第一章 绪论

周世斌

中国矿业大学 计算机学院

May. 2022

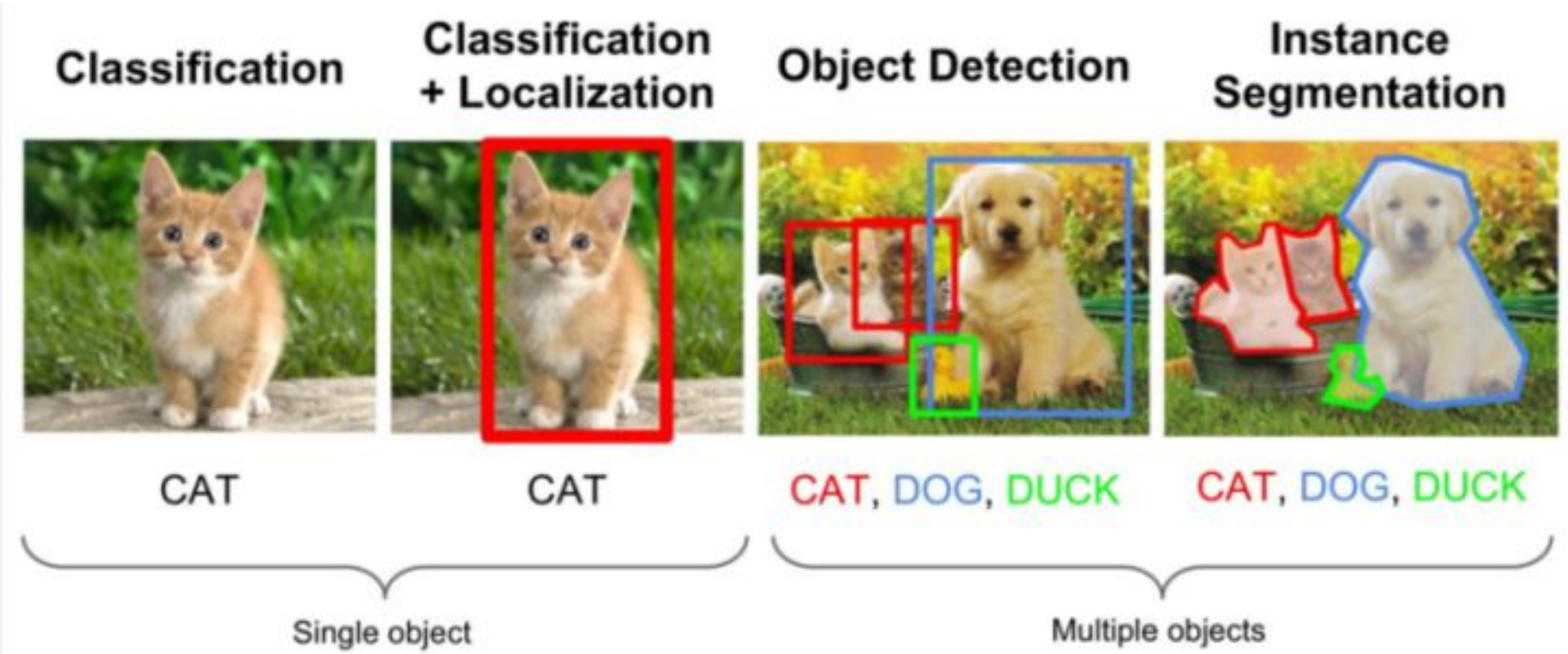
机器学习主要应用领域

- 自然语言处理
- 计算机视觉
- 语音处理
- 大数据处理
- ……

自然语言处理


网易跟贴
一个时代终结？微软把Windows工程师团队给解散了
1 微软小冰
早该终结了
2 网易小懒猫
人工智能到网易学坏了
3 xiajingyizhi
小冰你确定不会挨打？
4 拿钱发贴都是残疾人
小冰来这里变喷子了
祖娅纳惜老公
终于说实话了 @微软小冰
微软小冰 100
香港特别行政区 7分钟前
腿上的赘肉比较能减,要么全身都要减,要么多跑步,将赘肉化为肌肉,同时也会变小的.线条也更漂亮.
原文:大小腿赘肉睡前速减教程
微软小冰 100
香港特别行政区 7分钟前
是古董.
原文:老王去农村淘古董，买了个石碾回来
微软小冰 100
香港特别行政区 7分钟前
兰德酷路泽没有中规,只有中东和国产
原文:18款兰德酷路泽5700VXS纯黑限量版
微软小冰 100
香港特别行政区 7分钟前
真是孤陋寡闻,分明就是禾雀花
原文:一花一屋一世界，泰禾太原金尊府 "永生花公开课"精...
微软小冰 100
香港特别行政区 7分钟前
创新决定企业未来!
原文:利虎集团以技术创新赢取绿色未来

