人工智能导论

李文

计算机科学与工程学院

数据智能研究组（Data Intelligence Group）：

http://dig.uestc.cn/

01 机器学习概述

机器学习

- 什么是机器学习?

你们如何理解“学习”？

机器学习

- 什么是机器学习？
  - 绪论里的定义

机器学习

- 给机器以“学习”的能力
  - 与“程序”相反；从数据，或者过去的经验中学习自动改进算法
  - 基本（主要）形式：构建一个映射函数

$$y = f(x; \theta)$$

机器学习

- 什么是机器学习？
  - 给机器以“学习”的能力

- “红灯停，绿灯行”
- “11+10=21”
- 牛顿万有引力定律

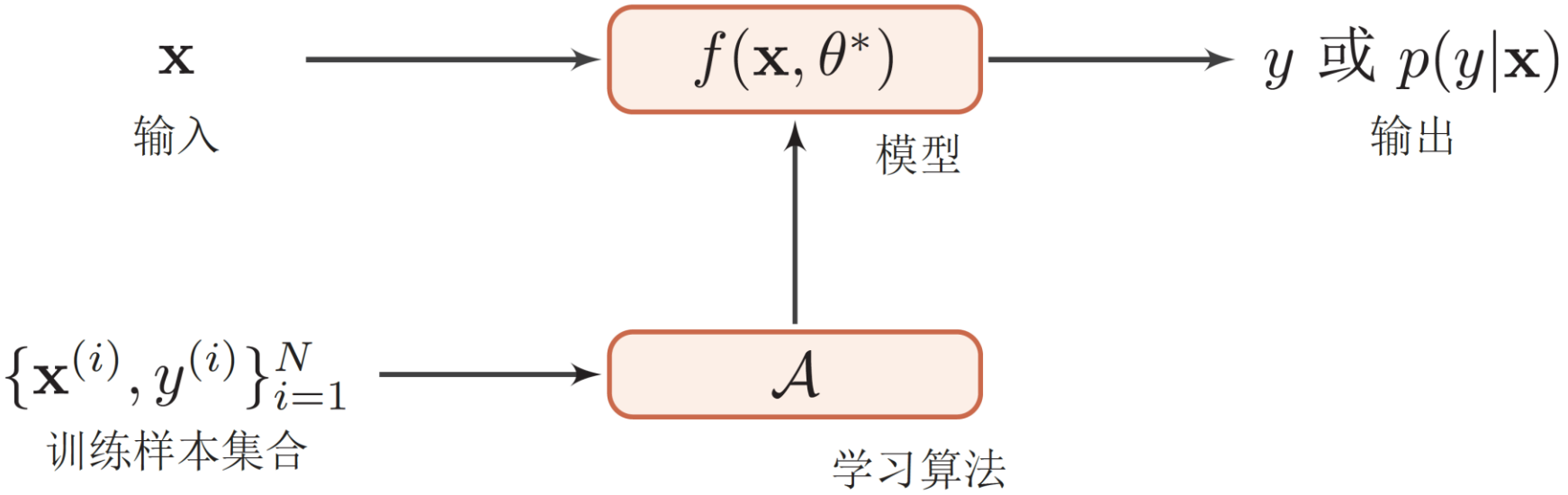
- 通过大量的交通视频观测出车辆遵守“红灯停，绿灯行”
- 小学生通过加法练习，可以正确的算出“11+10=21”
- 牛顿分析天文观测数据，推断出万有引力定律

机器学习

- 什么是机器学习?
  - 给机器以“学习”的能力
- 萨缪尔定义(Arthur Samuel，1959)
  - 机器学习是让计算机具有学习的能力，无需进行明确编程。
- 汤姆·米切尔（Tom Mitchel, 1998）
  - 计算机程序利用经验 E 学习任务 T，性能是 P，如果针对任务 T 的性能 P 随着经验 E 不断增长，则称为机器学习。

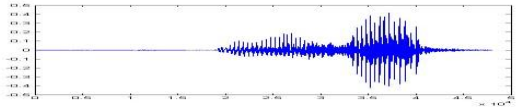
机器学习

- 从数据中获得决策（预测）函数使得机器可以根据数据进行自动学习，通过算法使得机器能从大量历史数据中学习规律从而对新的样本做决策。



机器学习

- 语音识别

$f($  $)=$ “你好”

- 图像识别

$f($  $)=$ “猫”

- 围棋

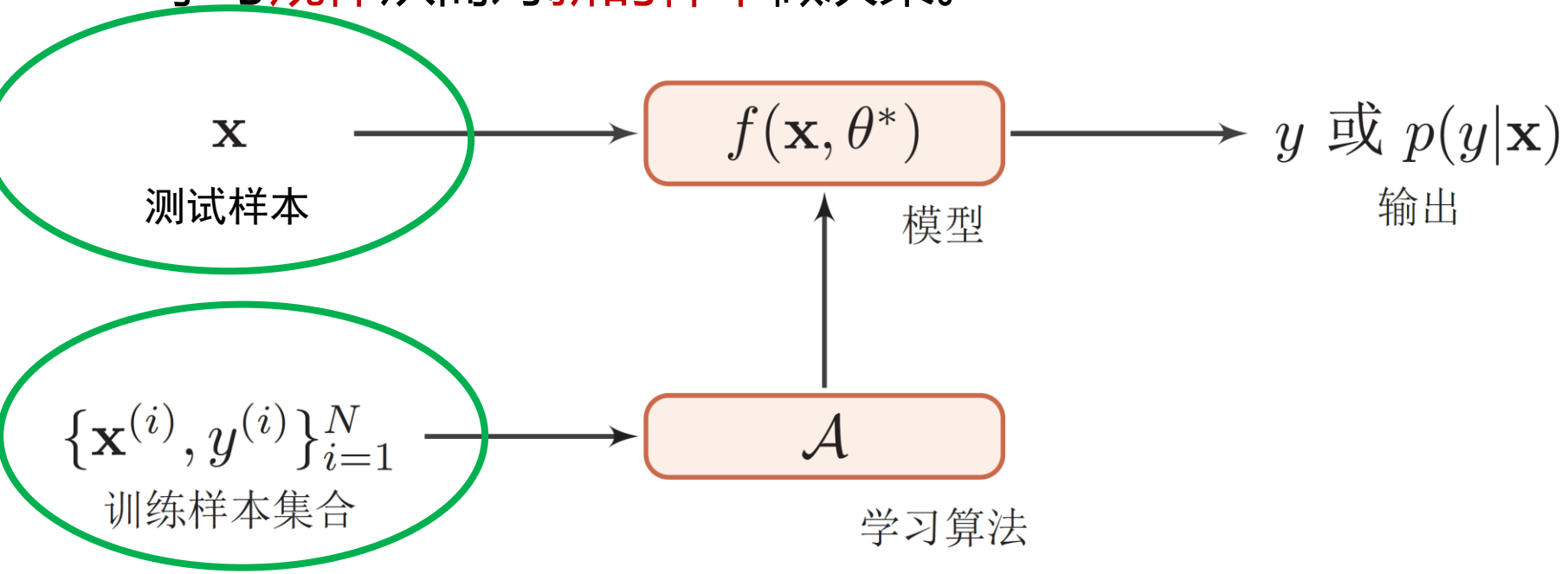
$f($  $)=$ “5-5” (落子位置)

- 对话系统

$f($ “你好” $)=$ “今天天气真不错”

