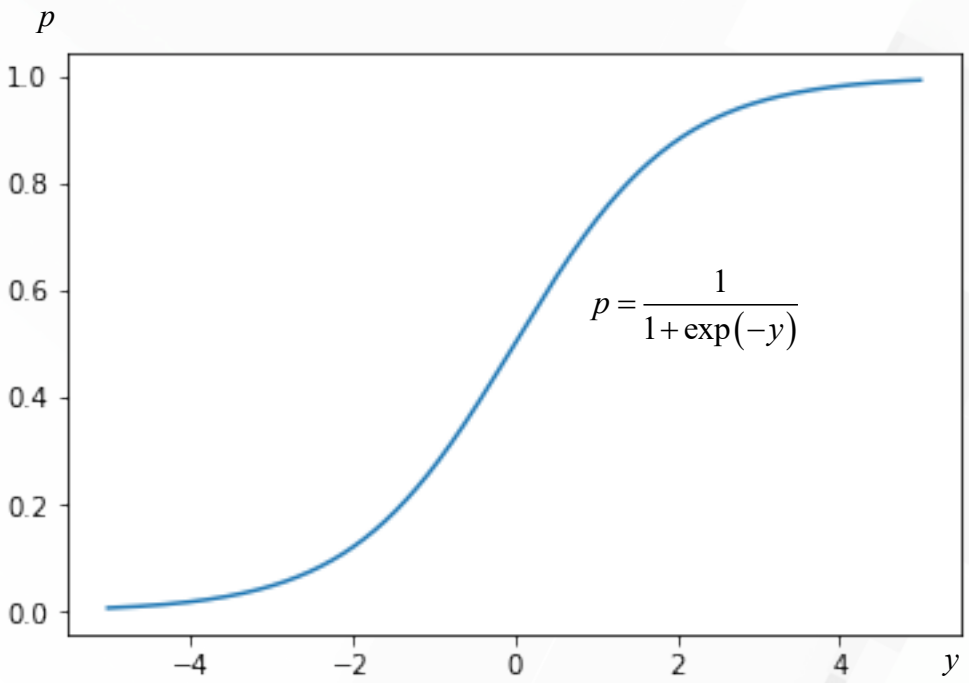
$$h(\mathbf{x}) = \frac{1}{1+\exp\left(-\mathbf{w}^{\mathrm{T}}\mathbf{x}+b\right)}$$

偏置项

权重向量

特征向量





$$y = \mathbf{w}^{\mathrm{T}}\mathbf{x}+b$$

根据一个人的信息预测出他/她是否会购买某一影片

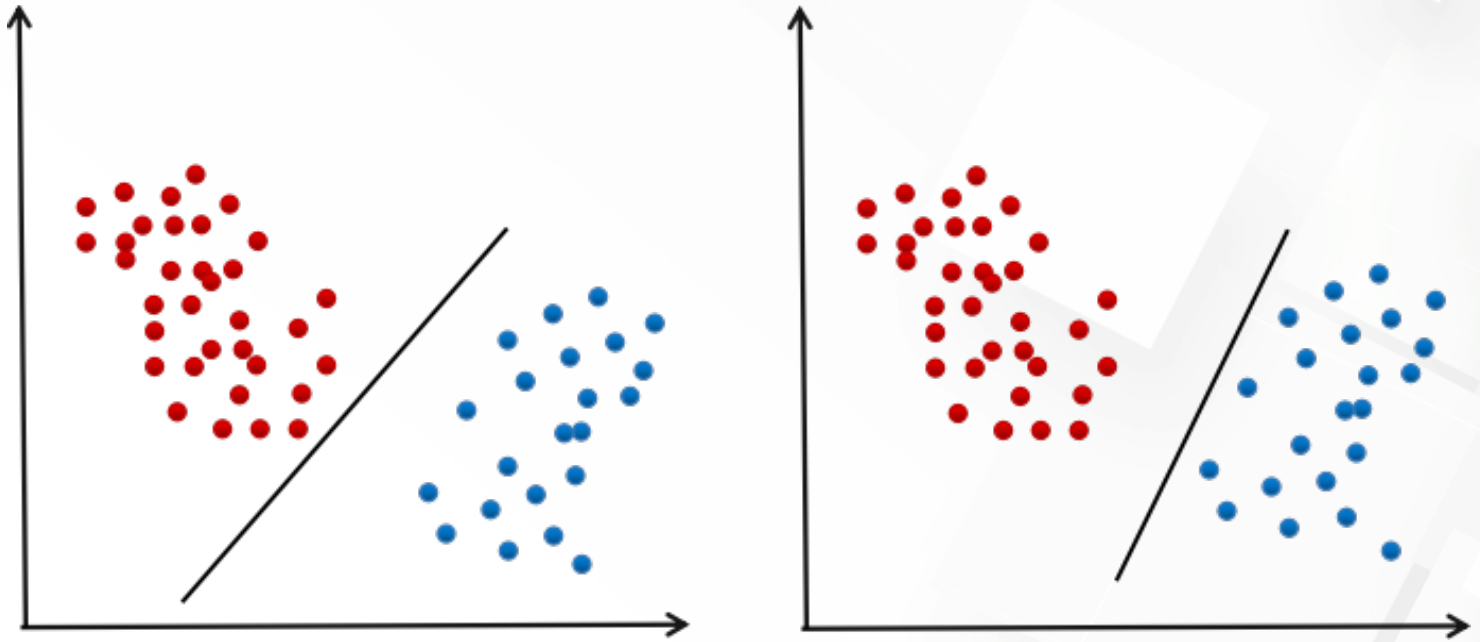
直播相关资料获取及回放查看地址： https://tianchi.aliyun.com/specials/promotion/activity/bookclub

支持向量机/SVM

直线/超平面方程 $\mathbf{w}^{\mathrm{T}}\mathbf{x}+b=0$

用超平面将两类样本分开

$$\operatorname{sgn}\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}+b\right)$$