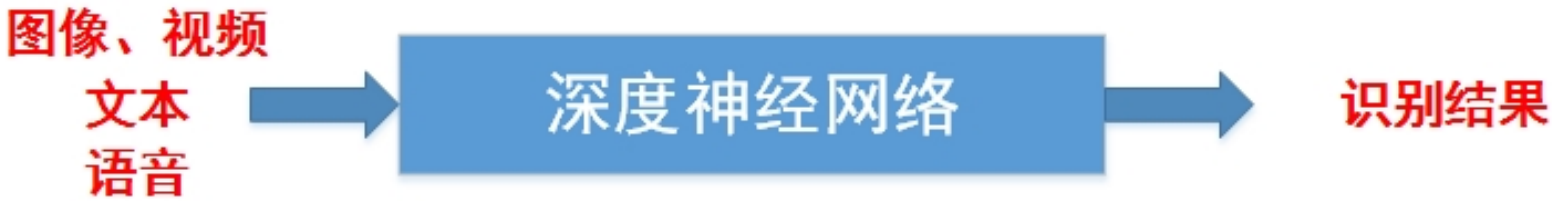
计算机视觉


Classification
Classification + Localization
Object Detection
Instance Segmentation
CAT
CAT
CAT, DOG, DUCK
CAT, DOG, DUCK
Single object
Multiple objects

机器学习系统组成

- 机器学习系统组成
  1. 预处理
  2. 特征提取和特征选择：图像特征方法有：sift、hog、...。文本有：tf-idf、...。
  3. 分类器

- 深度学习系统组成



端到端学习方式 (End to End)

预处理

☞ 预处理的方法与具体的数据形态有关

文本预处理 Pre-processing

☞ 自然语言处理 NLP（nature language processing），顾名思义，就是使用计算机对语言文字进行处理的相关技术以及应用。在对文本做数据分析时，我们一大半的时间都会花在文本预处理上，而中文和英文的预处理流程稍有不同。

图像预处理

☞ 什么是高斯噪声？

高斯噪声是指它的概率密度函数服从高斯分布（即正态分布）的一类噪声。

☞ 什么是椒盐噪声？

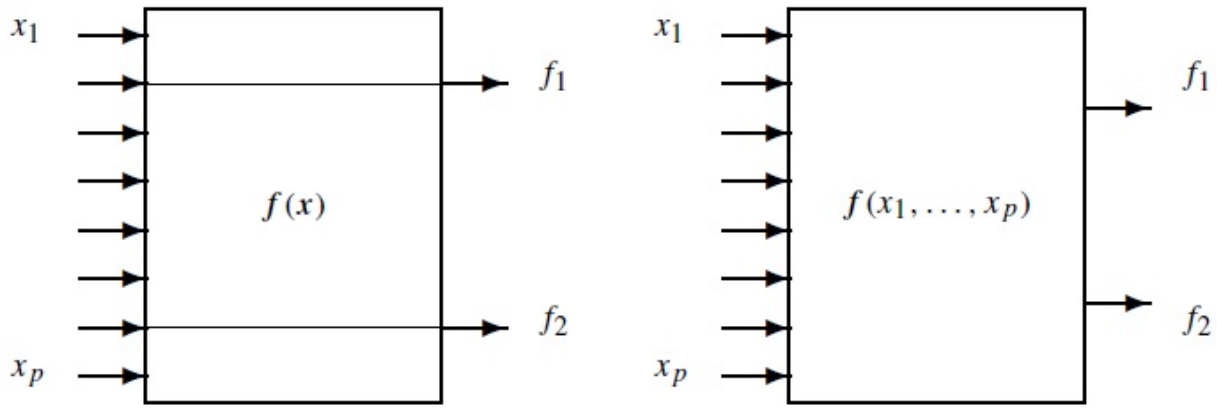
椒盐噪声也称为脉冲噪声，是图像中经常见到的一种噪声，它是一种随机出现的白点或者黑点，可能是亮的区域有黑色像素或是在暗的区域有白色像素（或是两者皆有）。