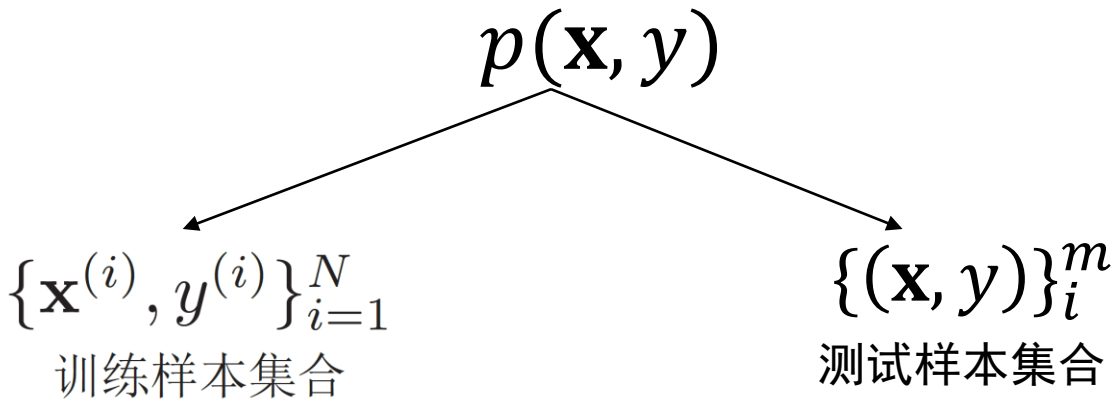
机器学习

- 从数据中获得决策（预测）函数使得机器可以根据数据进行自动学习，通过算法使得机器能从大量历史数据中学习规律从而对新的样本做决策。



机器学习

- 从数据中获得决策（预测）函数使得机器可以根据数据进行自动学习，通过算法使得机器能从大量历史数据中学习规律从而对新的样本做决策。
- 训练样本和测试样本满足独立同分布（independently sampled from an identical distribution）



常见的机器学习问题

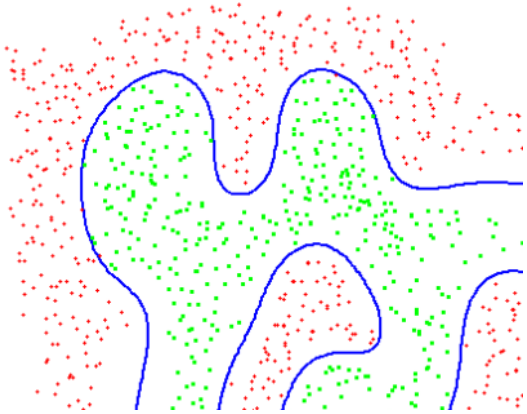
	监督学习	无监督学习	强化学习
训练样本	训练集 $\{(\mathbf{x}^{(n)}, y^{(n)})\}_{n=1}^{N}$	训练集 $\{\mathbf{x}^{n}\}_{n=1}^{N}$	智能体和环境交互的轨迹 τ 和累积奖励 G_τ
优化目标	$y = f(\mathbf{x})$ 或 $p(y\|\mathbf{x})$	$p(\mathbf{x})$ 或带隐变量 $\mathbf{z}$ 的 $p(\mathbf{x}\|\mathbf{z})$	期望总回报 $\mathbb{E}_\tau[G_\tau]$
学习准则	期望风险最小化 最大似然估计	最大似然估计 最小重构错误	策略评估 策略改进

常见的机器学习问题

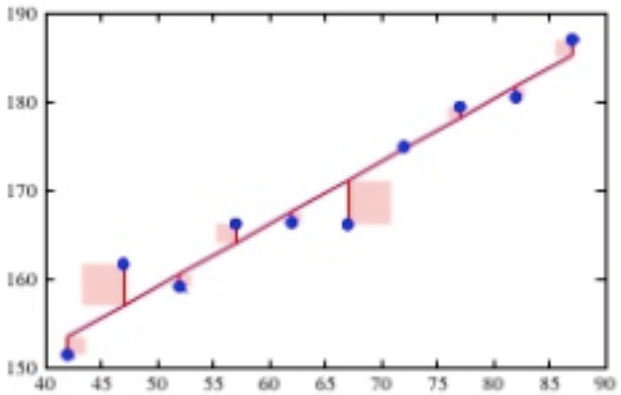
	监督学习	无监督学习	强化学习
训练样本	训练集 $\{(\mathbf{x}^{(n)}, y^{(n)})\}_{n=1}^{N}$	训练集 $\{\mathbf{x}^{n}\}_{n=1}^{N}$	智能体和环境交互的轨迹 τ 和累积奖励 G_τ
优化目标	$y = f(\mathbf{x})$ 或 $p(y\|\mathbf{x})$	$p(\mathbf{x})$ 或带隐变量 $\mathbf{z}$ 的 $p(\mathbf{x}\|\mathbf{z})$	期望总回报 $\mathbb{E}_\tau[G_\tau]$
学习准则	期望风险最小化 最大似然估计	最大似然估计 最小重构错误	策略评估 策略改进

常见的机器学习问题

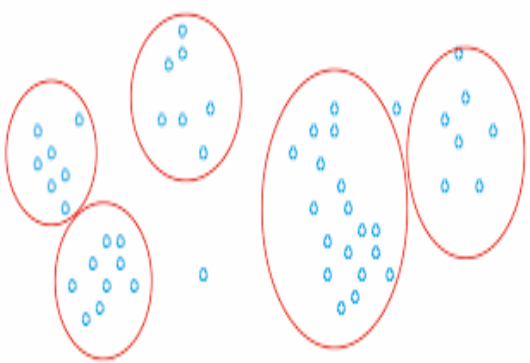
- 分类（Classification）
- 回归（Regression）
- 聚类（Clustering）



分类



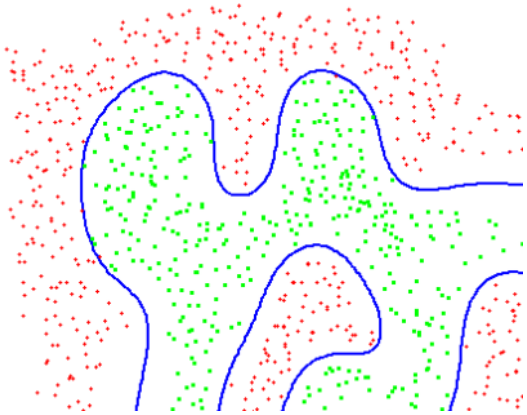
回归



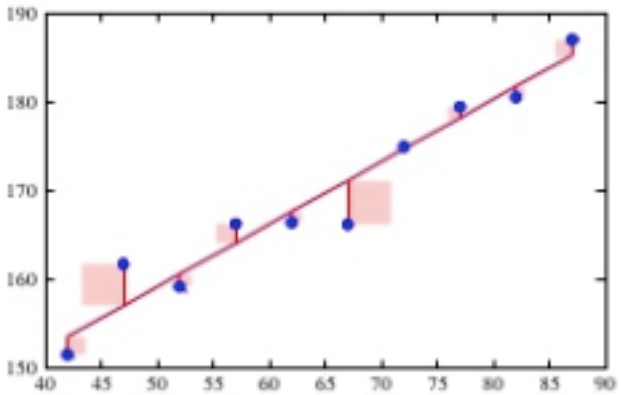
聚类

常见的机器学习问题

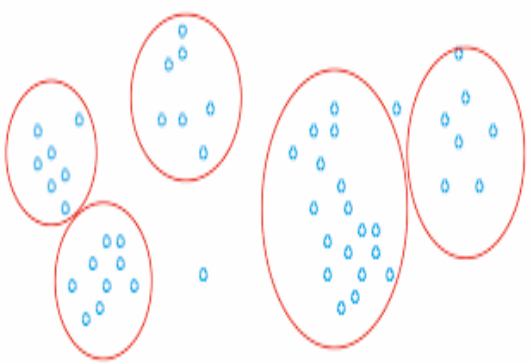
- 分类（Classification）
- 回归（Regression）
- 聚类（Clustering）



分类



回归

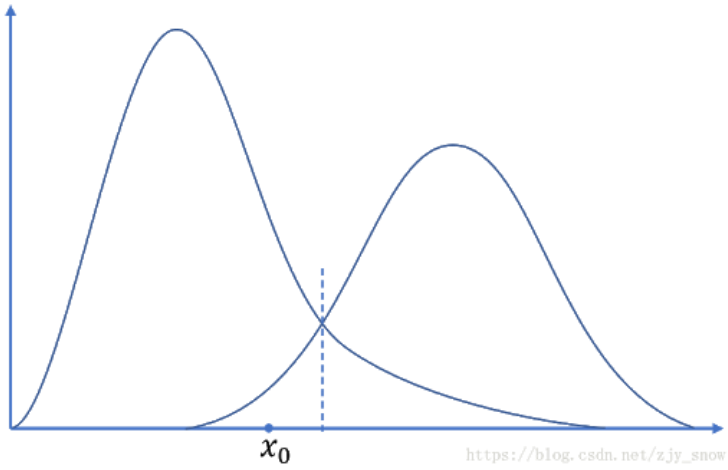


聚类

朴素贝叶斯分类器

- 后验概率
  - x为一张图像，y为标签（猫或者非猫）
  - 目标：求f(x)或者p(y|x)
- 贝叶斯公式

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



- 问题
  - 概率估计本身是一个难题
  - 我们通常面对的是样本

机器学习的三要素

模型

- 线性方法：$f(\mathbf{x}, \theta) = \mathbf{w}^{\mathrm{T}}\mathbf{x} + b$
- 非线性方法：$f(\mathbf{x}, \theta) = \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}) + b$
  - 如果ϕ(x)为可学习的非线性基函数，f(**x**,θ)就等价于神经网络