线性分类器

支持向量机

最大化分类间隔

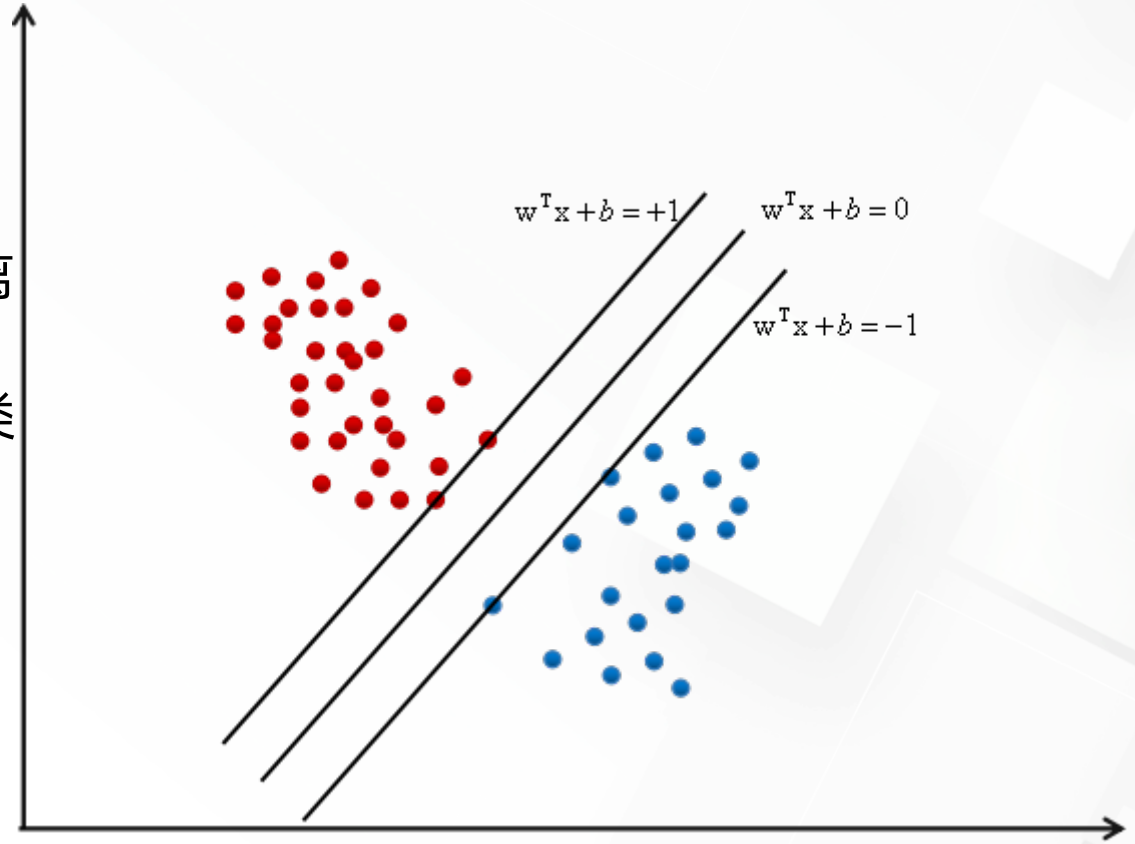
$$d = \frac{\left|\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b\right|}{\|\mathbf{w}\|}$$ ← 最大化样本与分界线的距离

$$y_i\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b\right) \geq 0$$ ← 所有样本都需要被正确分类

简化后的问题

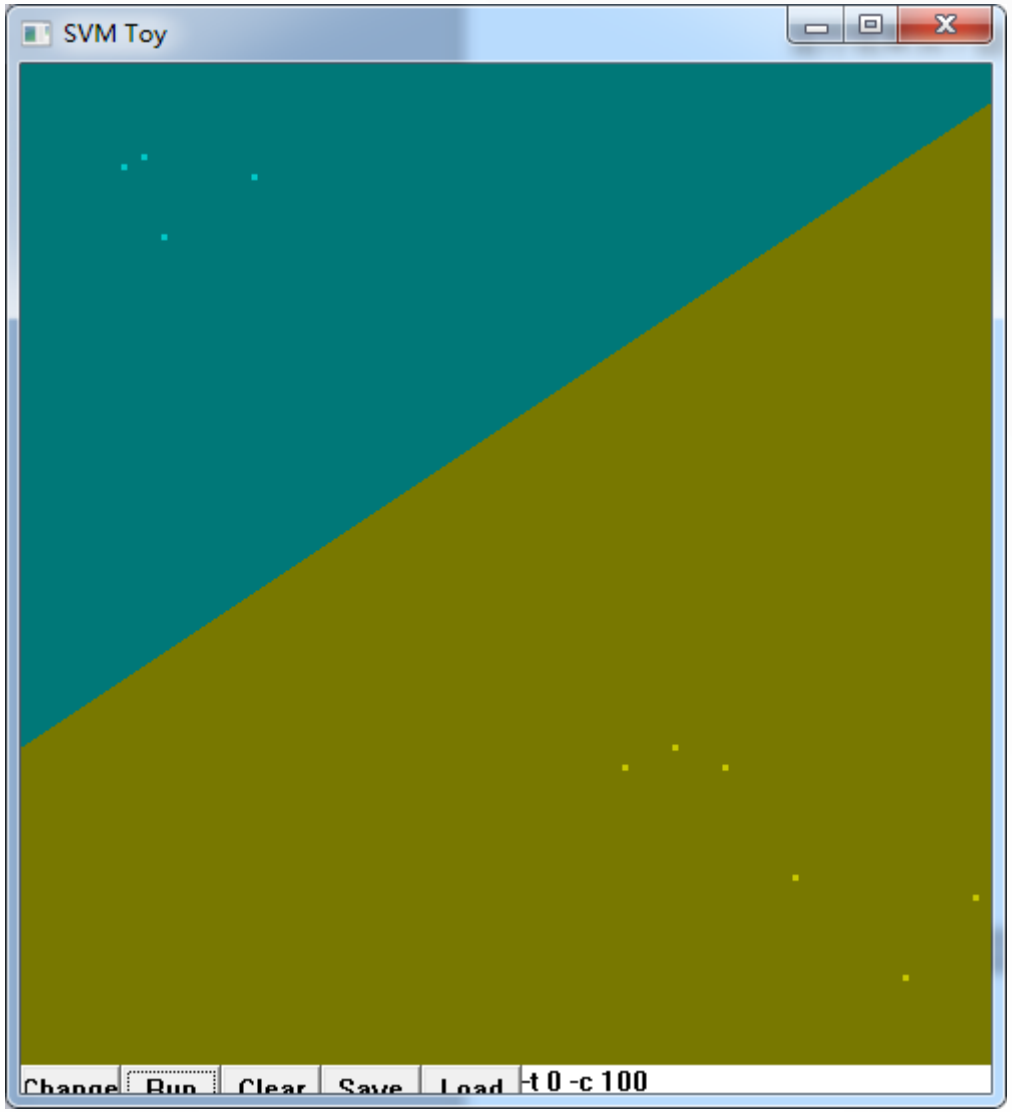
$$\min \frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}$$