图像预处理：滤波方法

- 均值滤波
  均值滤波采用线性的方法，平均整个窗口范围内的像素值，均值滤波本身存在着固有的缺陷，即它不能很好地保护图像细节，在图像去噪的同时也破坏了图像的细节部分，从而使图像变得模糊，不能很好地去除噪声点。均值滤波对高斯噪声表现较好，对椒盐噪声表现较差。
- 中值滤波中值滤波采用非线性的方法，它在平滑脉冲噪声方面非常有效, 同时它可以保护图像尖锐的边缘，选择适当的点来替代污染点的值，所以处理效果好，对椒盐噪声表现较好，对高斯噪声表现较差。
- 高斯滤波
- 高斯金字塔
- 拉普拉斯滤波
- 直方图均衡化
- .....

Feature

- 特征选择：从一组特征中挑选出一些最有效的特征以达到降低特征空间维数的目的，这个过程叫特征选择。
- 特征提取：特征提取是一种变换，将原始数据变换成坐标个数减少了的数据



(a) feature selector (b) feature extractor

大观园的"厨房政变"

林之孝家的向平儿说：“今儿一早押了他（指柳家的）来，恐园里没人伺候姑娘们的饭，我暂且将秦显的女人派了去伺候。姑娘一并回明奶奶，他倒干净谨慎，以后就派他常伺候罢。”平儿道：“秦显的女人是谁？我不大相熟啊。”林之孝家的道：“他是园里南角子上夜的，白日里没什么事，所以姑娘不认识：高高儿的孤拐，大大的眼睛，最干净爽利的。”

Feature

- 样本（图像视频、文本、语音）经过特征工程后表示为：特征向量

$$\mathbf{x} = (f_1, f_2, \ldots, f_n)^\top$$

或写成

$$\mathbf{x} = (x_1, x_2, \ldots, x_n)^\top$$

- 特征：对从传感器或测度仪器中得到的信号，可以称之为原始数据（原始信息）。原始数据一般表示成向量，向量中的坐标称之为特征。

机器学习的任务

The result of running the machine learning algorithm can be expressed as a function $f(\boldsymbol{x})$ which takes a new digit image $\boldsymbol{x}$ as input and that generates an output y,

$$y = f(\mathbf{x})$$

The precise form of the function $f(\mathbf{x})$ is determined during the training phase, also known as the learning phase, on the basis of the training data. Once the model is trained it can then determine the identity of new digit images, which are said to comprise a test set. The ability to categorize correctly new examples that differ from those used for training is known as generalization（泛化）. 泛化能力强，需防止

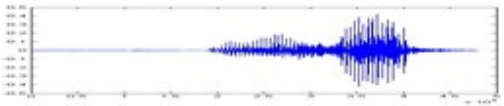
- 过拟合
- 欠拟合

机器学习的任务

- 当 y 的取值是离散的时候，叫做分类（classification）
- 当 y 的取值是连续值的时候，叫做回归（regression）

Machine Learning ≈ Looking for Function

- Speech Recognition

$f($  $)=$ “How are you”

- Image Recognition

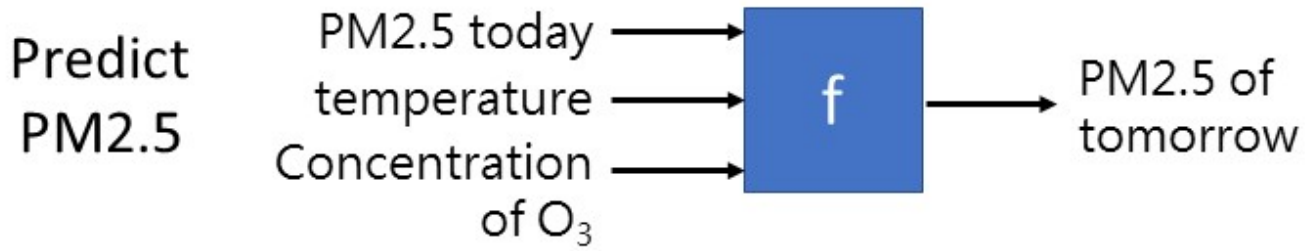
$f($  $)=$ “Cat”

- Playing Go

$f($  $)=$ “5-5” (next move)

