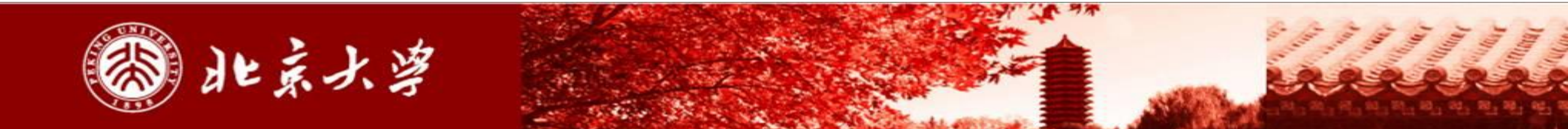
《物理与人工智能》

5. 机器学习基础和线性回归

授课教师：马滟青

2025/09/29（第四周）

鸣谢：基于计算机学院《人工智能引论》课程组幻灯片



什么是学习(Learning)?

- **“Learning is any process by which a system improves performance from experience.” - Herbert Simon**

Herbert A. Simon

American economist

Herbert Alexander Simon was an American economist, political scientist and cognitive psychologist, whose primary research interest was decision-making within organizations and is best known for the theories of "bounded rationality" and "satisficing". Wikipedia

1975年图灵奖
1978年诺贝尔经济学奖

机器学习(Machine Learning)

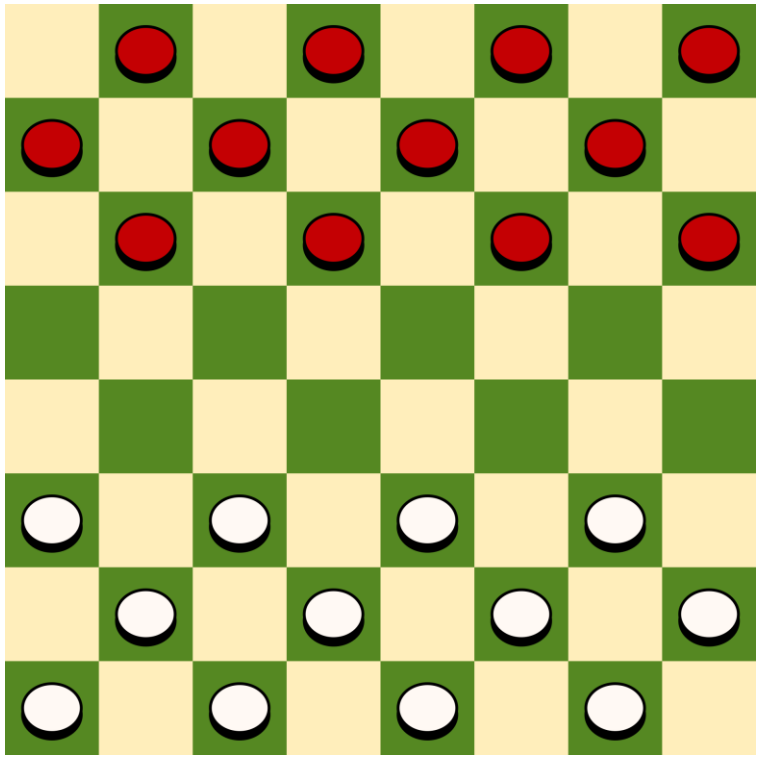
北京大学 PEKING UNIVERSITY

- **“A science of getting computers to learn without being explicitly programmed.” - Arthur Samuel**