$$y_i\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b\right) \geq 1$$



最大化分类间隔

支持向量机



线性可分的支持向量机

支持向量机

通过核函数将向量映射到高维空间，

可以解决非线性分类问题

加上核函数之后的训练问题为

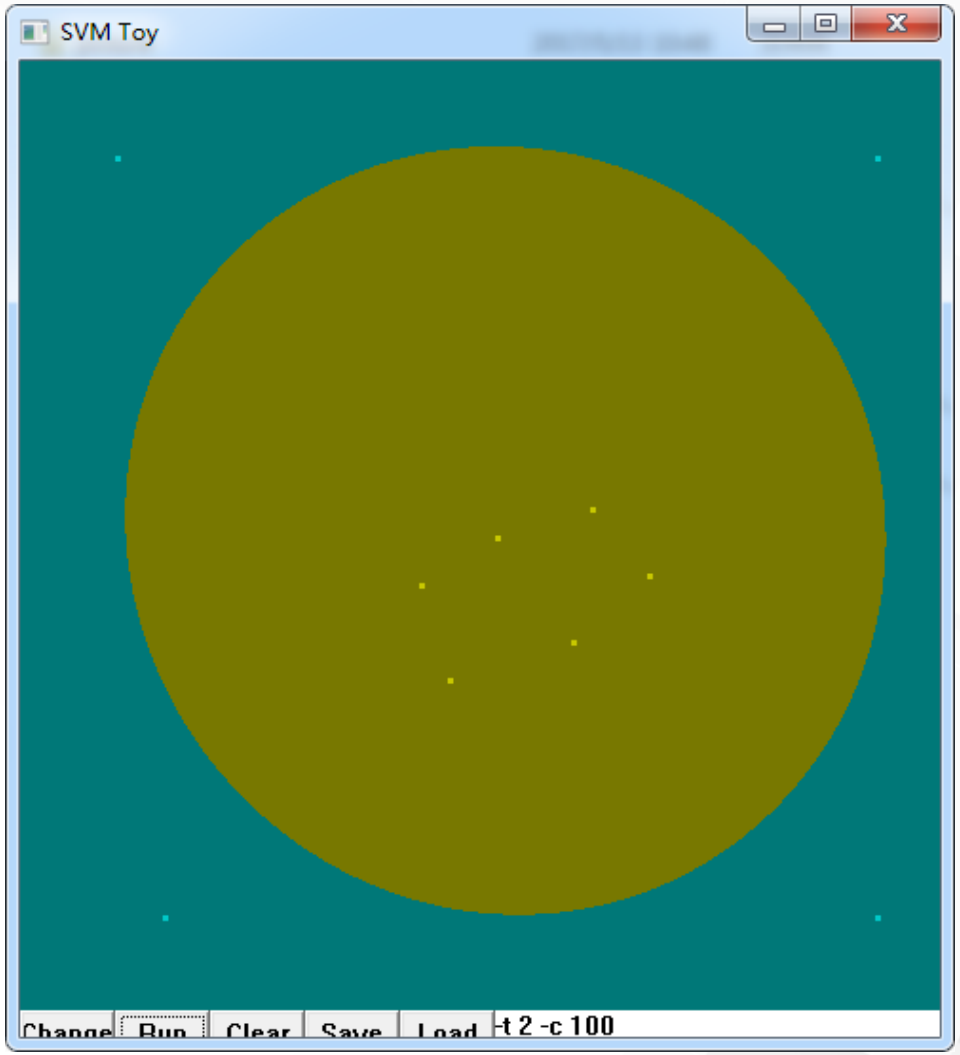
$$\min_{\alpha} \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j y_i y_j K\left(\mathrm{x}_i,\mathrm{x}_j\right)-\sum_{i=1}^{l}\alpha_i$$

$$0 \le \alpha_i \le C$$

$$\sum_{j=1}^{l}\alpha_j y_j = 0$$

预测函数为

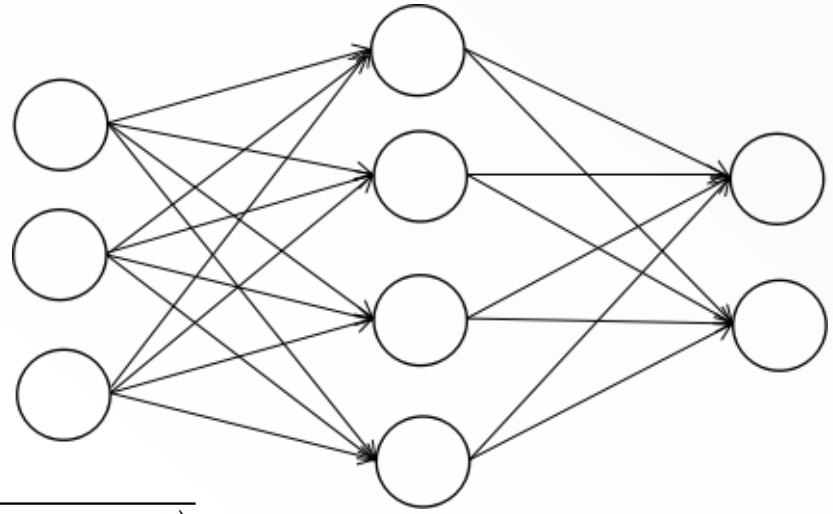
$$\operatorname{sgn}\left(\sum_{i=1}^{l}\alpha_i y_i K\left(\mathrm{x}_i,\mathrm{x}\right)+b\right)$$



用核函数解决线性不可分的问题

人工神经网络

输入层　　隐含层　　输出层



$$y_1 = \frac{1}{1+\exp\left(-\left(w_{11}^{(1)}x_1 + w_{12}^{(1)}x_2 + w_{13}^{(1)}x_3 + b_1^{(1)}\right)\right)}$$

$$y_2 = \frac{1}{1+\exp\left(-\left(w_{21}^{(1)}x_1 + w_{22}^{(1)}x_2 + w_{23}^{(1)}x_3 + b_2^{(1)}\right)\right)}$$

$$y_3 = \frac{1}{1+\exp\left(-\left(w_{31}^{(1)}x_1 + w_{32}^{(1)}x_2 + w_{33}^{(1)}x_3 + b_3^{(1)}\right)\right)}$$

$$y_4 = \frac{1}{1+\exp\left(-\left(w_{41}^{(1)}x_1 + w_{42}^{(1)}x_2 + w_{43}^{(1)}x_3 + b_4^{(1)}\right)\right)}$$

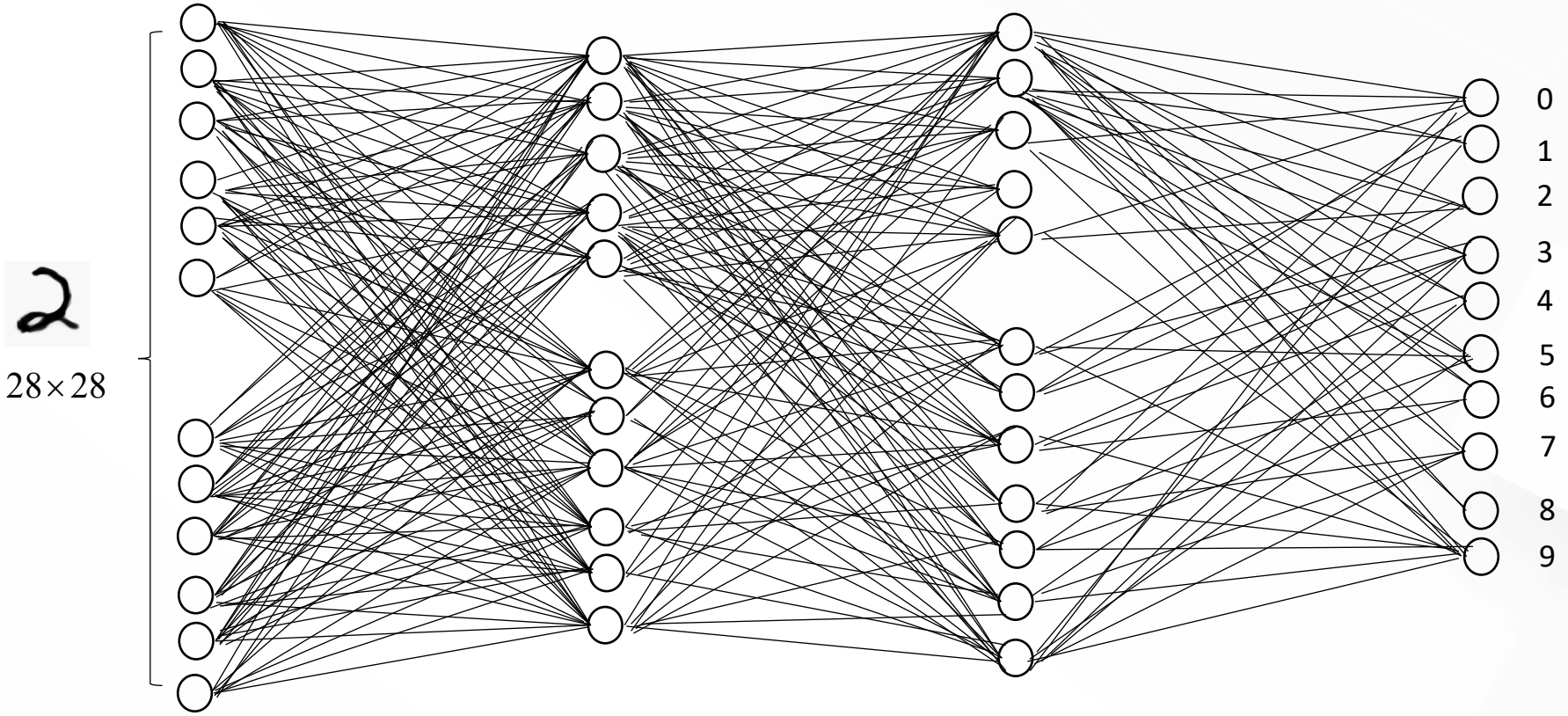
$$z_1 = \frac{1}{1+\exp\left(-\left(w_{11}^{(2)}y_1 + w_{12}^{(2)}y_2 + w_{13}^{(2)}y_3 + w_{14}^{(2)}y_4 + b_1^{(2)}\right)\right)}$$