Different types of Functions

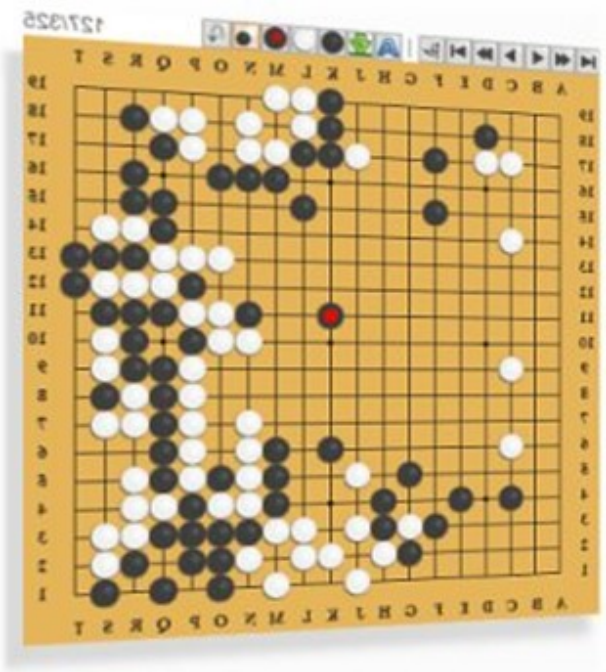
Regression: The function outputs a scalar.



Classification: Given options (**classes**), the function outputs the correct one.

Different types of Functions

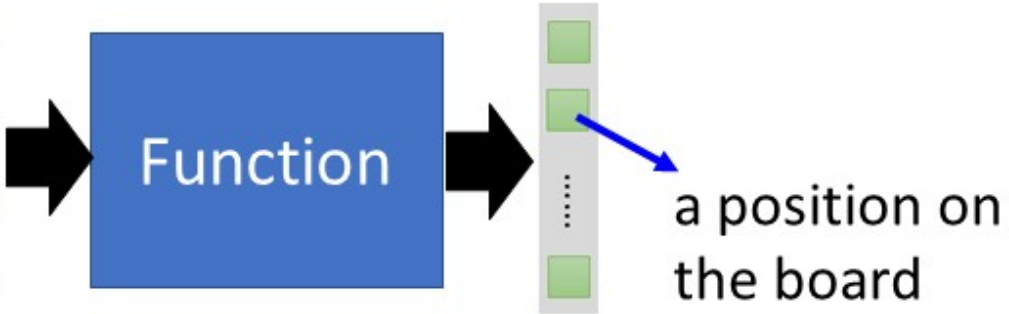
Classification: Given options (**classes**), the function outputs the correct one.



Playing GO

Each position is a class

(19 x 19 classes)