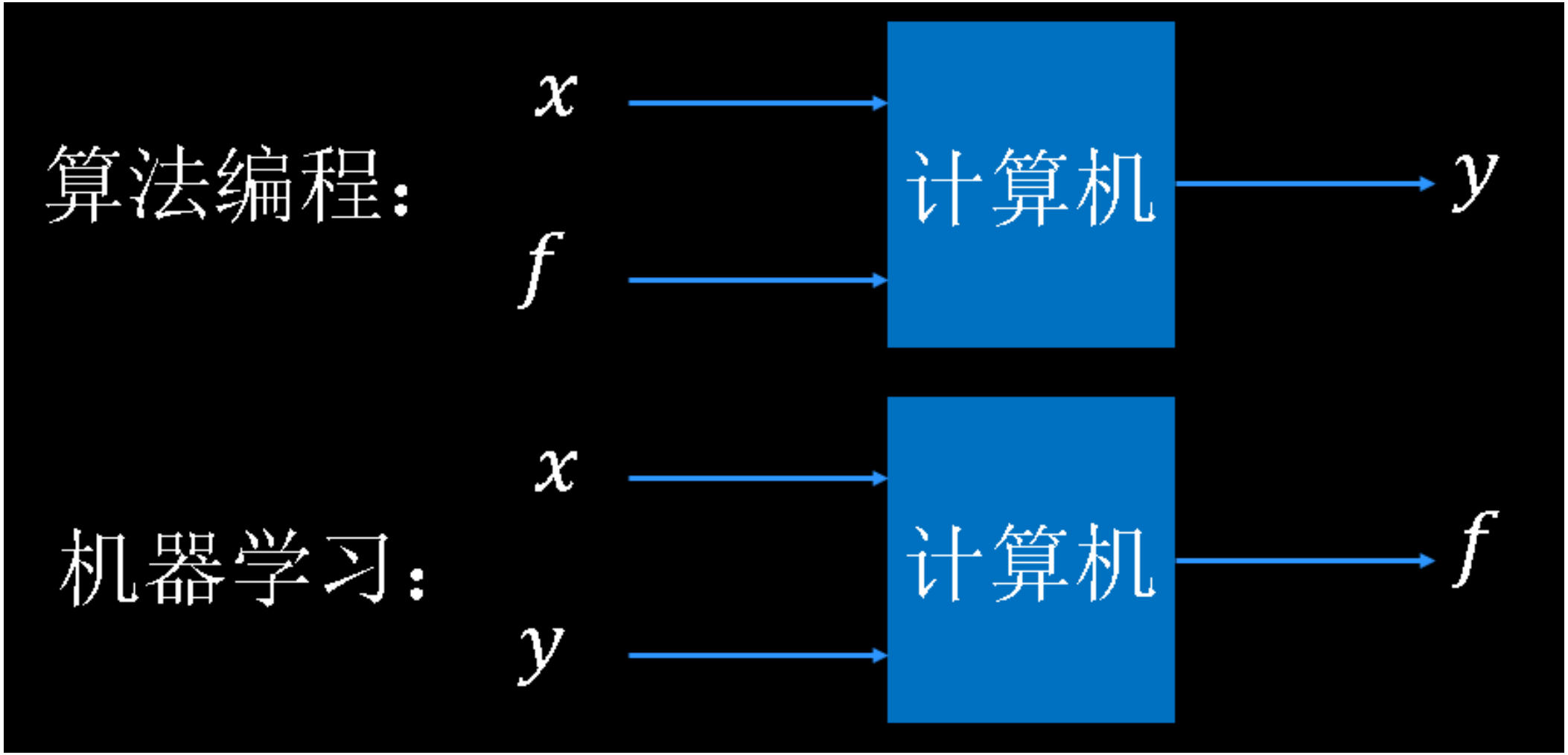
Arthur Samuel

Computer scientist

Arthur Lee Samuel was an American pioneer in the field of computer gaming and artificial intelligence. He popularized the term "machine learning" in 1959.
Wikipedia



算法编程 vs 机器学习

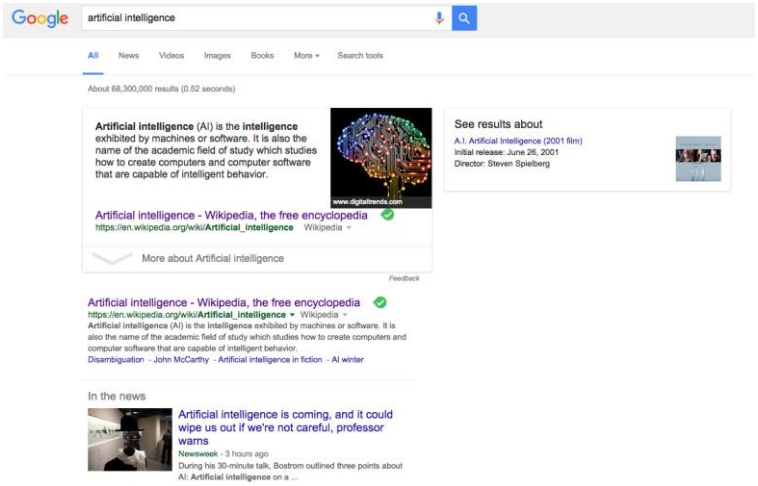


Slide credit: Pedro Domingos

机器学习日常应用

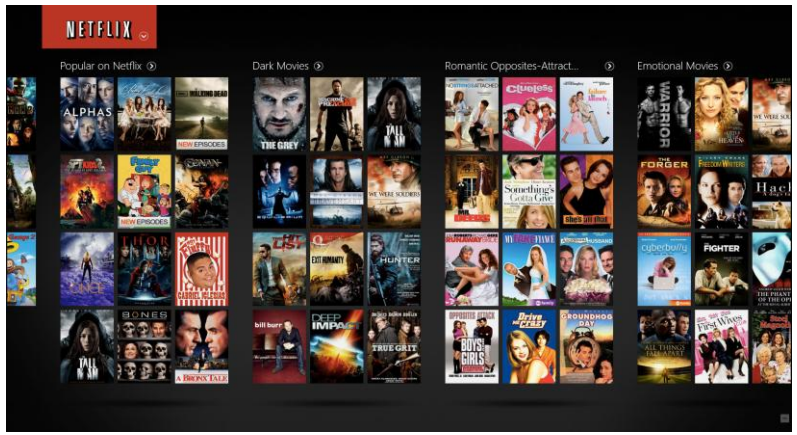
北京大学 PEKING UNIVERSITY

搜索引擎

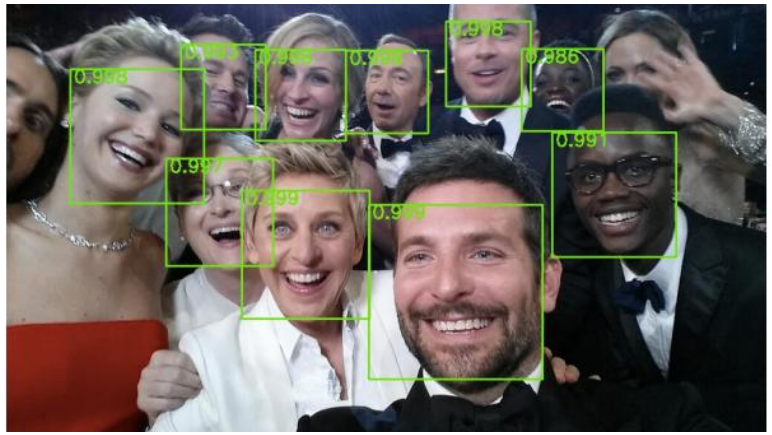


垃圾邮件分类器

推荐系统

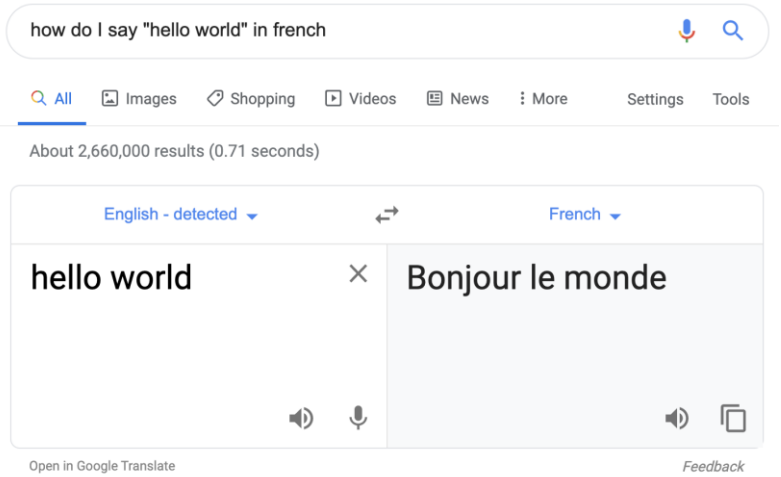


人脸识别



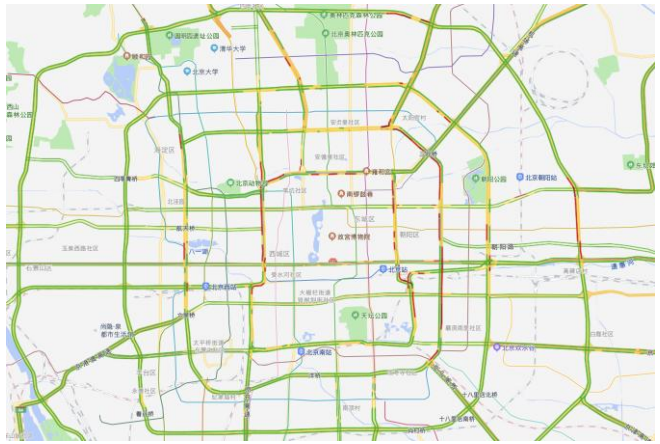
机器学习日常应用

机器翻译



数字助理

交通预测



语音识别

机器学习前沿应用

北京大学 PEKING UNIVERSITY

内容生成

Let us watch a video!