$$z_2 = \frac{1}{1+\exp\left(-\left(w_{21}^{(2)}y_1 + w_{22}^{(2)}y_2 + w_{23}^{(2)}y_3 + w_{24}^{(2)}y_4 + b_2^{(2)}\right)\right)}$$

激活函数　　线性加权

神经网络本质上是一个复合函数

人工神经网络-手写数字图像识别



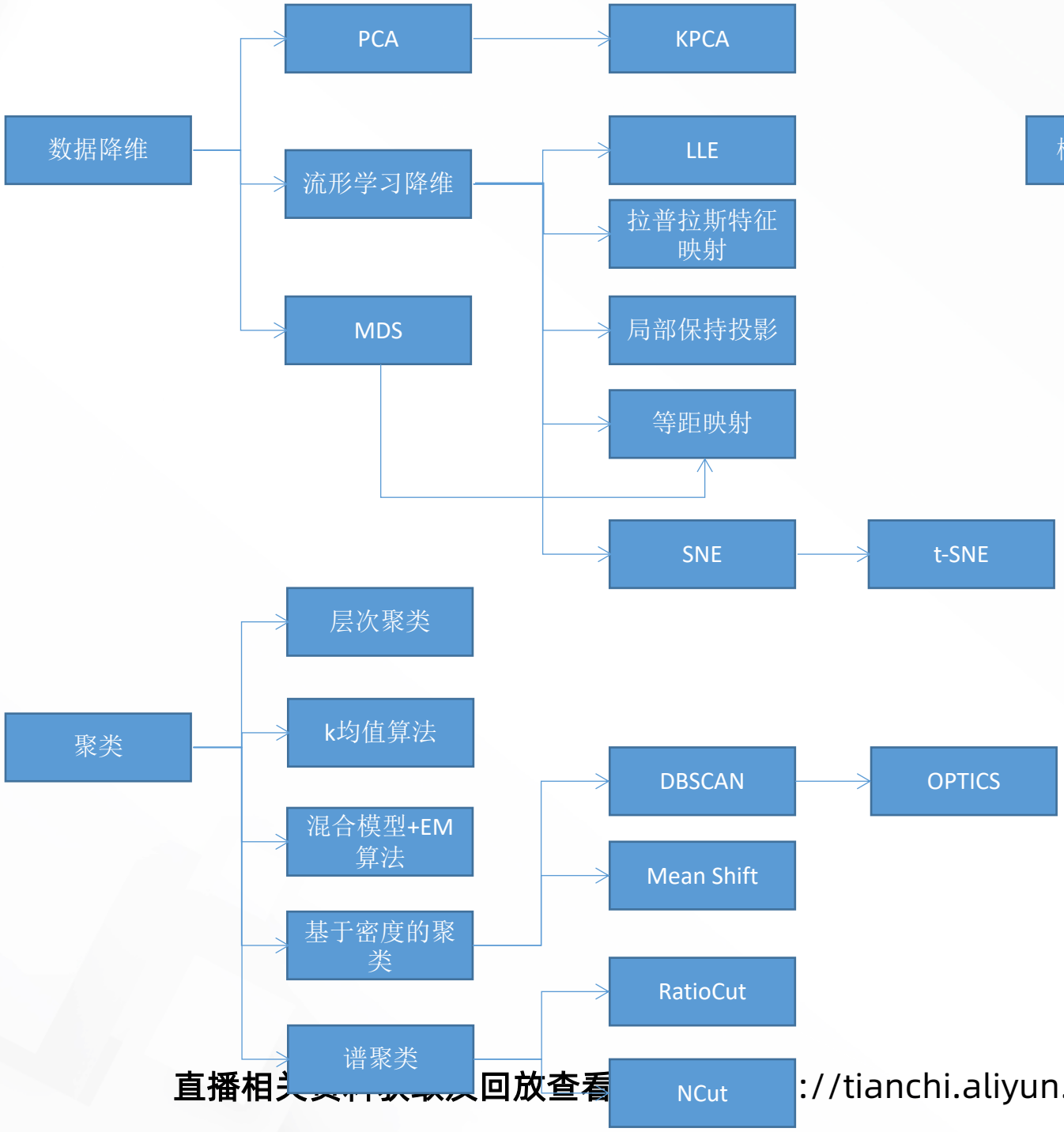
输入层有784个神经元

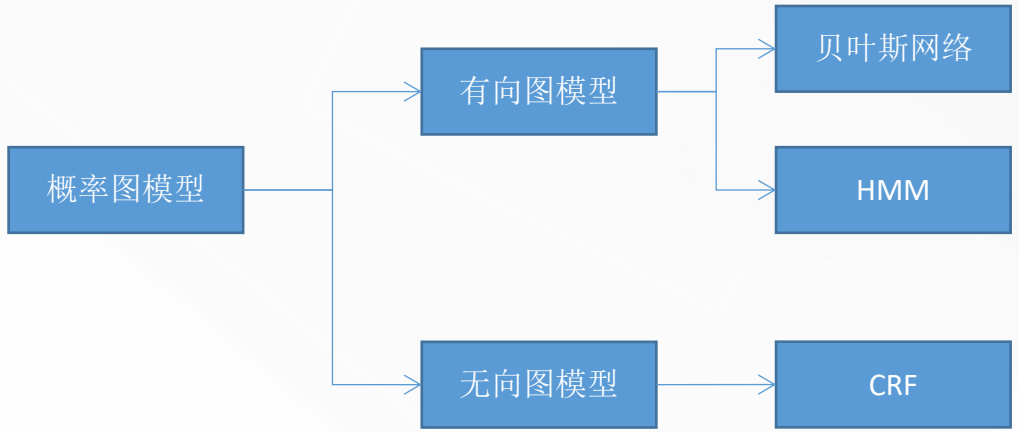
隐含层的神经元数量根据需要设定

输出层有10个神经元

TIANCHI天池

第4部分-无监督学习

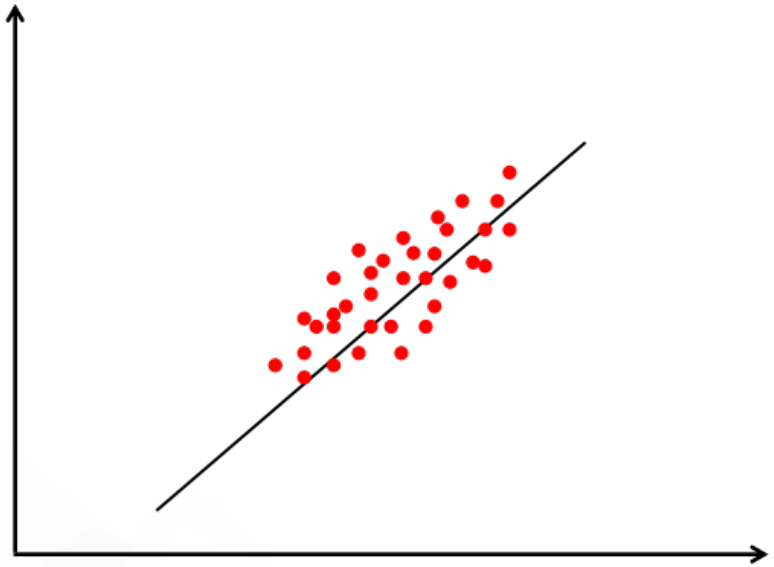




直播相关资料获取及回放查看 ://tianchi.aliyun.com/specials/promotion/activity/bookclub

主成分分析/PCA

向数据的主要变化方向投影



低维空间中的坐标

最小化重构误差

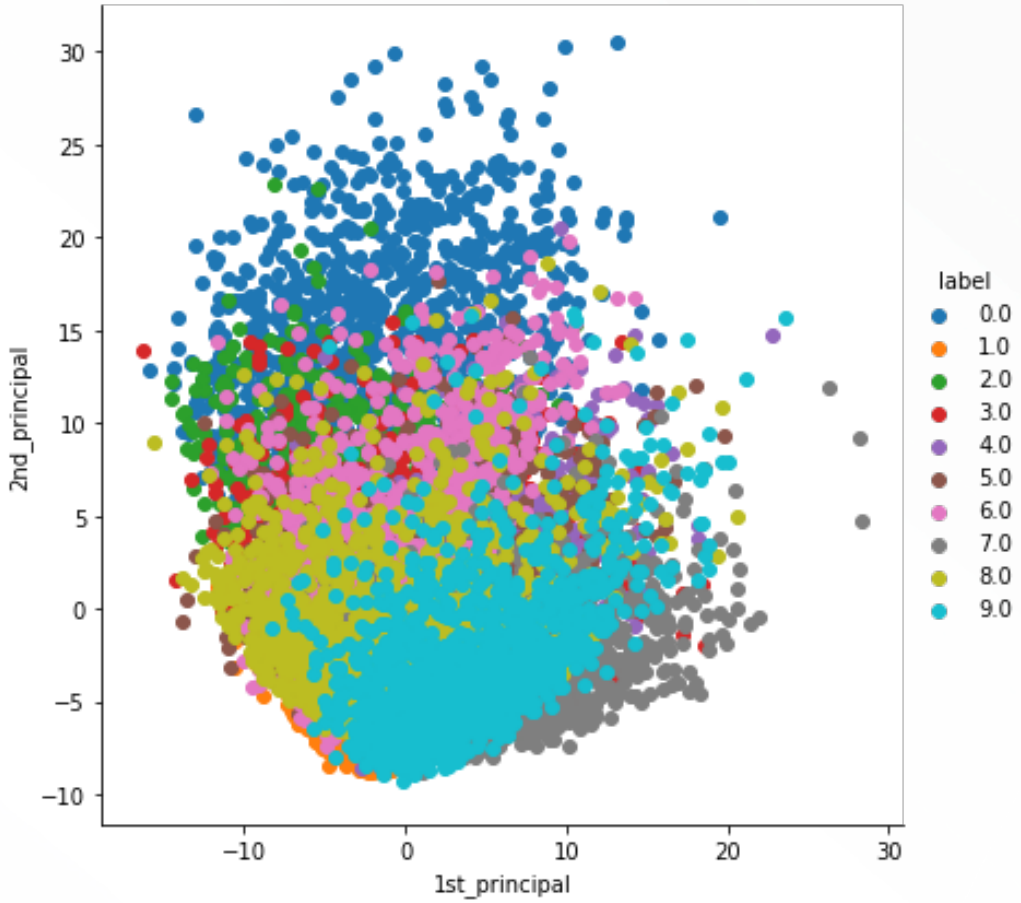