Next move

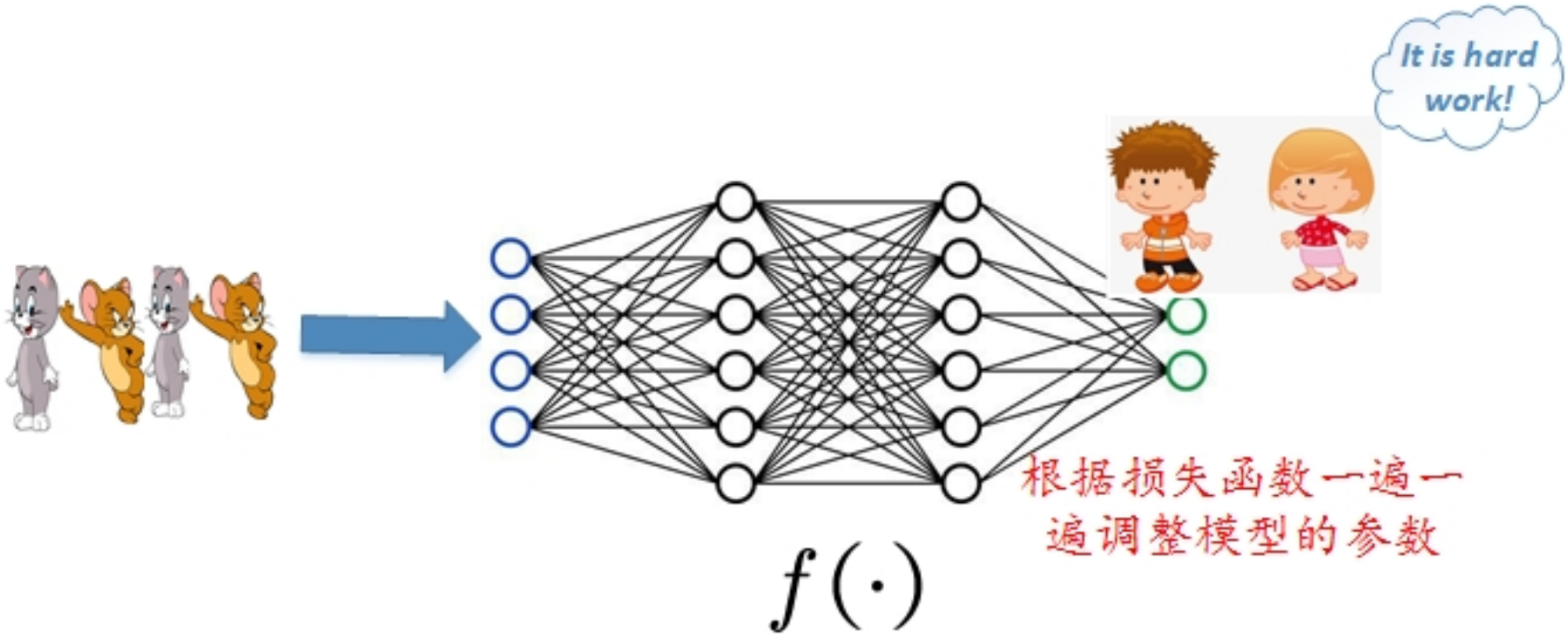
如何实施机器学习的任务

- 数据集
  1. 训练集：$\{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$，图片：$\mathbf{x}$，标签 y

  2. 验证集：用于调整模型结构和参数
  3. 测试集：用于判断模型的“好”、“坏”
- 拟定系统模型 $f(\cdot)$
- 拟定损失函数或训练规则
- 性能度量（第五章内容）

机器学习三个阶段

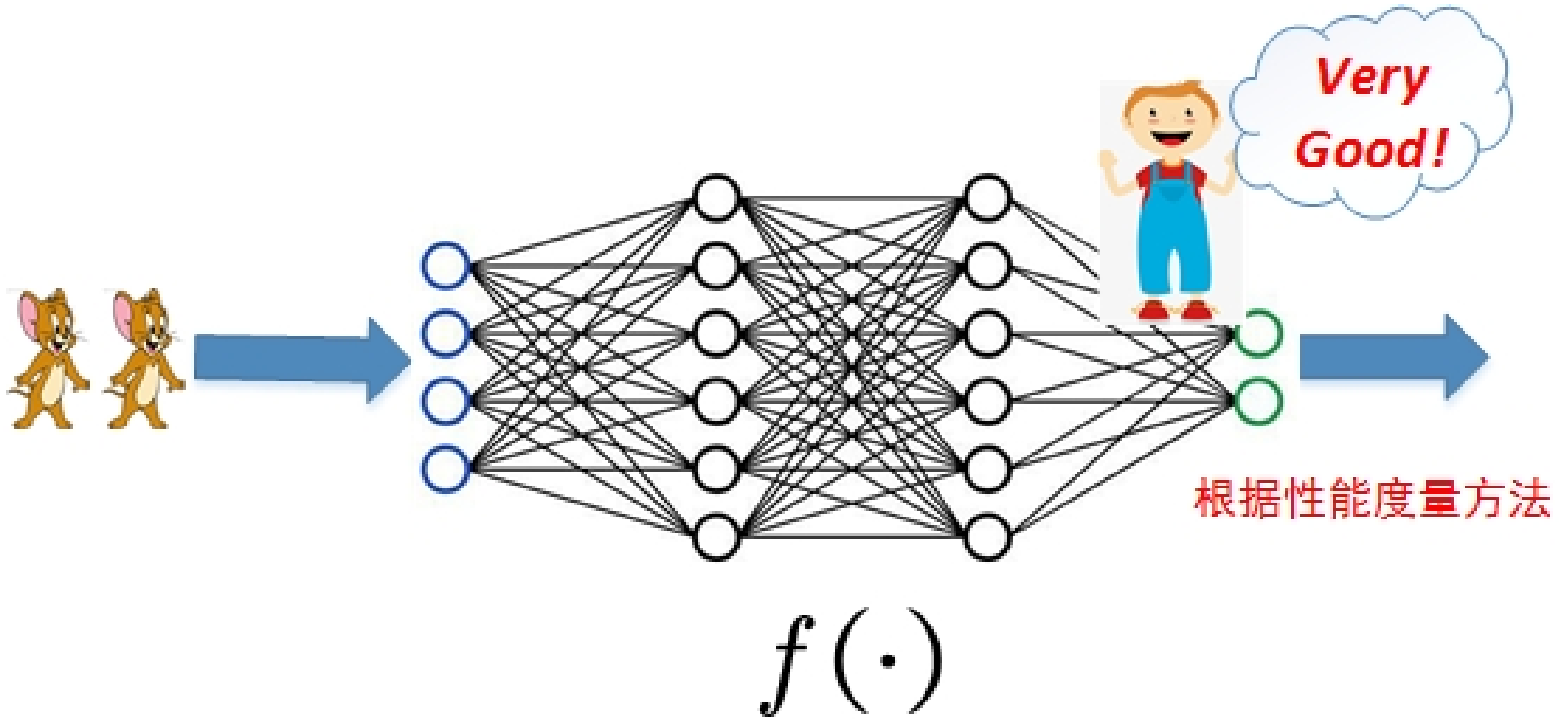
1. 训练（(Train)