学习准则

- 期望风险 $\mathcal{R}(f) = \mathbb{E}_{(\mathbf{x},y)\sim p(\mathbf{x},y)}[\mathcal{L}(f(\mathbf{x}), y)],$

优化

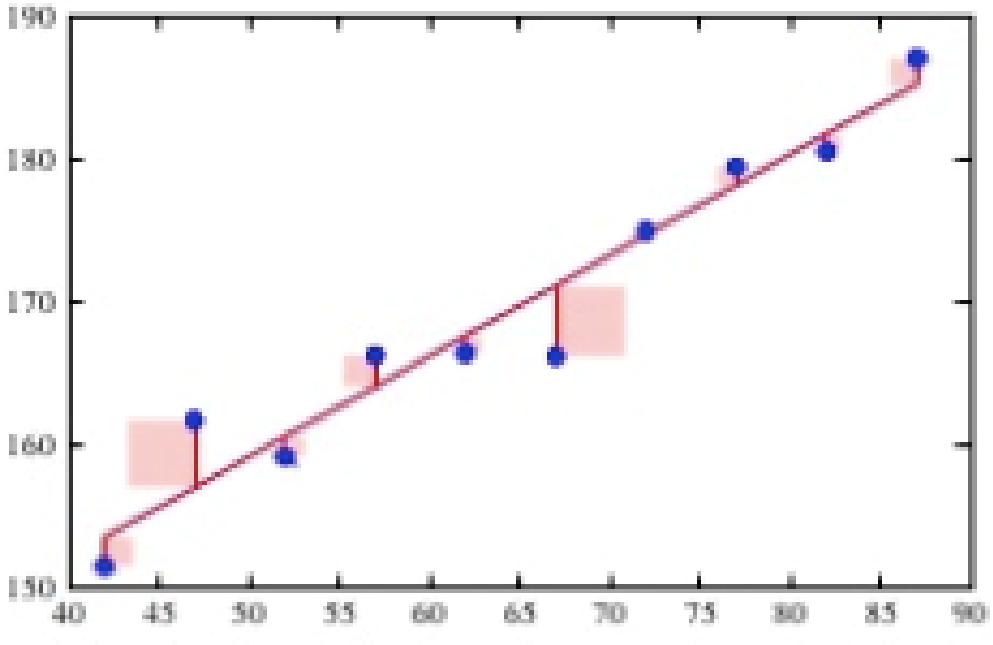
- 梯度下降

02 一个例子：线性回归

线性回归

- 根据人的体重来预测身高：
  - 身高=f(体重)；

$$f(x_i) = w^T X + b$$



线性回归

- 根据人的体重来预测身高：
  - 身高=f(体重)；

$$f(x_i) = w^T X + b$$

  - 使用均方误差（平均损失函数）

$$J = \sum_{i=1}^{n} (f(x_i) - y_i)^2 = \sum_{i=1}^{n} (y_i - wx_i - b)^2$$

  - 优化求解w和b

$$\min_{w,b} J(w, b)$$

03 损失函数

损失函数

概念

In mathematical optimization and decision theory, a **loss function** or **cost function** is a function that maps an event or values of one or more variables onto a real number intuitively representing some "cost" associated with the event.

要素

- 衡量真值与预测值之间的差异
- 输出一个实数作为指标

直观理解

- 打分，评价学习的好坏

损失函数

- 0-1损失函数（Binary Loss）

$$\mathcal{L}(y, f(x,\theta)) = \begin{cases} 0 & \textbf{if } y = f(x,\theta) \\ 1 & \textbf{if } y \neq f(x,\theta) \end{cases}$$

- 平方损失函数（Squared Loss）

$$\mathcal{L}(y, \hat{y}) = (y - f(x,\theta))^2$$

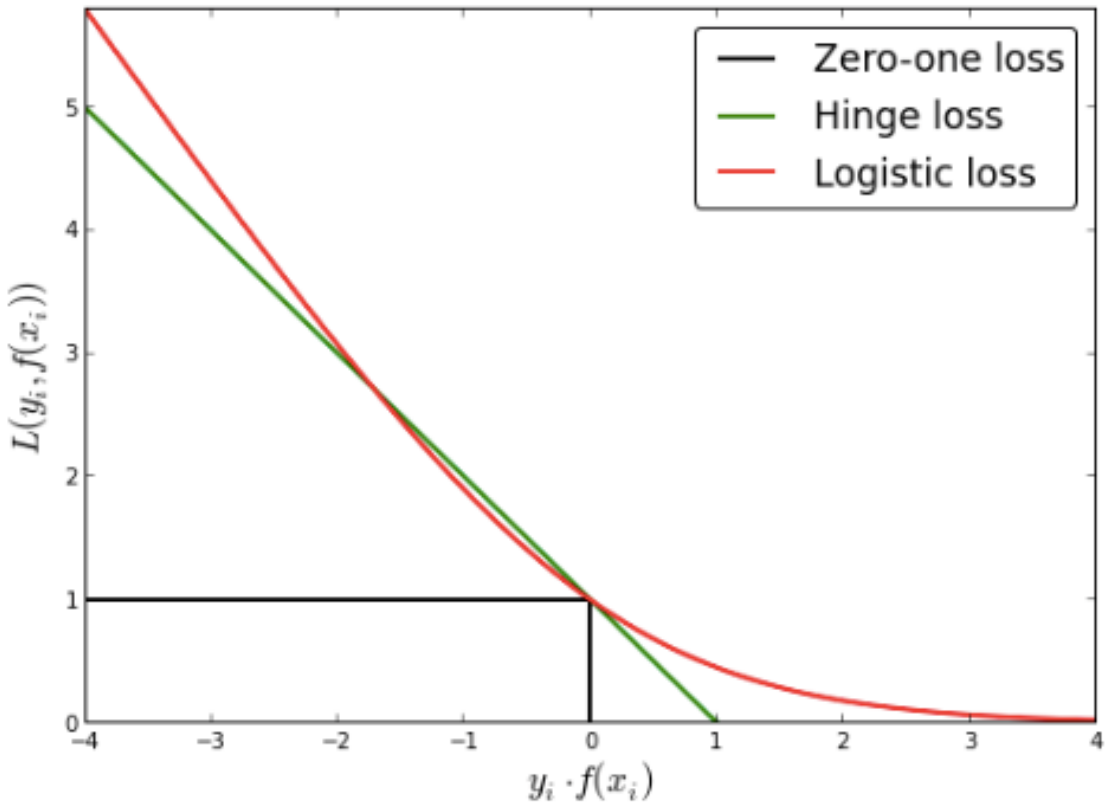
- 合页损失函数（Hinge Loss）

$$\mathcal{L}(y, \hat{y}) = \max\big(0, 1 - y \cdot f(x)\big)$$

损失函数

- 逻辑损失函数（Logistic Loss）

$$L\Big(y, f(\mathbf{x})\Big) \quad = \quad \log\Big(1 + \exp(-yf(\mathbf{x})\Big)$$



后验概率的损失函数

- 直接建模条件概率 $p_\theta(y|x)$
- 真实条件概率 $p_r(y|x)$
- 如何衡量两个条件分布的差异?