低维空间的基

$$\begin{cases} L(a,\mathbf{e}) = \sum_{i=1}^{n} \left\| \mathbf{m} + \sum_{j=1}^{d'} a_{ij}\mathbf{e}_j - \mathbf{x}_i \right\|^2 \\ \mathbf{e}_i^{\mathrm{T}}\mathbf{e}_i = 1 \\ \mathbf{e}_i^{\mathrm{T}}\mathbf{e}_j = 0, i \neq j \end{cases} \Rightarrow \mathbf{S}\mathbf{e} = \lambda\mathbf{e}$$

标准正交基

等价于求解样本协方差矩阵的
特征值和特征向量

主成分分析/PCA

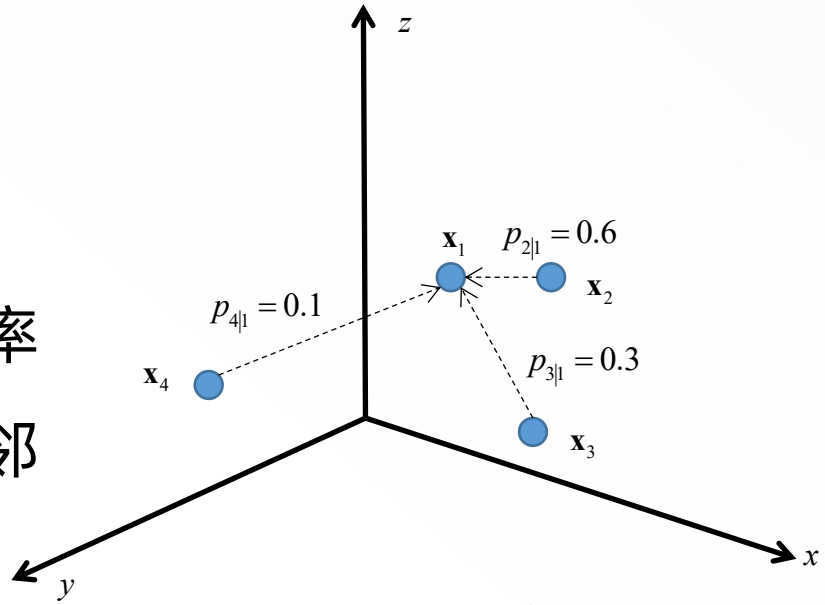


PCA对MNIST数据集的降维结果

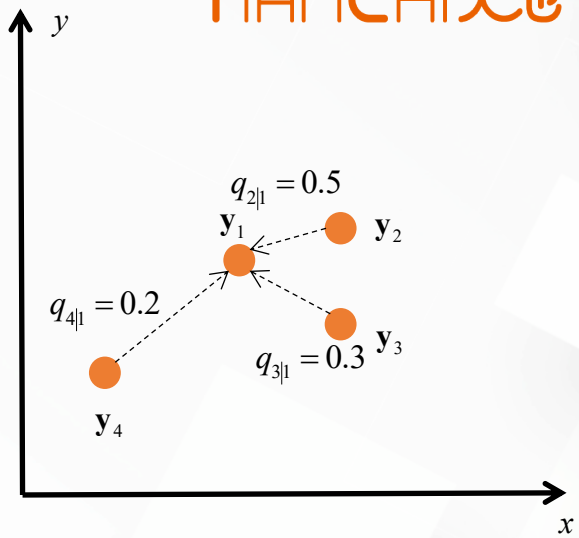
直播相关资料获取及回放查看地址：https://tianchi.aliyun.com/specials/promotion/activity/bookclub