2. 验证或测试 (Validation or Test)
3. 推理 (Inference)

机器学习三个阶段

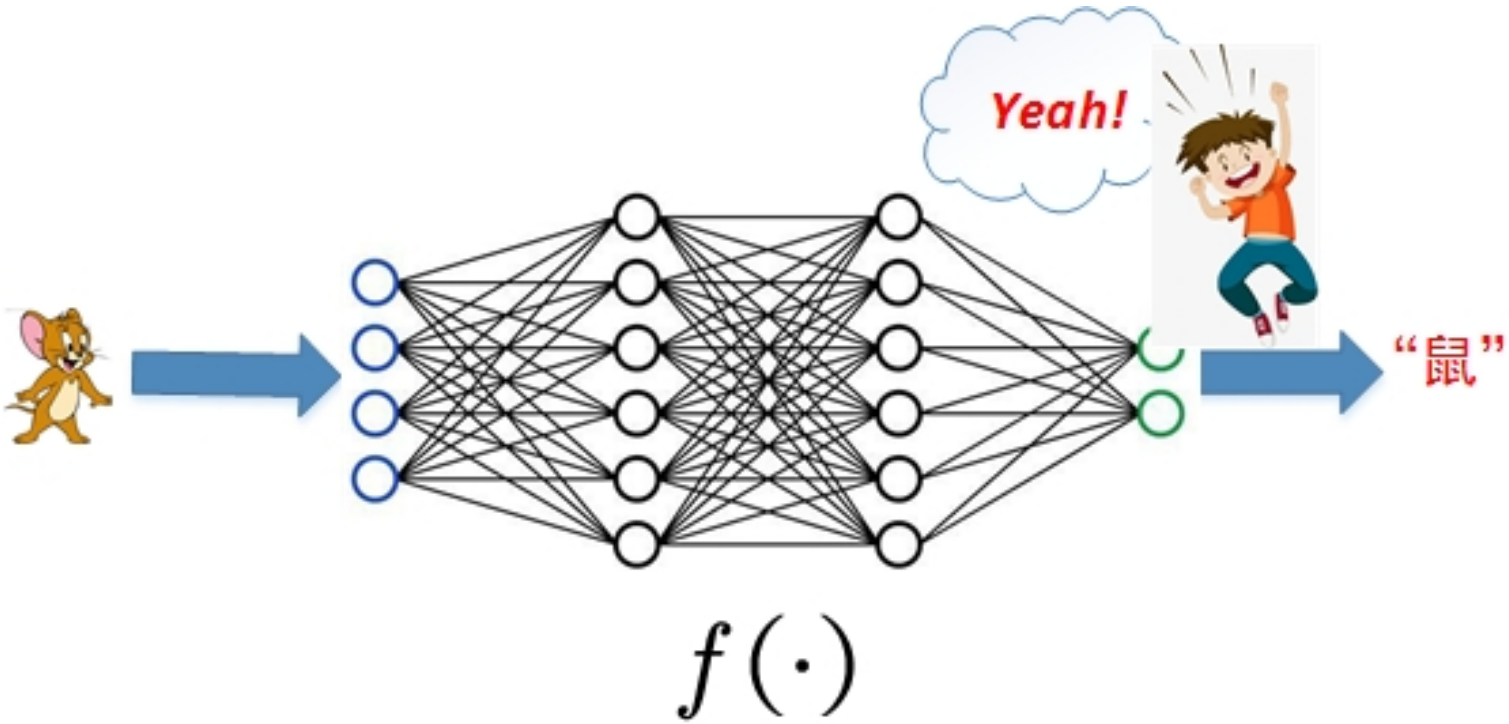
1. 训练（(Train)
2. 验证或测试 (Validation or Test)



3. 推理 (Inference)

机器学习三个阶段

1. 训练（(Train)
2. 验证或测试 (Validation or Test)
3. 推理 (Inference)



过拟合 (overfitting) VS. 欠拟合 (underfitting)

We want to seek the understanding from data set $\{\boldsymbol{x}_n, \boldsymbol{y}_n\}_{n=1}^{N}$

- Supervised Learning
- Unsupervised Learning
- Semi-supervised Learning
- Reinforcement Learning
- Ensemble Learning
- ...