机器学习前沿应用

北京大学 PEKING UNIVERSITY

人类水平游戏

AlphaGo vs. Lee Sedol 4:1

Let us watch a video!

机器学习前沿应用

北京大学 PEKING UNIVERSITY

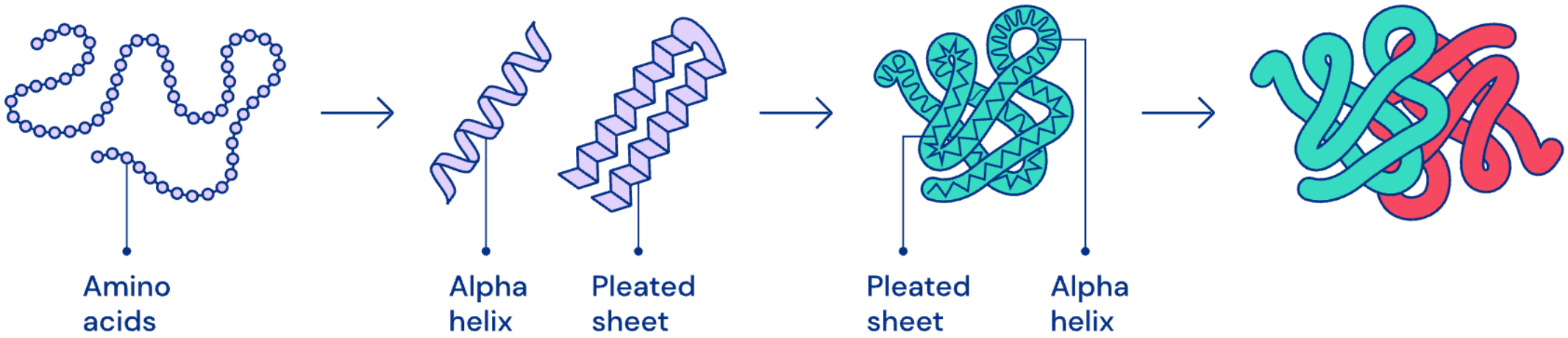
自动驾驶？仍然充满困难

Let us watch a video!

机器学习前沿应用

北京大学 PEKING UNIVERSITY

蛋白质结构预测



Let us watch a video!

机器学习模型分类

北京大学 PEKING UNIVERSITY

Three types of models:

- 判别式模型 (Discriminative Models)
  - Given input x, predict label y
  - Learn p(y|x)
- 描述式模型 (Descriptive Models)
  - Given input x, describe its probability.
  - Learn p(x)
- 生成式模型 (Generative Models)
  - Given a latent vector z, generate x from z
  - Learn p(x|z)

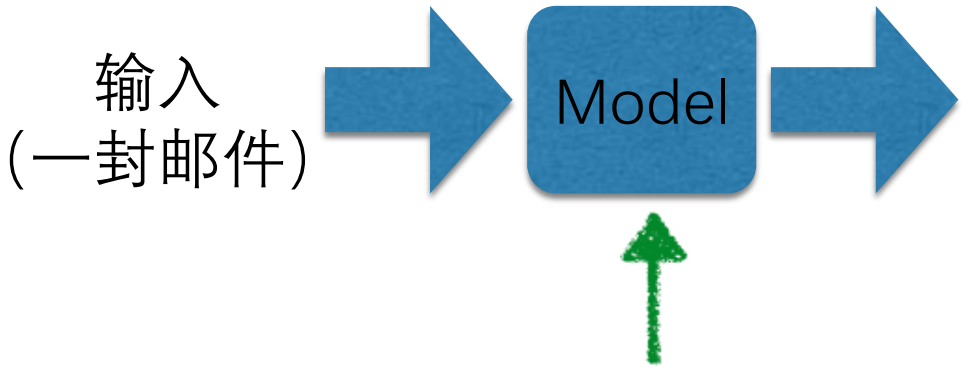
Two settings:

- 监督式学习 (Supervised Learning)
  - Need labels y
- 非监督式学习 (Unsupervised Learning)
  - Do not need labels y
- 强化学习 (Reinforcement Learning)
  - Interact with the environment to gather data

判别式模型

模型 (Model): 一个包含参数 (parameters) 的函数

标签 (Label): 要预测的类别或数值



模型训练/学习 (training/learning): 通过调整模型参数来拟合 (fit) 训练数据 (training data)

分类 vs 回归

- 回归 (regression): 标签 (label) 是连续 (continuous) 值
  - 比如，预测一个房子的房价
- 分类 (classification): 标签 (label) 是离散 (discrete) 值
  - 比如，预测狗的品种