流形学习 - t-SNE

样本之间存在邻居概率关系，相距越近，是邻居的概率越大

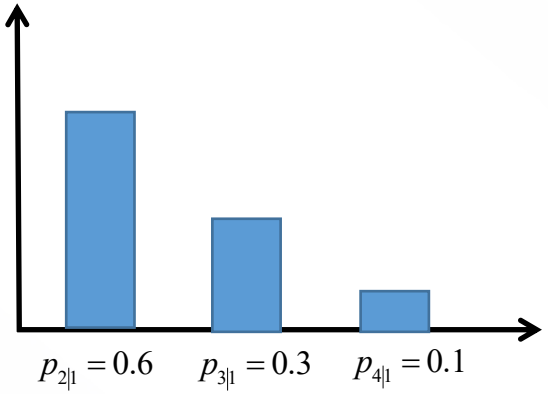


三维空间中的样本

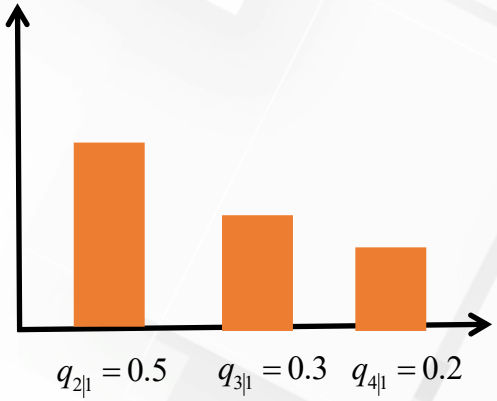


降到二维空间之后的样本

投影到低维空间中后保持邻居概率关系



三维空间中的邻居关系概率分布



降到二维空间之后邻居关系概率分布

直播相关资料获取及回放查看地址：https://tianchi.aliyun.com/specials/promotion/activity/bookclub

流形学习 - t-SNE

TIANCHI天池

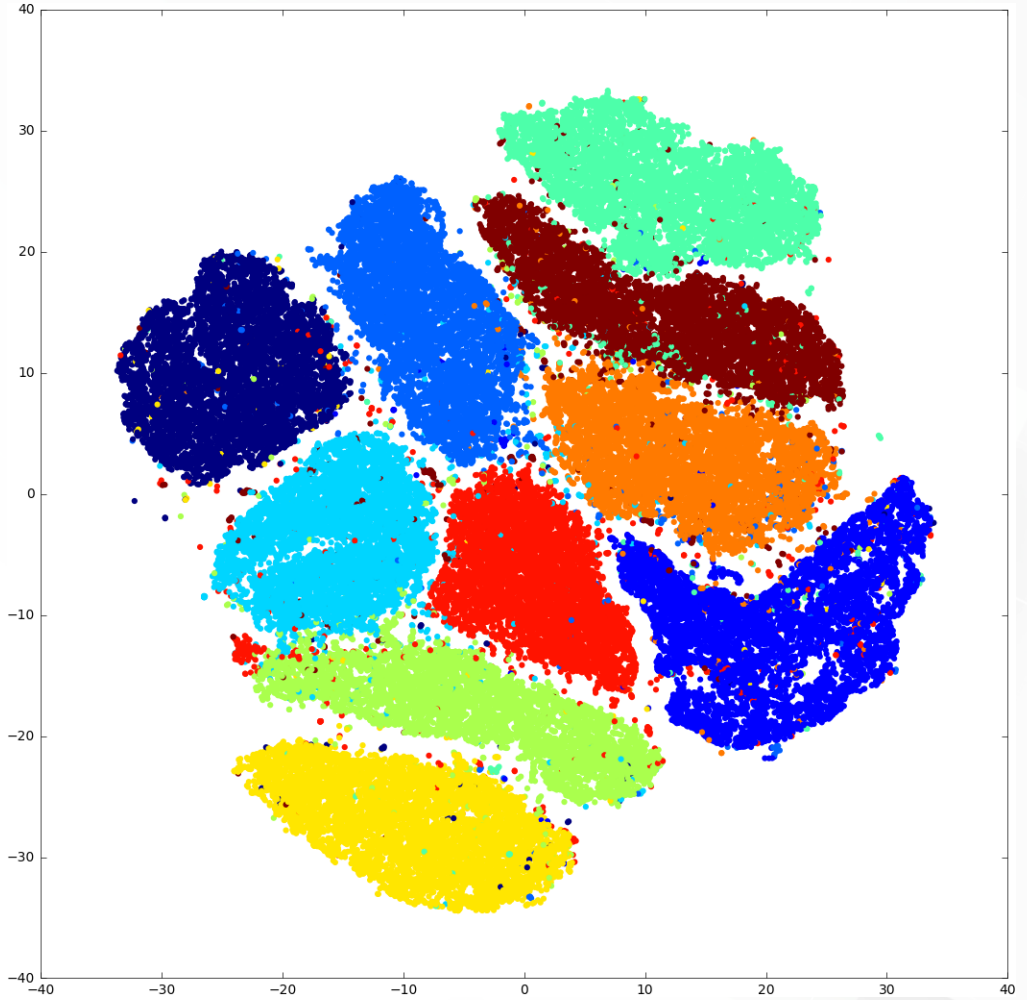
高维空间中的邻居概率 $$p_{ij}=\frac{\exp\left(-\left\|\mathbf{x}_i-\mathbf{x}_j\right\|^2/2\sigma^2\right)}{\sum_{k\neq l}\exp\left(-\left\|\mathbf{x}_k-\mathbf{x}_l\right\|^2/2\sigma^2\right)}$$

低维空间中的邻居概率 $$q_{ij}=\frac{\left(1+\left\|\mathbf{y}_i-\mathbf{y}_j\right\|^2\right)^{-1}}{\sum_{k\neq l}\left(1+\left\|\mathbf{y}_k-\mathbf{y}_l\right\|^2\right)^{-1}}$$

KL散度用于衡量两个概率分布之间的差距 $$KL\left(p\|q\right)=\sum_x p\left(x\right)\log\frac{p\left(x\right)}{q\left(x\right)}$$

目标函数 $$L\left(\mathbf{y}_i\right)=\sum_{i=1}^{l}KL\left(P_i|Q_i\right)=\sum_{i=1}^{l}\sum_{j=1}^{l}p_{j|i}\log\frac{p_{j|i}}{q_{j|i}}$$