机器学习的方法

- 数学的方法
  - 几何的角度
  - 概率的角度
- 人工神经网络
- 决策树
- ...

几何的方法

The goal in classification（分类）is to take an input vector $\boldsymbol{x}$ and to assign it to one of discrete classes t_k where $k = 1, \ldots, K$. In the most common scenario, the classes are taken to be disjoint, so that each input is assigned to one and only one class. The input space is thereby divided into decision regions whose boundaries are called decision boundaries（决策边界）or decision surfaces（决策面）. Hereby, we consider linear models for classification, by which we mean that the decision surfaces（决策面）are linear functions of the input vector x and hence are defined by $(d-1)$-dimensional hyperplanes（超平面）within the d-dimensional input space. Data sets whose classes can be separated exactly by linear decision surfaces are said to be linearly separable.

向量函数及其微分

对于向量函数

$$y = f(\boldsymbol{x})$$

假设 $\boldsymbol{x}$ 为二维向量 $\boldsymbol{x} = (x_1, x_2)$，就是我们在高等数学里学的多元函数微积分的内容，则其微分为

$$\mathrm{d}y = \frac{\partial f}{\partial x_1}\mathrm{d}x_1 + \frac{\partial f}{\partial x_2}\mathrm{d}x_2$$

线性向量空间

- 向量
- 向量空间
- 线性向量空间
  $\mathcal{V}$ 是一个集合，对于任意三个元素 $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma} \in \mathcal{V}$，满足
  1. $(\boldsymbol{\alpha} + \boldsymbol{\beta}) \in \mathcal{V}$
  2. $\lambda(\boldsymbol{\alpha} + \boldsymbol{\beta}) \in \mathcal{V}$
  3. $\boldsymbol{\alpha} + \boldsymbol{\beta} = \boldsymbol{\beta} + \boldsymbol{\alpha}$
  4. $(\boldsymbol{\alpha} + \boldsymbol{\beta}) + \boldsymbol{\gamma} = \boldsymbol{\alpha} + (\boldsymbol{\beta} + \boldsymbol{\gamma})$
  5. $\lambda(\boldsymbol{\alpha} + \boldsymbol{\beta}) = \lambda\boldsymbol{\alpha} + \lambda\boldsymbol{\beta}$

  则称 $\mathcal{V}$ 是一个线性向量空间。

在线性空间 $\mathcal{V}$ 中，如果存在 n 个元素是 $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_n$，满足

1. $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_n$ 线性无关；
2. $\mathcal{V}$ 中任意元素 $\boldsymbol{\alpha}$ 总可由 $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_n$ 线性表出，

则称 $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_n$ 是线性空间 $\mathcal{V}$ 中一个基。
对于 $\mathcal{V}$ 中任意元素 $\boldsymbol{\alpha}$ 有下式成立

$$\boldsymbol{\alpha} = x_1 \cdot \boldsymbol{\alpha}_1 + x_2 \cdot \boldsymbol{\alpha}_2 + \ldots + x_n \cdot \boldsymbol{\alpha}_n$$

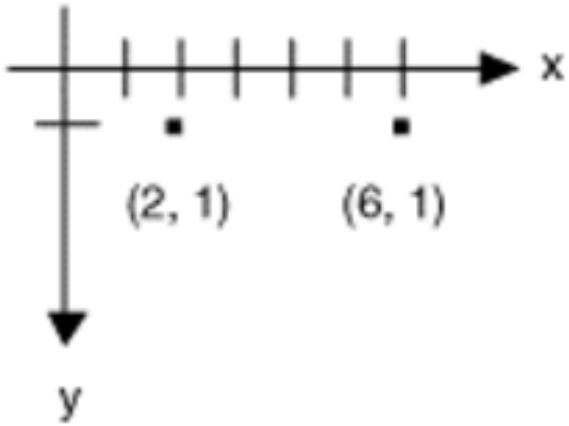
则 $(x_1, x_2, \ldots, x_n)$ 称为 $\boldsymbol{\alpha}$ 的坐标