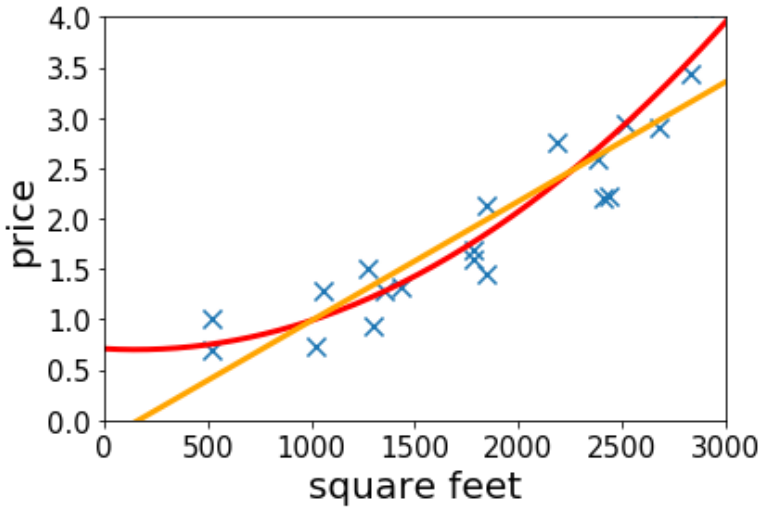
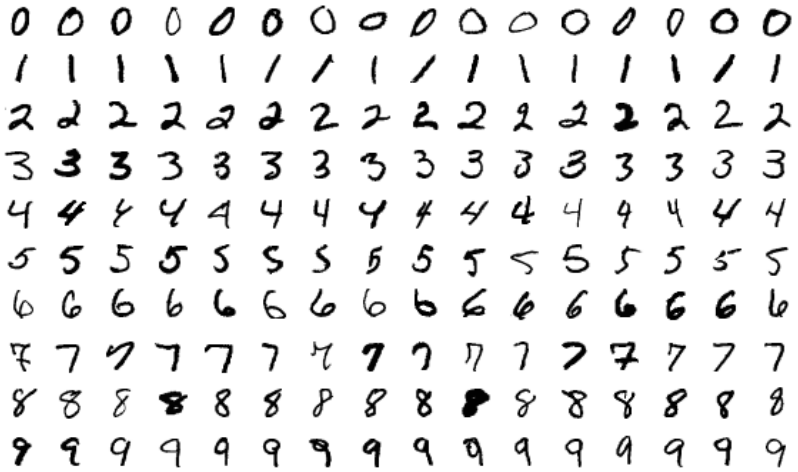


Image credit: Stanford CS229 and Internet

分类 or 回归?

- 手写数字识别 Classification
- 垃圾邮件识别 Classification
- 预测某只股票明天的股价 Regression
- 预测豆瓣电影评分 Can be either classification or regression!

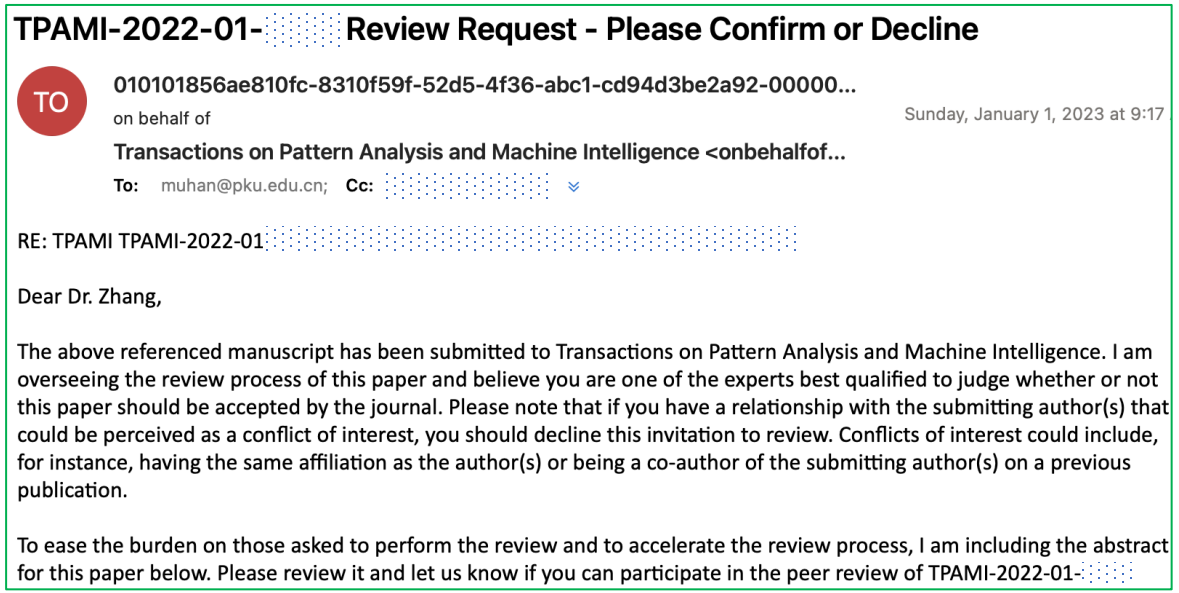


MNIST handwritten recognition dataset

Douban movie rating: 1 star to 5 stars

示例：垃圾邮件分类

输入: x = 邮件



TPAMI-2022-01- Review Request - Please Confirm or Decline

TO

010101856ae810fc-8310f59f-52d5-4f36-abc1-cd94d3be2a92-00000...

on behalf of

Sunday, January 1, 2023 at 9:17

Transactions on Pattern Analysis and Machine Intelligence <onbehalfof...

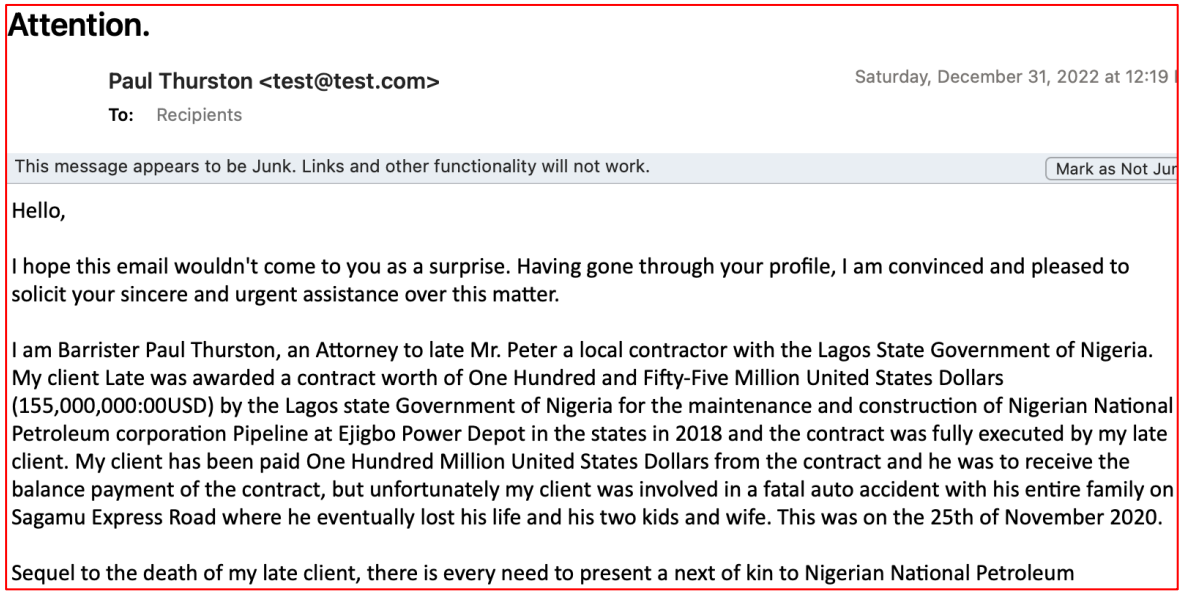
To: muhan@pku.edu.cn; Cc:

RE: TPAMI TPAMI-2022-01

Dear Dr. Zhang,

The above referenced manuscript has been submitted to Transactions on Pattern Analysis and Machine Intelligence. I am overseeing the review process of this paper and believe you are one of the experts best qualified to judge whether or not this paper should be accepted by the journal. Please note that if you have a relationship with the submitting author(s) that could be perceived as a conflict of interest, you should decline this invitation to review. Conflicts of interest could include, for instance, having the same affiliation as the author(s) or being a co-author of the submitting author(s) on a previous publication.

To ease the burden on those asked to perform the review and to accelerate the review process, I am including the abstract for this paper below. Please review it and let us know if you can participate in the peer review of TPAMI-2022-01-



Attention.

Paul Thurston <test@test.com>

Saturday, December 31, 2022 at 12:19

To: Recipients

This message appears to be Junk. Links and other functionality will not work. Mark as Not Jur

Hello,

I hope this email wouldn't come to you as a surprise. Having gone through your profile, I am convinced and pleased to solicit your sincere and urgent assistance over this matter.

I am Barrister Paul Thurston, an Attorney to late Mr. Peter a local contractor with the Lagos State Government of Nigeria. My client Late was awarded a contract worth of One Hundred and Fifty-Five Million United States Dollars (155,000,000:00USD) by the Lagos state Government of Nigeria for the maintenance and construction of Nigerian National Petroleum corporation Pipeline at Ejigbo Power Depot in the states in 2018 and the contract was fully executed by my late client. My client has been paid One Hundred Million United States Dollars from the contract and he was to receive the balance payment of the contract, but unfortunately my client was involved in a fatal auto accident with his entire family on Sagamu Express Road where he eventually lost his life and his two kids and wife. This was on the 25th of November 2020.

Sequel to the death of my late client, there is every need to present a next of kin to Nigerian National Petroleum

输出: $y \in \{$正常邮件，垃圾邮件$\}$

目标：训练模型 f，使得 $f(x) = y$

模型训练过程 (Training)

训练数据 (Training Data)

(email1, not_spam)
(email2, spam)
(email3, spam)
……
(email97, not_spam)
(email98, spam)
(email99, not_spam)
(email100, spam)

特征提取
(Feature extraction)

x_i: 特征向量 (feature vector)

y_i : 标签 (label)

(x_1, y_1)
(x_2, y_2)
$\cdots$
(x_{100}, y_{100})

x_i 词频向量

attention:	1
million:	2
dollars:	1
prize:	0
payment:	3
government:	1
…	

$y_i = \{-1, 1\}$

模型训练
(Model Training)

Learning a function f

$$y_i \approx f(x_i) \quad \forall i$$

模型评估 (Testing/Evaluation)

- 训练误差 (training error): 在训练集上的平均误差
  - 通过**最小化**训练误差来训练模型
  - 对分类问题，通常用错误率来衡量训练误差。即，分类错误的样本 / 总样本数

- 测试误差 (testing error): 在测试集上的平均误差
  - 训练完成后，用来真正衡量模型在**新数据**上的好坏
  - 衡量模型的泛化 (generalization) 能力
  - 训练误差低不代表测试误差一定也低——过拟合!

- 训练集（training set）、测试集（test set）划分
  - 给定全部的数据，按一定比例（如90%-10%）随机划分为训练集和测试集
  - 若按时间顺序排列，将前90%划分为训练集，后10%划分为测试集

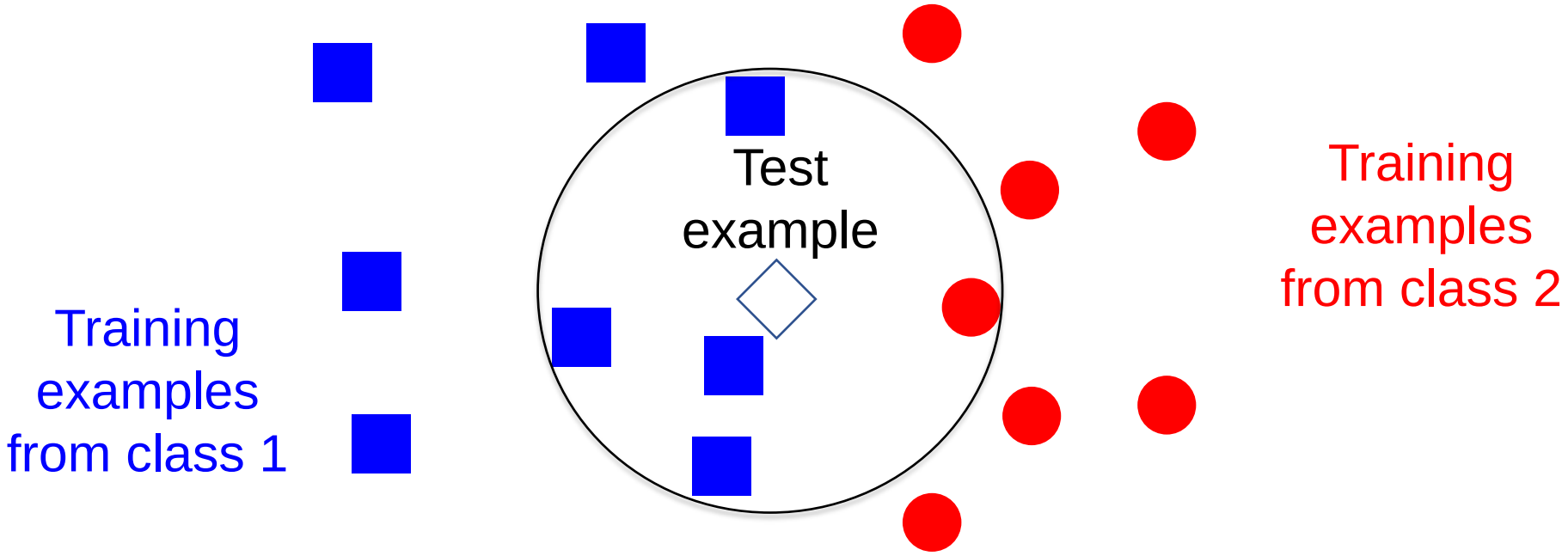
模型评估 (Testing/Evaluation)