流形学习 - t-SNE



t-SNE对MNIST数据集的降维结果

聚类 - k-均值算法

用类中心向量表示一个类

将每个样本分配到最近的那个类

然后更新每个类的类中心向量

初始化 k 个类的中心向量 $\boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_k$

循环，直到收敛

分配阶段。根据当前的类中心估计值确定每个样本所属的类：

循环，对每个样本 $\mathbf{x}_i$

计算样本离每个类中心 $\boldsymbol{\mu}_j$ 的距离：

$$d_{ij} = \left\|\mathbf{x}_i - \boldsymbol{\mu}_j\right\|$$

将样本分配到距离最近的那个类

结束循环

更新阶段。更新每个类的类中心：

循环，对每个类

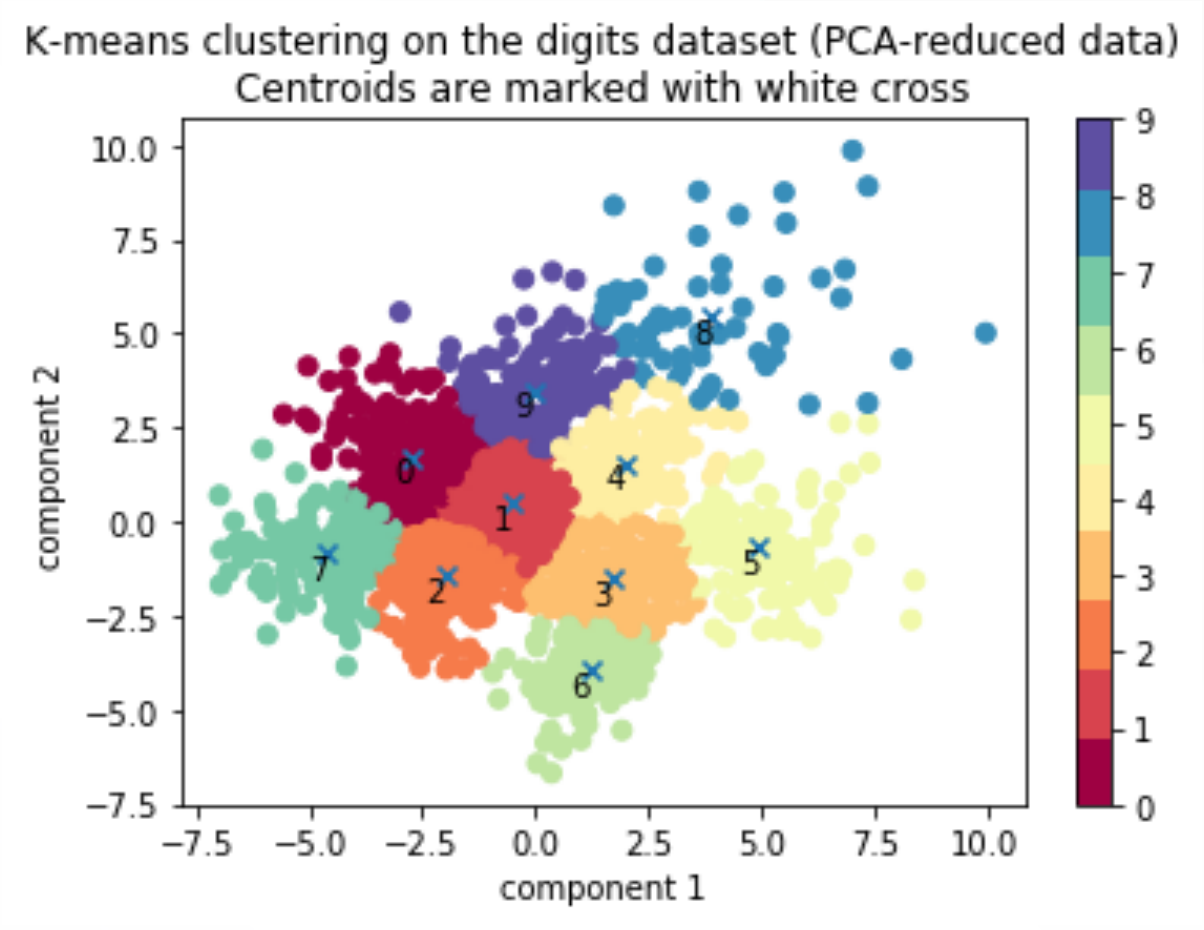
根据上一步的分配方案更新每个类的中心：

$$\boldsymbol{\mu}_i = \sum_{j=1, y_j=i}^{l} \mathbf{x}_j / N_i$$

结束循环

结束循环

聚类 - k-均值算法



k均值算法对MNIST数据集的聚类结果

直播相关资料获取及回放查看地址：https://tianchi.aliyun.com/specials/promotion/activity/bookclub

TIANCHI天池

第5部分-数据生成问题

数据生成问题

TIANCHI天池

从一个数据集学习出一个模型，这个模型可以生成随机样本，它们与训练集相似但又不完全相同

用深度生成模型生成的逼真图像

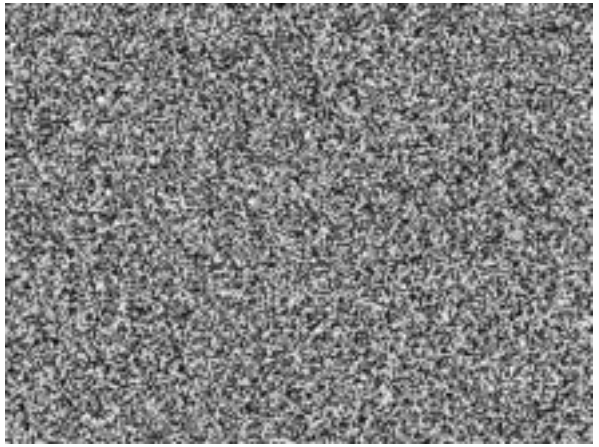
数据生成问题

已知一个样本集，其样本服从概率分布 $p_{gt}(\mathbf{x})$

从该样本集学习出一个概率分布 $p(\mathbf{x})$ ，它概率分布 $p_{gt}(\mathbf{x})$

尽可能类似

从 $p(\mathbf{x})$采样出一些样本，就是我们想要的数据