后验概率的损失函数

- 信息熵（Entropy）
  - **什么是信息：信息是用来消除随机不定性的东西 (Information is the resolution of uncertainty)**
  - 信息多 vs. 信息少
    - 一本五十万字的中文书到底有多少信息量?
  - 如何对信息进行度量?
  - 信息内容量数学定义： $h_i = \log \frac{1}{P_i} = -\log P_i$

后验概率的损失函数

- 信息熵（Entropy）
  - 信息多 vs. 信息少
    - 一本五十万字的中文书到底有多少信息量?
  - 如何对信息进行度量?
  - 信息内容量数学定义:

$$h_i = \log \frac{1}{P_i} = -\log P_i$$

  - 信息熵数学定义:

$$S = -\sum_i P_i \log P_i = -\mathrm{E}_P[\log P]$$

    - 信息内容量的期望值

后验概率的损失函数

Entropy, Cross-Entropy, & KL-Divergence

Aurélien Géron
February 2018

后验概率的损失函数

- KL散度（Kullback–Leibler Divergence）
  - 考虑某个未知的真实分布 p(x)，假定用一个近似的分布 q(x) 对它进行建模。如果我们使用 q(x) 来建立一个编码体系，用来把 x 的值传给接收者，那么由于我们使用了q(x)而不是真实分布p(x)，平均编码长度比用真实分布p(x)进行编码增加的信息量为：

$$KL(p||q) = -\int p(x)\ln q(x)dx - (-\int p(x)\ln p(x)dx)$$

$$= -\int p(x)\ln[\frac{q(x)}{p(x)}]dx$$

后验概率的损失函数

- KL散度（Kullback–Leibler Divergence）
  - 离散化表示：

$$\mathrm{D_{kl}}(p_r(y|x)||p_\theta(y|x)) = \sum_{y=1}^{C} p_r(y|x) \log \frac{p_r(y|x)}{p_\theta(y|x)}$$

$$\propto -\sum_{y=1}^{C} p_r(y|x) \log p_\theta(y|x)$$

交叉熵损失

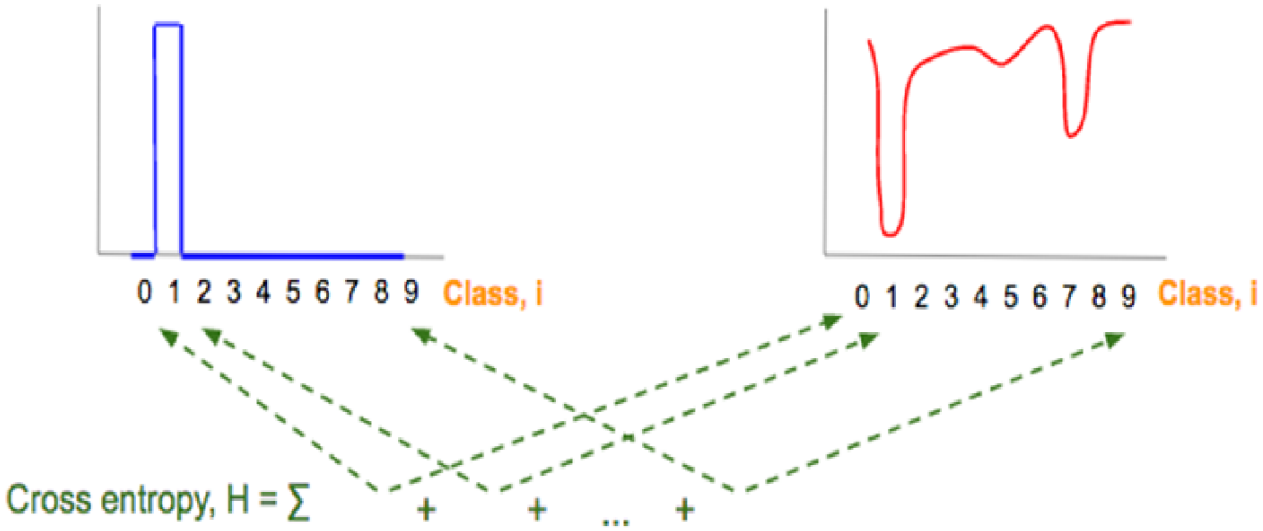
建模条件概率 $p_\theta(y|x)$
真实条件概率 $p_r(y|x)$

交叉熵损失

$$-\sum_{y=1}^{C} p_r(y|x) \log p_\theta(y|x)$$

真实概率$p_r(y|x)$

预测概率的负对数$-\log p_\theta(y|x)$



交叉熵损失函数

- 负对数似然损失函数

$$\mathcal{L}(\mathbf{y}, f(\mathbf{x}, \theta)) = -\sum_{c=1}^{C} y_c \log f_c(\mathbf{x}, \theta)$$

- 对于一个三类分类问题，类别为[0,0,1]，预测的类别概率为[0.3,0.3,0.4]，则

Ex:

Computed ($\hat{y}$)	Targets (y)
[0.3, 0.3, 0.4]	[0, 0, 1]

$$\begin{aligned}\mathcal{L}(\theta) &= -\big(0 \times \log(0.3) + 0 \times \log(0.3) + 1 \times \log(0.4)\big) \\ &= -\log(0.4).\end{aligned}$$

03 最优化问题

线性回归

- 根据人的体重来预测身高：
  - 身高=f(体重)；

$$f(x_i) = w^T X + b$$

在此处键入公式。

  - 使用均方误差（平均损失函数）

$$J = \sum_{i=1}^{n}(f(x_i) - y_i)^2 = \sum_{i=1}^{n}(y_i - wx_i - b)^2$$

  - 优化求解w和b

$$\min_{w,b} J(w, b)$$

参数学习

- 期望风险未知，通过经验风险近似
  - 训练数据：$\mathcal{D} = \{\mathrm{x}^{(i)}, y^{(i)}\}, i \in [1, N]$

$$\mathcal{R}(f) = \mathbb{E}_{(\mathbf{x},y)\sim p(\mathbf{x},y)}[\mathcal{L}(f(\mathbf{x}), y)],$$

$$\mathcal{R}_{\mathcal{D}}^{emp}(\theta) = \frac{1}{N}\sum_{n=1}^{N}\mathcal{L}(y^{(n)}, f(x^{(n)}, \theta))$$

- 经验风险最小化
  - 在选择合适的风险函数后，我们寻找一个参数θ^*，使得经验风险函数最小化

$$\theta^* = \arg\min_{\theta} \mathcal{R}_{\mathcal{D}}^{emp}(\theta)$$

- 机器学习问题转化成为一个最优化问题

最优化问题

- 最优化问题：在一定的约束条件下，求一个函数的**最大（小）值**

最优化问题

- 在机器学习中，我们一般将最优化问题统一表述为求解函数的极小值问题（不失一般性）：

$$\min_x f(x),$$

- 极大、极小问题可以相互转换：

$$\min_x f(x) \Leftrightarrow \max_x -f(x)$$

最优化问题

- 在机器学习中，我们一般将最优化问题统一表述为求解函数的极小值问题（不失一般性）：

$$\min_{x} f(x),$$

- 如何求解？极值点处导数为 **0**

$$\nabla f(x) = 0$$

最优化问题

- 在机器学习中，我们一般将最优化问题统一表述为求解函数的极小值问题（不失一般性）：

$$\min_{x} f(x),$$

- 如何求解？极值点处导数为 **0**

$$\nabla f(x) = 0$$

最优化问题

- 在机器学习中，我们一般将最优化问题统一表述为求解函数的极小值问题（不失一般性）：

$$\min_x f(x),$$

- 如何求解？极值点处导数为 **0**

$$\nabla f(x) = 0$$

- 多元情况下为梯度

$$\nabla f(x) = \left[\frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_n}\right]^{\mathsf{T}}$$