- 过拟合 (overfitting): 测试误差远远大于训练误差
  - 比如，训练样本中的10封垃圾邮件恰好都包含dollars这个词，而90封正常邮件都不含
  - 那么训练一个只靠这个单词判断垃圾邮件的模型将在训练集上达到0误差
  - 思考：10封垃圾邮件有代表性吗？未来所有新邮件都可以只靠这个单词来判断吗？
  - 错把训练样本中找到的特殊规律当做了普遍规律——应避免这种现象

- 欠拟合 (underfitting): 训练完成后，训练误差仍然很大
  - 说明连训练样本都没有拟合好
  - 即，在训练集中都难以做到 $f(x_i) \approx y_i$

k近邻 (k-Nearest Neighbor, k-NN) 算法

- 一个最简单机器学习算法
  - 对于一个测试样本，用训练样本中距离它最近的k个样本中占多数的标签来预测测试样本



k近邻 (k-Nearest Neighbor, k-NN) 算法

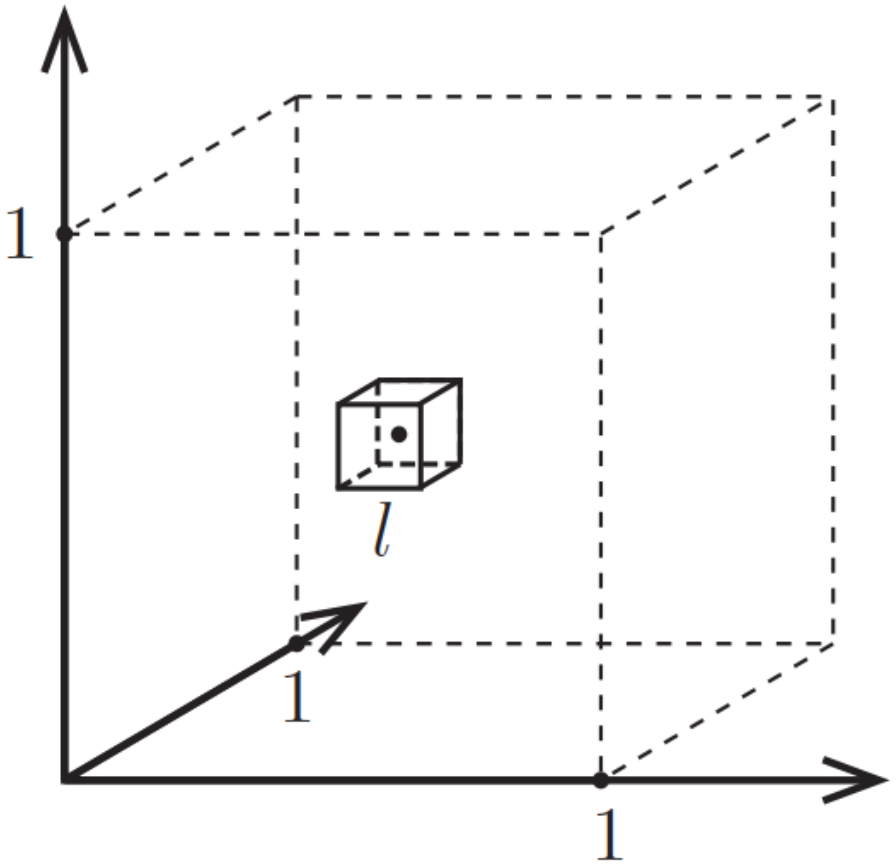
- 优点：
  - 不需要训练
  - 只需要一个距离函数即可，如欧氏距离 $\|x - x'\| = \sqrt{\sum_{j\in[d]}\left(x_1^{(j)} - x_2^{(j)}\right)^2}$，d为维度
- 缺点：
  - 需要存储所有训练样本
  - 在测试时需要计算测试样本到所有训练样本的距离
  - 有时很难找到一个好的距离函数

$\|$  $-$  $\|$

  - 欧氏距离 $\|x_1 - x_2\|$？两幅图片的欧氏距离不一定能反映其相似度

维度灾难（Curse of Dimensionality）

- 假设 k=10，总共有n个训练样本从 $[0,1]^d$ 中均匀采样



- 对某个测试样本，设 l 为能包含住其所有k近邻的超立方体的边长
- 则 $l^d \approx \frac{k}{n} \rightarrow l \approx \sqrt[d]{\frac{k}{n}}$
- 当 $n = 1000$，我们来计算 l

d	ℓ
2	0.1
10	0.63
100	0.955
1000	0.9954

- 距离在高维空间失去意义，不再能很好地衡量远近
- k-NN不适合高维空间

Image credit: Kilian Weinberger

非参数化模型 vs 参数化模型

- 非参数化模型：
  - 例如：k-NN
  - 模型不包含参数 (parameters)
  - 需要保留训练样本，以对测试样本做出预测

- 参数化模型：
  - 例如：线性回归、神经网络等
  - 模型包含可训练的参数，通过拟合训练数据来估算模型参数
  - 训练好模型参数后，可以丢弃训练数据，仅依靠模型参数去预测新样本
  - 可以写成 $y \approx f(x), f$ 为包含参数的模型

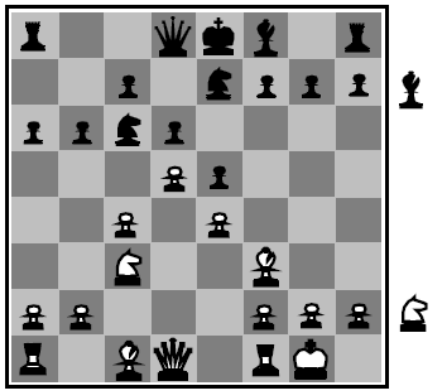
最简单的参数化模型——线性模型

- 线性模型 (Linear Models)