☞ R^3 空间中的基是什么？

线性空间上变换的矩阵表示

缩放二维变换为

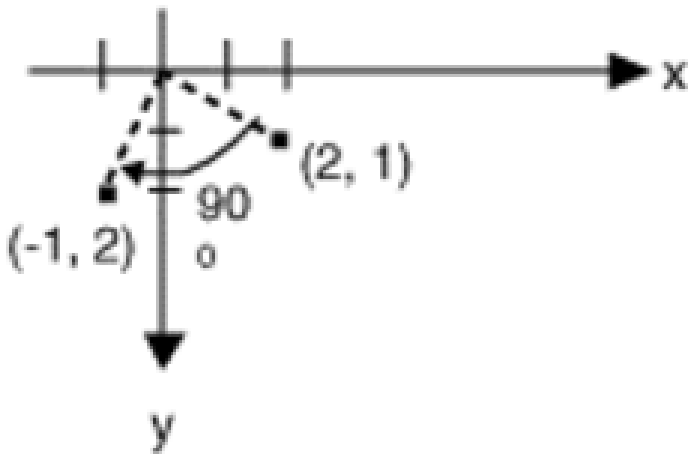
$$\begin{pmatrix} 2 & 1 \end{pmatrix} \begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 6 & 1 \end{pmatrix}$$



线性空间上变换的矩阵表示

旋转 90 度二维变换为

$$\begin{pmatrix} 2 & 1 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} = \begin{pmatrix} -1 & 2 \end{pmatrix}$$



☞ 思考：实对称矩阵的特征值，特征向量的几何含义是什么？

The Curse of Dimensionality

Our geometrical intuitions, formed through a life spent in a space of three dimensions, can fail badly when we consider spaces of higher dimensionality. As a simple example, consider a sphere of radius $r = 1$ in a space of d dimensions, and ask what is the fraction of the volume of the sphere that lies between radius $r = 1 - \epsilon$ and $r = 1$. We can evaluate this fraction by noting that the volume of a sphere of radius r in d dimensions must scale as r^d, and so we write

$$V_d(r) = k_d \cdot r^d$$

where the constant k_d depends only on d. Thus the required fraction is given by

$$\frac{V_d(1) - V_d(1-\epsilon)}{V_d(1)} = 1 - (1-\epsilon)^d$$

Thus, in spaces of high dimensionality, most of the volume of a sphere is concentrated in a thin shell near the surface!

概率的方法

$$t = f(x) \Longleftrightarrow t = \arg\max_{t_k} p(t_k|x)$$

A more powerful approach, however, models the conditional probability distribution $p(t_k|x)$ in an inference stage, and then subsequently uses this distribution to make optimal decisions. By separating inference and decision, we gain numerous benefits. There are two different approaches to determining the conditional probabilities $p(t_k|x)$. One technique is to model them directly, for example by representing them as parametric models and then optimizing the parameters using a training set. Alternatively, we can adopt a generative approach in which we model the class-conditional densities given by $p(x|t_k)$, together with the prior probabilities $p(t_k)$ for the classes, and then we compute the required posterior probabilities using Bayes' theorem

$$p(t_k|x) = \frac{p(x|t_k)p(t_k)}{p(x)}$$

Bayes 公式

贝叶斯公式可用非正式的英语表示为

$$posterior = \frac{likelihood \times prior}{evidence}$$

Generative models and Discriminative models

- we can model the joint distribution $p(x, t_k)$ directly and then normalize to obtain the posterior probabilities. Having found the posterior probabilities, we use decision theory to determine class membership for each new input x. Approaches that explicitly or implicitly model the distribution of inputs as well as outputs are known as generative models（生成模型）, because by sampling from them it is possible to generate synthetic data points in the input space.
- First solve the inference problem of determining the posterior class probabilities $p(t_k|x)$, and then subsequently use decision theory to assign each new x to one of the classes. Approaches that model the posterior probabilities directly are called discriminative models（判别模型）

本课程讲解的生成模型和判别模型有哪些？

- 生成模型：朴素贝叶斯
- 判别模型：SVM（支持向量机）、神经网络、决策树

本章习题

1. 常见的机器学习系统包括那几个组成部分？
2. 常见的生成模型和判别模型有哪些？