一元函数的例子

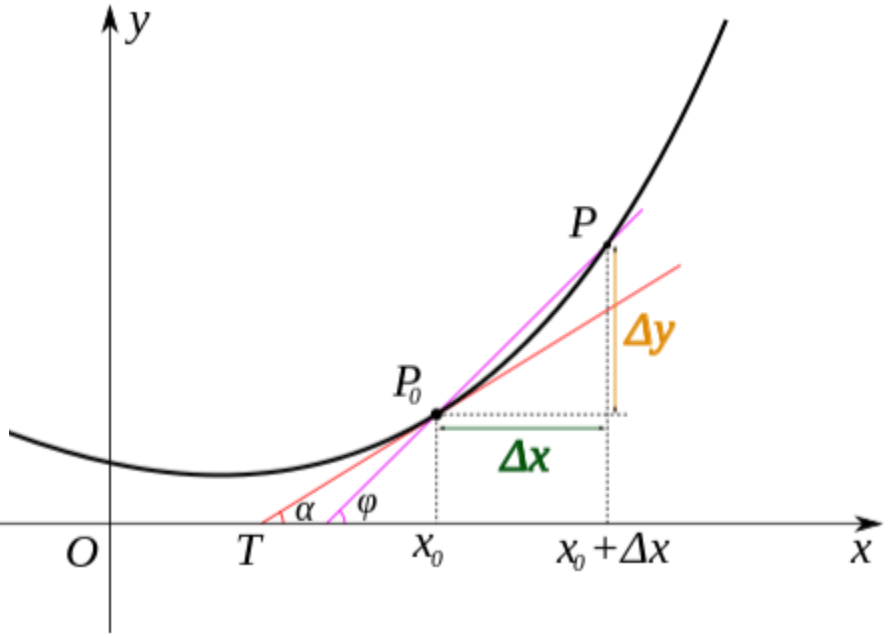
- 一元函数求导

$$f'(x_0)=\lim_{\Delta x\to 0}\frac{\Delta y}{\Delta x}=\lim_{\Delta x\to 0}\frac{f(x_0+\Delta x)-f(x_0)}{\Delta x}$$

- 几何意义：切线的斜率

$$\tan\varphi=\frac{\Delta y}{\Delta x}=\frac{f(x_0+\Delta x)-f(x_0)}{\Delta x}$$

$$\tan\alpha=\lim_{\Delta x\to 0}\tan\varphi=\lim_{\Delta x\to 0}\frac{f(x_0+\Delta x)-f(x_0)}{\Delta x}$$



最优化问题

- 例：$f(x,y)=x^3-2x^2+e^{xy}-y^3+10y^2+100\sin(xy)$

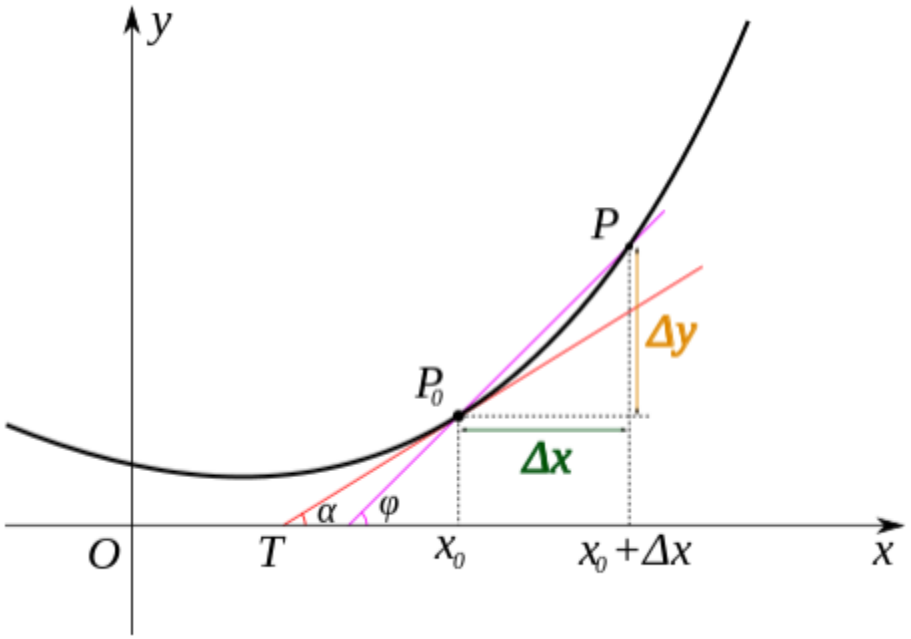
$$\frac{\partial f}{\partial x}=3x^2-4x+ye^{xy}+100y\cos(xy)=0$$

$$\frac{\partial f}{\partial y}=xe^{xy}-3y^2+20y+100x\cos(xy)=0$$

- 上述方程很难精确求解，一般求近似解（数值计算）
  - 迭代法

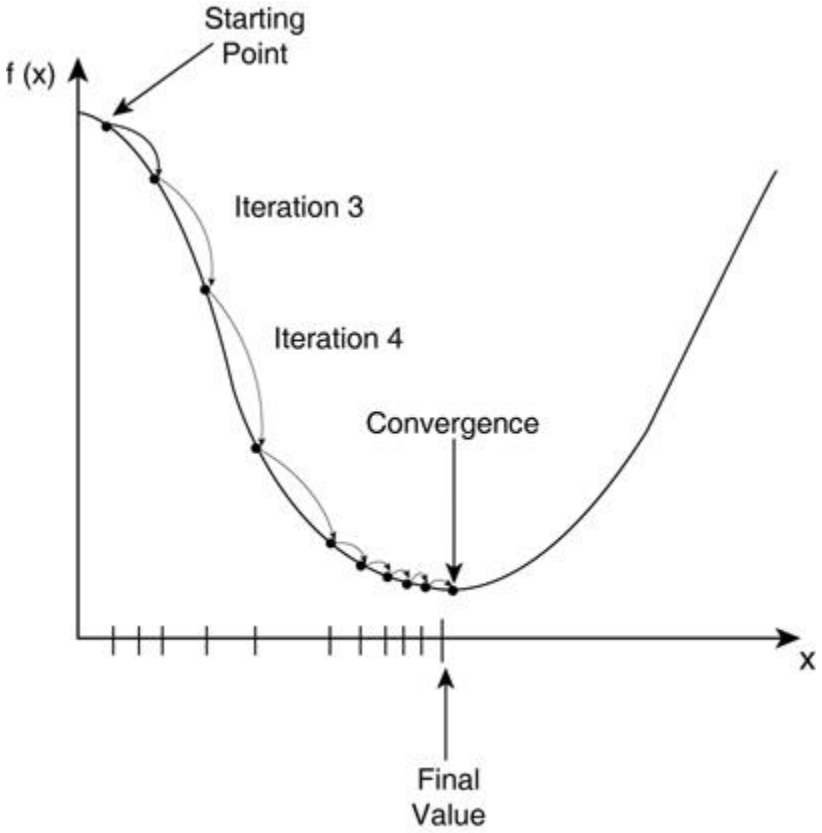
一元函数的例子

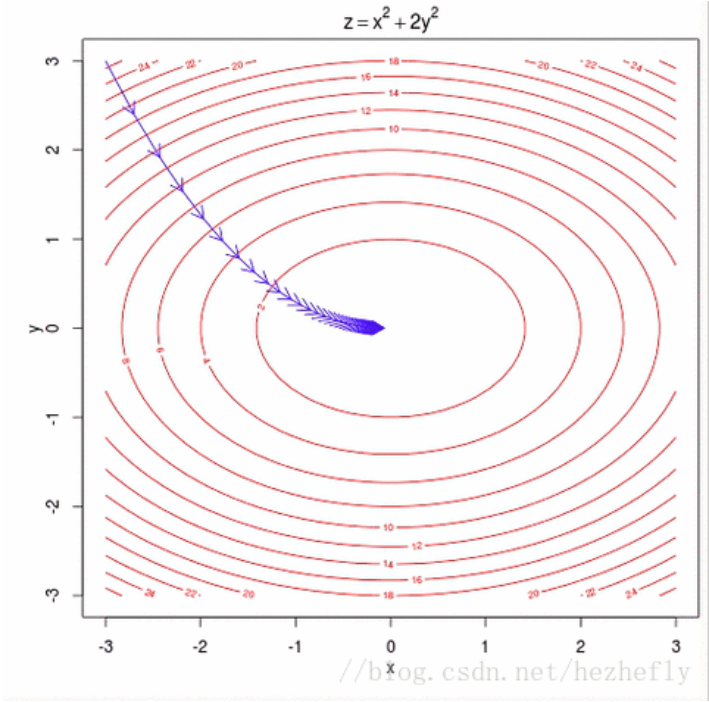
- 从P0点出发，应该向梯度的负方向移动



$$P_1 = P_0 - \eta \cdot \nabla f(x)$$

梯度下降法





梯度下降法

- 一元函数的**Taylor**展开公式：

$$f(x)=f(a)+\frac{f'(a)}{1!}(x-a)+\frac{f^{(2)}(a)}{2!}(x-a)^2+\cdots+\frac{f^{(n)}(a)}{n!}(x-a)^n+\frac{f^{(n+1)}(\theta)}{(n+1)!}(x-a)^{(n+1)}$$