  $f(x) = w^T x + b$

  - 输入 $x \in \mathbb{R}^d$
  - 参数 $w \in \mathbb{R}^d, b \in \mathbb{R}$，分别称为 权重 (weight) 和 偏置 (bias)
  - 输出 $f(x) \in \mathbb{R}$
  - 为每个特征维度学习一个权重，反映其重要性
- 线性回归 (Linear Regression)
  - 使用线性模型做回归任务
  - 标签 $y \in \mathbb{R}$

E.g. 盘面评估模型



Black to move

White slightly better

$$Eval(s) = w_1 f_1(s) + w_2 f_2(s) + \ldots + w_n f_n(s)$$

e.g. $f_1(s)$ = (num white queens – num black queens)

线性回归例子

- 预测房价
  - 输入特征维度：面积（平米）、房龄、楼层、学区等级（1-5）、物业水平（1-5）
  - 学习到的权重 $w = [10, -10, -5, 50, 2]^T$
  - 学习到的偏置 $b = 100$
  - 输入向量 $x = [100, 10, 5, 5, 5]^T$
  - 线性模型：$f(x) = w^T x + b$
  - 输出预测：$100 \times 10 + 10 \times (-10) + 5 \times (-5) + 5 \times 50 + 5 \times 2 + 100 = 1085$
  - 偏置的作用：提供一个基础预测值，所有特征在其基础上加权累加

线性回归训练

- 如何训练?
  - 通过最小化损失函数 (loss function)
- 平方损失函数 (squared loss)

$$L(f(x_i), y_i) = (f(x_i) - y_i)^2$$

  - 例如，$f(x_i) = 8,\ y_i = 9,\ L(f(x_i), y_i) = (8-9)^2 = 1$
  - 惩罚那些预测值 (prediction) 偏离真实值 (groundtruth) 太大的情况
  - 也称最小二乘法 (least squares)
- 通过在训练集上最小化平均损失函数来优化参数w, b

$$\min_{w,b} \frac{1}{n} \sum_{i \in [n]} L(f(x_i), y_i)$$

  - $[n] = \{1,2,\dots,n\}$，n 为训练样本的总数

线性回归训练

- 如何对以下最优化问题求解?

$$\min_{w,b} \frac{1}{n} \sum_{i \in [n]} L(f(x_i), y_i)$$

  - 将要优化的目标 (objective) 写作变量 w, b 的函数

$$J(w,b) = \frac{1}{n} \sum_{i \in [n]} L(f(x_i), y_i)$$

  - 随机选定 w, b 的初始值，逐步调整 w, b 以降低 $J(w,b)$
  - 每次沿着使 $J(w,b)$ 变小最快的方向使 w, b 走一小步

$$w \leftarrow w - \alpha \cdot \frac{\partial J(w,b)}{\partial w}, \qquad b \leftarrow b - \alpha \cdot \frac{\partial J(w,b)}{\partial b}$$

  - 梯度下降 (Gradient Decent)

梯度下降

$$w \leftarrow w - \alpha \cdot \frac{\partial J(w,b)}{\partial w}, \qquad b \leftarrow b - \alpha \cdot \frac{\partial J(w,b)}{\partial b}$$

- $\frac{\partial J(w,b)}{\partial w} = \begin{bmatrix} \frac{\partial J(w,b)}{\partial w_1} \\ \cdots \\ \frac{\partial J(w,b)}{\partial w_d} \end{bmatrix} \in \mathbb{R}^d$, $\frac{\partial J(w,b)}{\partial b} \in \mathbb{R}$，两者一同构成了 $J(w,b)$ 的梯度 (gradient)

$$\nabla J(w,b) = \begin{bmatrix} \frac{\partial J(w,b)}{\partial w} \\ \frac{\partial J(w,b)}{\partial b} \end{bmatrix} \in \mathbb{R}^{d+1}$$

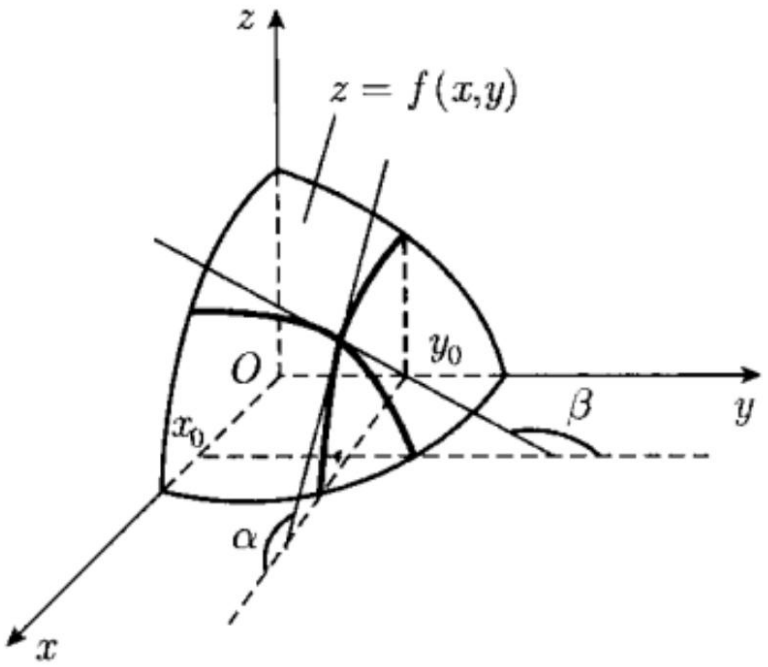
- $\alpha \in \mathbb{R}$
  步长 (step size)，是一个事先指定的超参数 (hyperparameter)，不随w, b 一起优化
- 梯度下降：沿梯度的相反方向走一步，步长为 α
- 相当于在每个维度上，沿着负偏导的方向移动一定距离，距离正比于偏导的绝对值，比例为 α