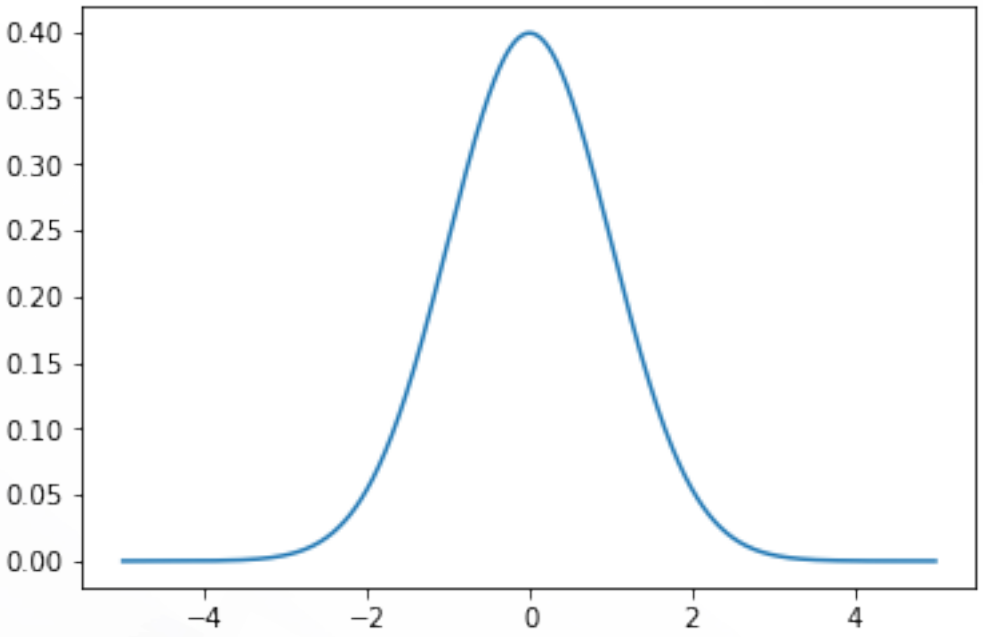


如何评价生成的样本，以指导算法进行生成

数据生成问题

TIANCHI天池

基于均匀分布/正态分布，可以变换出任意的概率分布



标准正态分布

$$\begin{bmatrix} z_1 \\ z_2 \\ \dots \\ z_n \end{bmatrix}$$

采样出随机向量值 $\mathbf{z}$

分布变换 $g(\mathbf{z})$

$\mathbf{x}$ 服从真实图像的概率分布

概率分布变换

深度生成模型

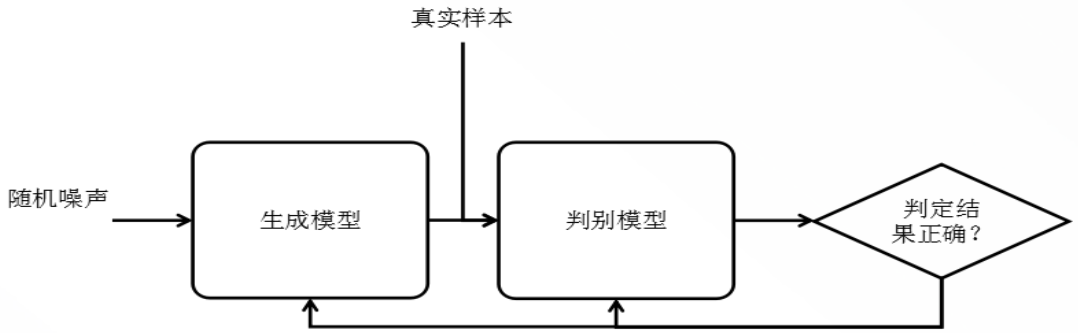
- 生成对抗网络（GAN）
- 变分自动编码器 （VAE）

生成对抗网络/GAN

TIANCHI天池

生成器：G，根据随机噪声生成样本数据，实现概率分布变换

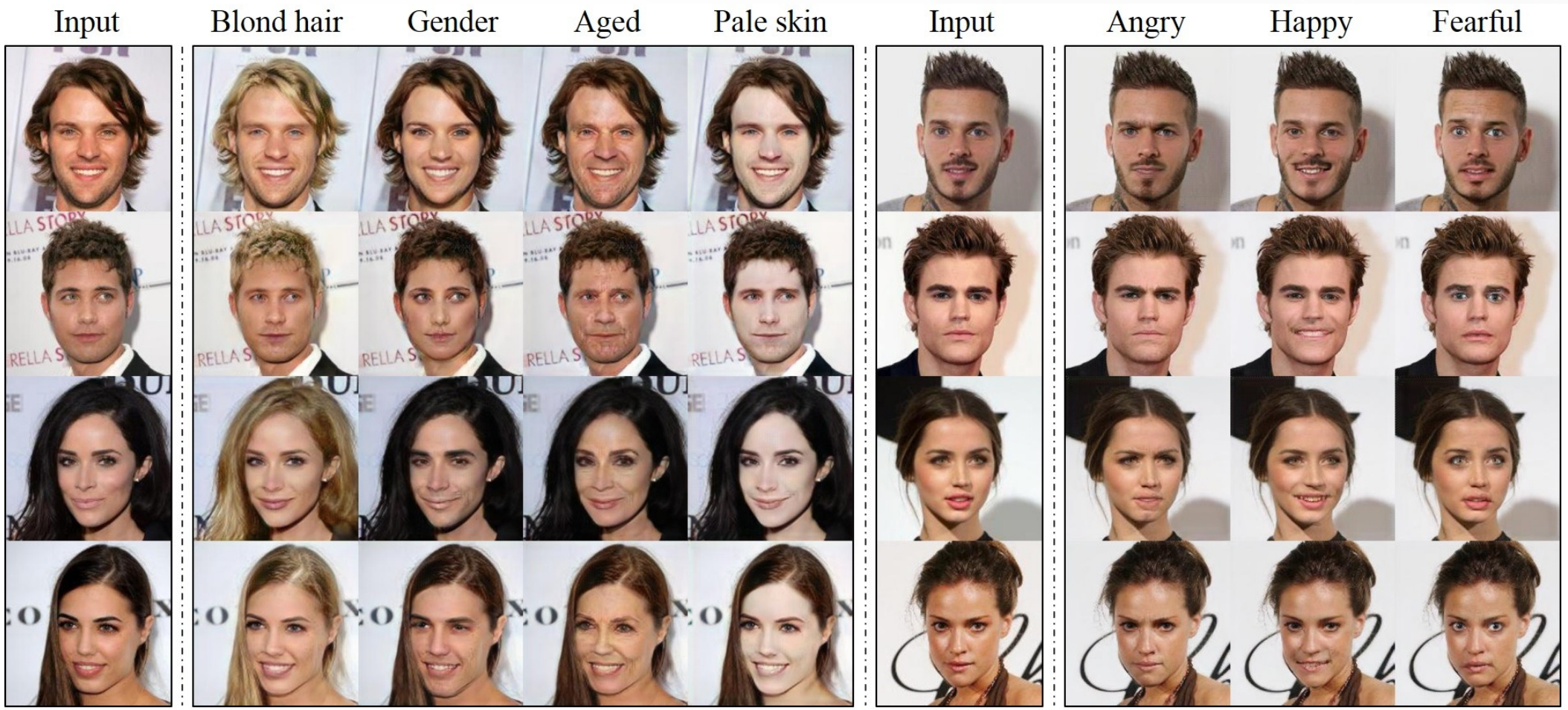
判别器：D，判定样本是真实的还是由生成器生成的



$$\min_G \max_D V(D,G) = \mathrm{E}_{\mathbf{x}\sim p_{data}(\mathbf{x})}\left[\ln D(\mathbf{x})\right] + \mathrm{E}_{\mathbf{z}\sim p_{\mathbf{z}}(\mathbf{z})}\left[\ln\left(1-D(G(\mathbf{z}))\right)\right]$$

真实的样本　　生成的样本

生成对抗网络-GAN



GAN生成的人脸图像

直播相关资料获取及回放查看地址： https://tianchi.aliyun.com/specials/promotion/activity/bookclub

TIANCHI天池

Q&A

天池读书会

TIANCHI天池

清華大學出版社
TSINGHUA UNIVERSITY PRESS

《机器学习的原理：算法与应用》

本书全面系统地讲述了深度学习、机器学习的主要算法。

直播嘉宾：SIGAI CEO 雷明

直播时间：2月24日20:00～21:00

扫码领取读书会相关
学习资源