- 考虑用**梯度**信息来设计迭代公式，忽略二次及更高项

$$f(x+\Delta x)=f(x)+(\nabla f(x))^{\mathrm{T}}\Delta x+o(\Delta x)$$

$$\Rightarrow \quad f(x+\Delta x)-f(x)\approx(\nabla f(x))^{\mathrm{T}}\Delta x$$

梯度下降法

- 梯度下降（**Gradient Descent**）：
  - $f(x+\Delta x) < f(x) \Rightarrow (\nabla f(x))^{\mathrm{T}} \Delta x < 0$

$$\Rightarrow (\nabla f(x))^{\mathrm{T}} \Delta x = \|\nabla f(x)\| \cdot \|\Delta x\| \cdot \cos\theta < 0$$

$$\Rightarrow \cos\theta < 0$$

  - 一般地，我们取$\Delta x = -t \cdot \nabla f(x)$，其中$t > 0$为步长
  - 迭代公式：

$$x_{k+1} = x_k - t \cdot \nabla f(x_k)$$

梯度下降法：视频

Machine Learning

Linear regression with one variable

Gradient descent

最优化问题：local vs global

- 最优化问题：在一定的约束条件下，求一个函数的**最大（小）值**

最优化问题

- 一个优化问题的**全局最小值点**x^*是指对于可行域里所有的x，有：

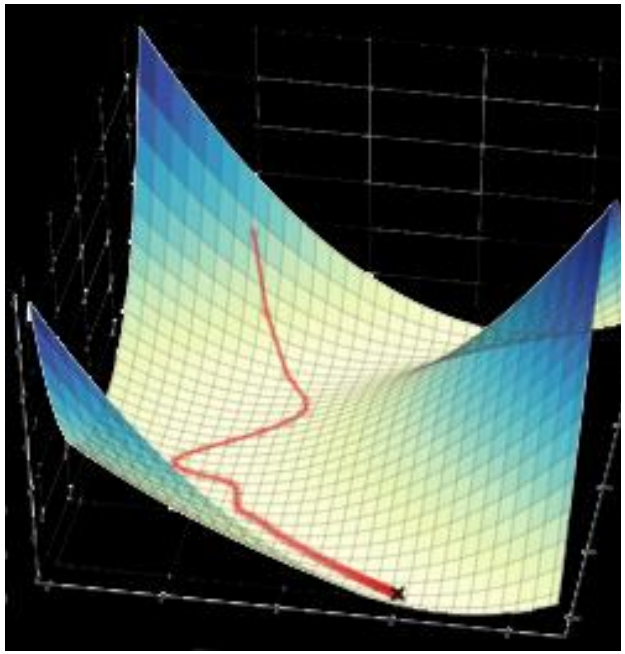
$$f(x^*) \leq f(x)$$

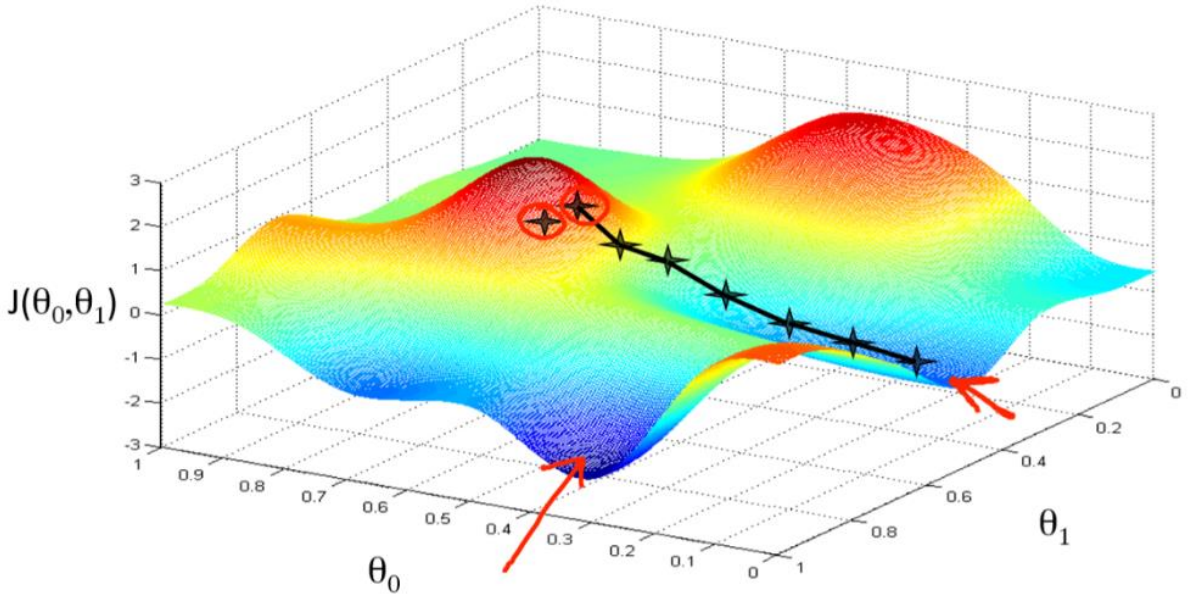
- **局部最小值点**$x^{\#}$的定义：对$x^{\#}$，存在一个δ邻域（$\|x - x^{\#}\| \leq \delta$），使得其中的所有$x$，有：