理解梯度的意义

- $\left[\frac{\partial f(x,y)}{\partial x}, \frac{\partial f(x,y)}{\partial y}\right]^T$ 定义了f 在 (x, y) 点的梯度
- 在任意方向$[\Delta x, \Delta y\ ]^T$上移动无穷小固定距离， 函数大小变化为

$$\Delta f = f(x+\Delta x, y+\Delta y) - f(x,y)$$

$$= \frac{\partial f(x,y)}{\partial x}\Delta x + \frac{\partial f(x,y)}{\partial y}\Delta y = < \begin{bmatrix} \frac{\partial f(x,y)}{\partial x} \\ \frac{\partial f(x,y)}{\partial y} \end{bmatrix}, \begin{bmatrix} \Delta x \\ \Delta y \end{bmatrix} > = \left\| \begin{bmatrix} \frac{\partial f(x,y)}{\partial x} \\ \frac{\partial f(x,y)}{\partial y} \end{bmatrix} \right\| \cdot \left\| \begin{bmatrix} \Delta x \\ \Delta y \end{bmatrix} \right\| \cos(\theta)$$



- Δf 最小当且仅当夹角 $\theta = \pi$， 即$[\Delta x, \Delta y\ ]^T$指向梯度$\left[\frac{\partial f(x,y)}{\partial x}, \frac{\partial f(x,y)}{\partial y}\right]^T$反方向
- 梯度给出了在所有可能方向上移动相同小距离使函数值上升最快的方向(steepest ascent)。
- 梯度下降：在使函数下降最快的方向（负梯度）上走一个小步长 α

image from 《数学分析》， 伍胜健

梯度下降

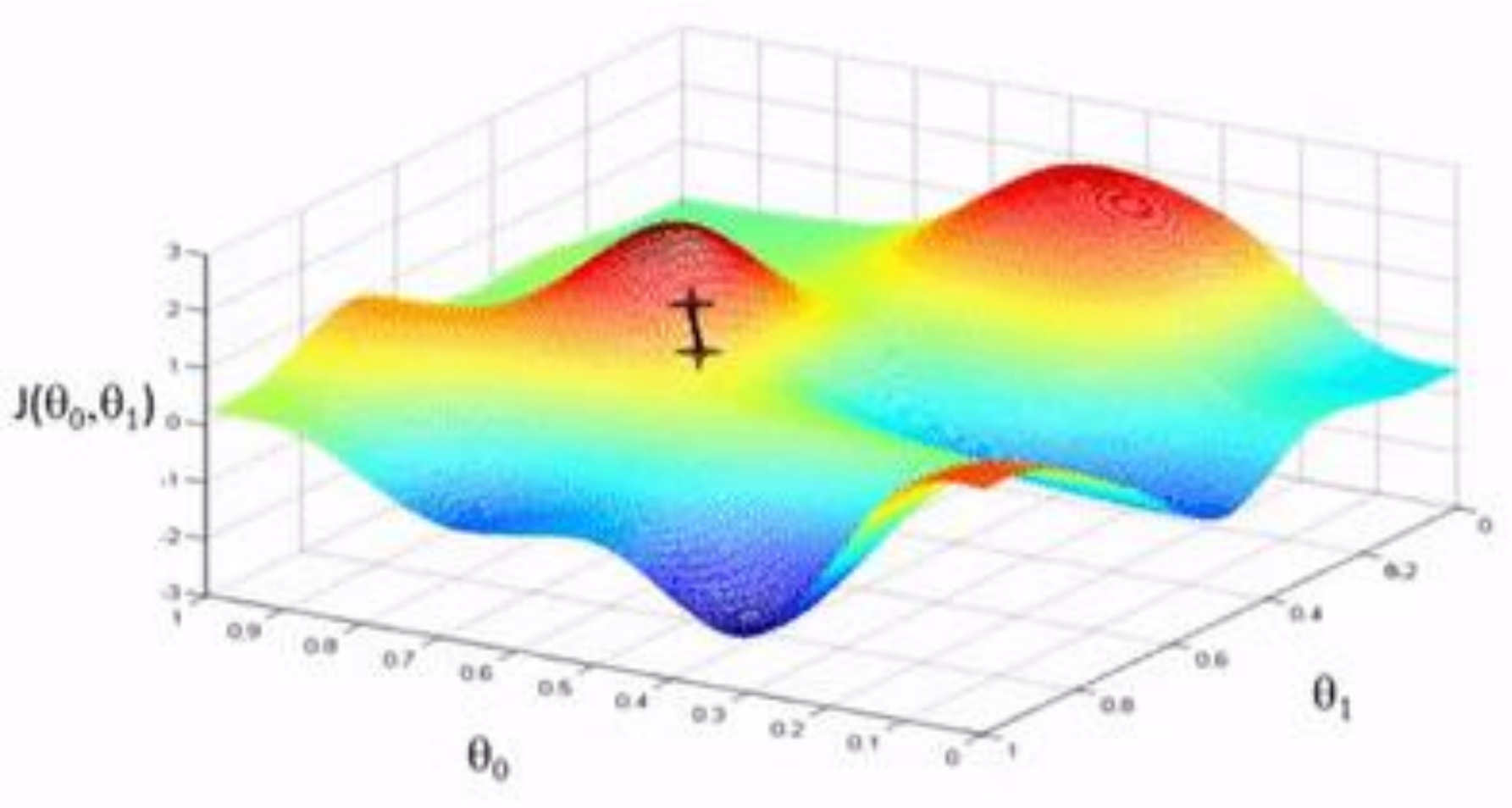


image from Andrew Ng

线性回归的梯度下降

- $J(w,b)=\frac{1}{n}\sum_{i\in[n]}L(f(x_i),y_i)=\frac{1}{n}\sum_{i\in[n]}(w^Tx_i+b-y_i)^2$
- $\frac{\partial J(w,b)}{\partial w}=\frac{2}{n}\sum_{i\in[n]}(w^Tx_i+b-y_i)\,x_i\in\mathbb{R}^d$
- $\frac{\partial J(w,b)}{\partial b}=\frac{2}{n}\sum_{i\in[n]}(w^Tx_i+b-y_i)\in\mathbb{R}$

- 定义 $w^Tx_i+b-y_i=e_i\in\mathbb{R}$，线性回归的梯度下降公式为：

$$w\leftarrow w-\alpha\cdot\frac{2}{n}\sum_{i\in[n]}e_i\,x_i$$

$$b\leftarrow b-\alpha\cdot\frac{2}{n}\sum_{i\in[n]}e_i$$

- 迭代直到 $J(w,b)$ 无法再下降（比如两次的差小于1e-4）或达到预设的最大次数

谢谢

北京大学
PEKING UNIVERSITY