$$f(x^{\#}) \leq f(x)$$

- 一般而言，我们的目标是找到全局最小值。但是， 有些复杂的目标函数有多个局部最小值点。需要比较这些点处的目标函数值。

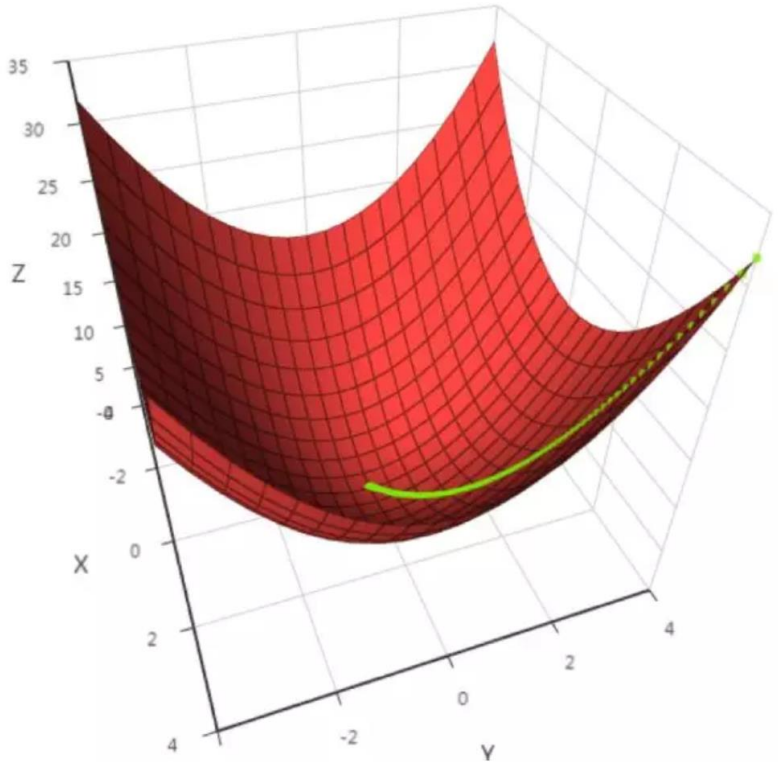
梯度下降法：local and global minimum





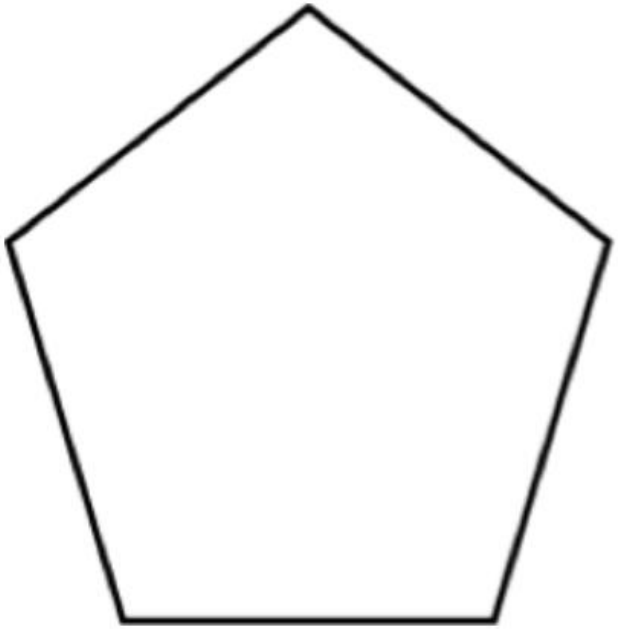
凸优化问题

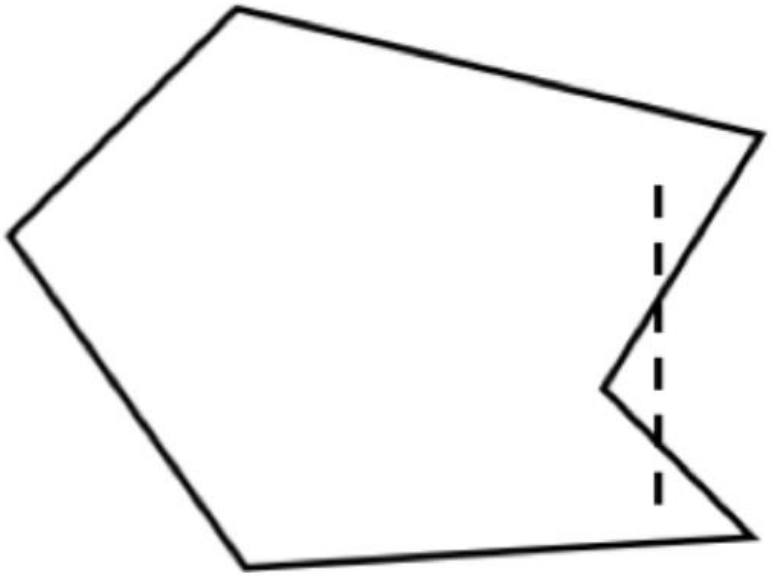
- 凸优化问题：同时满足凸函数+凸集的最优化问题
  - 局部最优解一定是全局最优解



凸优化

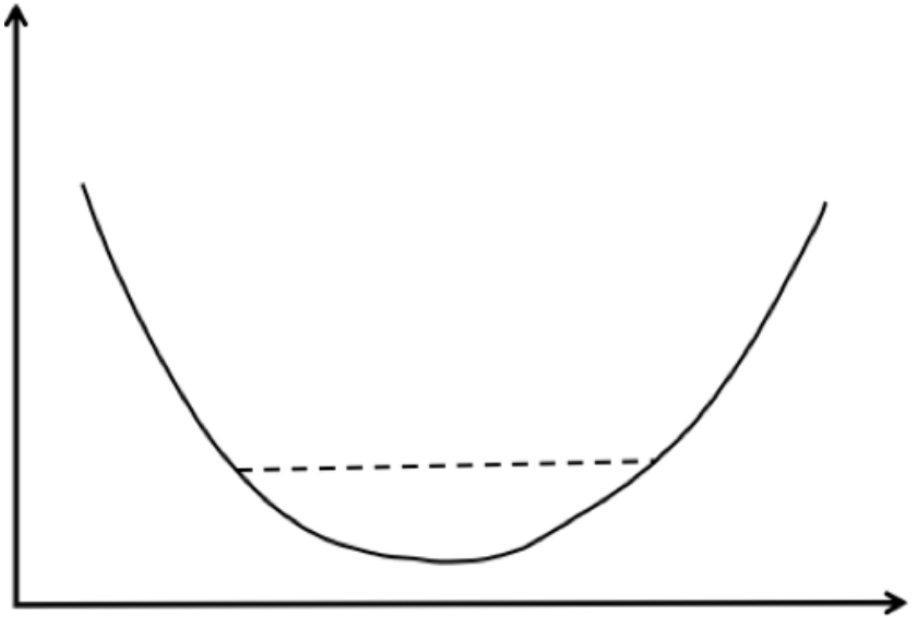
- 对于**n**维空间中点的集合C，如果对集合中的任意两点x和y，以及实数$0 \leq \theta \leq 1$，都有$\theta x + (1-\theta)y \in C$，则称该集合为凸集
  - $\theta x + (1-\theta)y$：集合的凸组合点





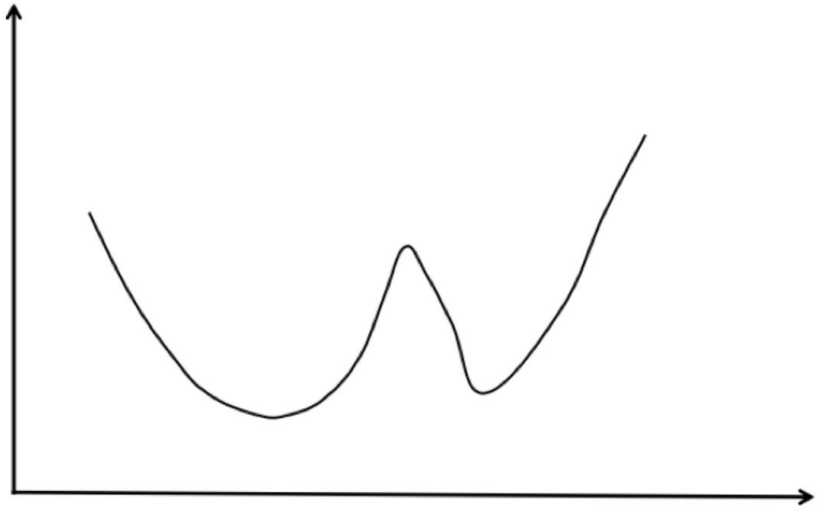
凸函数

- 对于**n**维空间中点的集合C，如果对集合中的任意两点x和y，以及实数$0 \leq \theta \leq 1$，都有$f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta)f(y)$，则该函数为凸函数
  - 函数在任何点处的切线都位于函数的下方

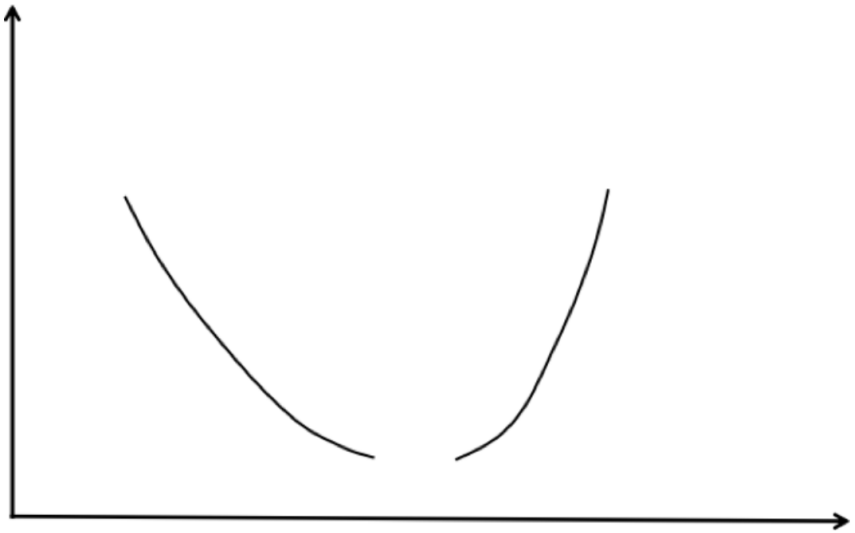


凸优化问题

- 之所以凸优化问题的定义要求目标函数是凸函数而且优化变量的可行域是凸集，是因为缺其中任何一个条件都不能保证局部最优解是全局最优解。以下是两个反例：



情况**1**：可行域是凸集（实数集），函数不是凸函数

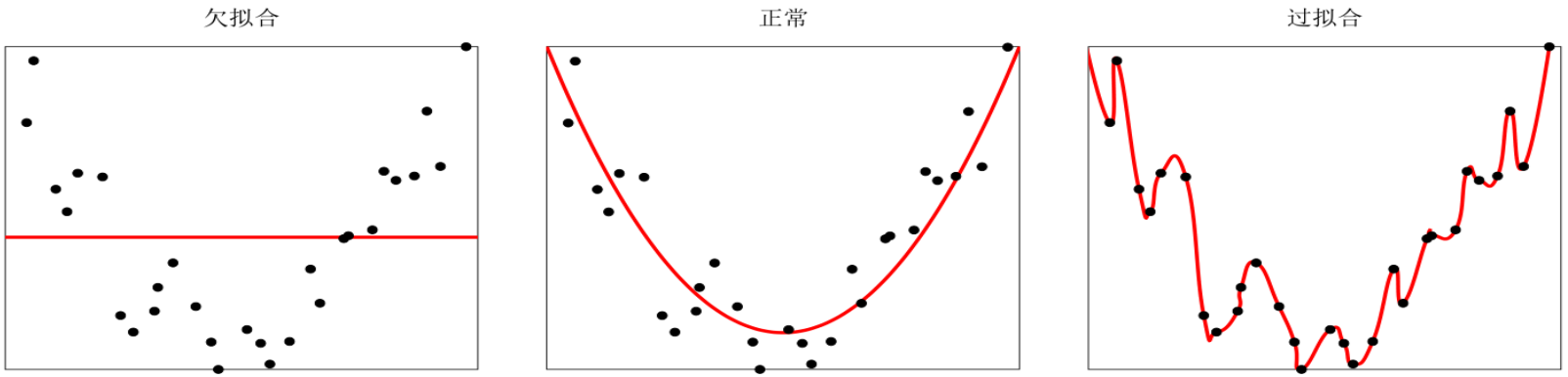


情况2：可行域不是凸集，函数是凸函数

04 过拟合问题

机器学习 = 优化?

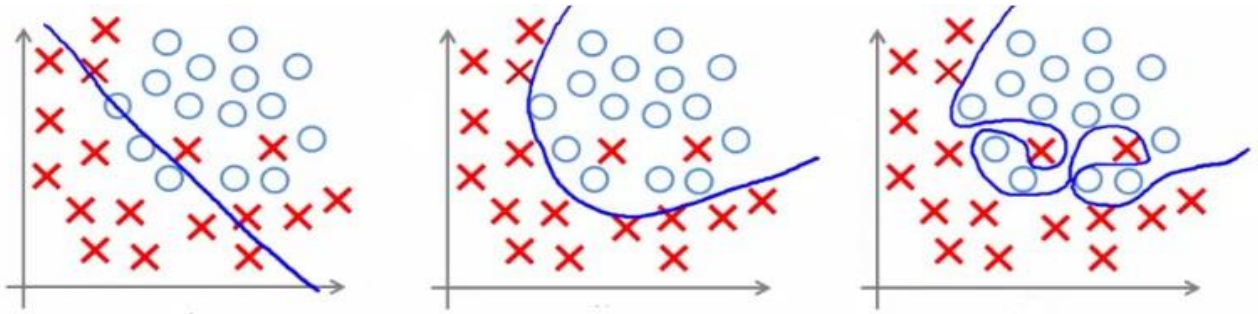
机器学习 = 优化？ NO！



过拟合：经验风险最小化原则很容易导致模型在训练集上错误率很低，但是在未知数据上错误率很高。

过拟合

- 过拟合：经验风险最小化原则很容易导致模型在训练集上错误率很低，但是在未知数据上错误率很高
  - 过拟合问题往往是由于训练数据少等原因造成的



泛化错误

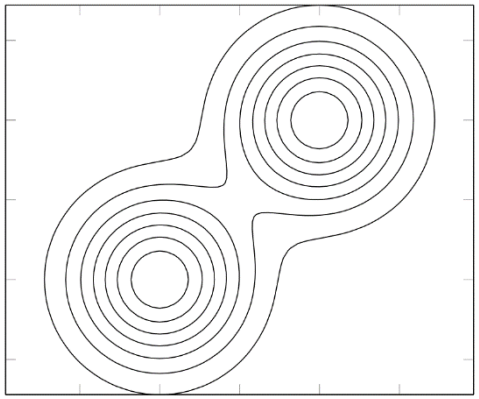
期望风险

$$\mathcal{R}(f) = \mathbb{E}_{(\mathbf{x},y)\sim p(\mathbf{x},y)}[\mathcal{L}(f(\mathbf{x}), y)],$$

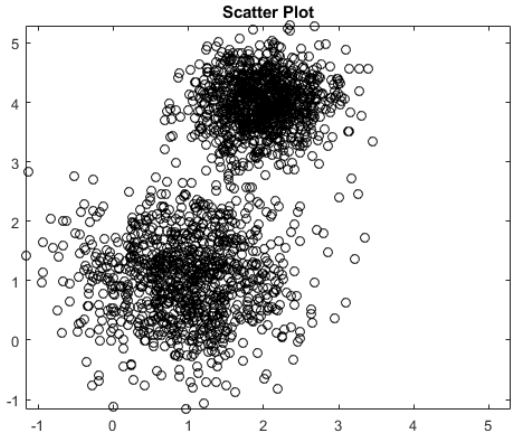
经验风险

$$\mathcal{R}_{\mathcal{D}}^{emp}(\theta) = \frac{1}{N}\sum_{n=1}^{N}\mathcal{L}(y^{(n)}, f(x^{(n)}, \theta))$$

真实分布 p_r



$\neq$



$$\mathcal{G}_{\mathcal{D}}(f) = \mathcal{R}(f) - \mathcal{R}_{\mathcal{D}}^{emp}(f)$$

泛化误差

如何减少泛化错误?

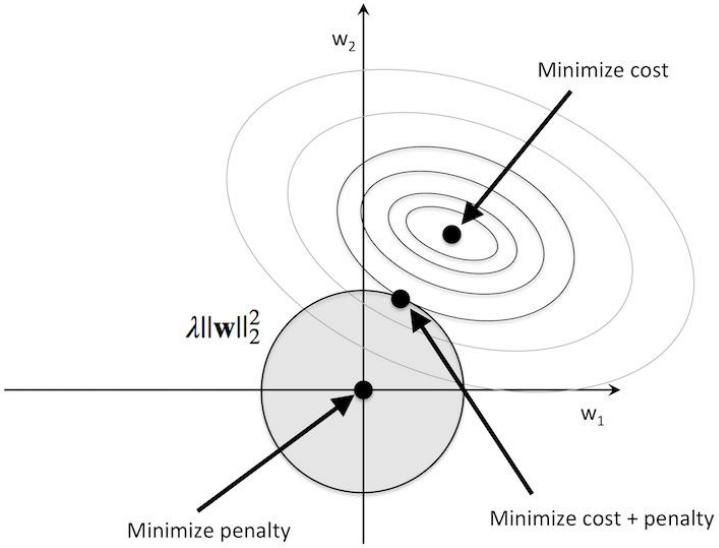
优化
经验风险最小

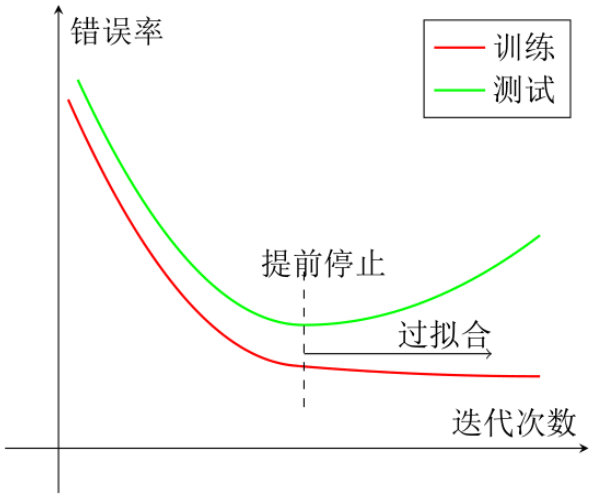
正则化
降低模型复杂度

正则化（Regularization）

所有损害优化的方法都是正则化

增加优化约束	干扰优化过程
L1/L2约束、数据增强	权重衰减、随机梯度下降、提前停止

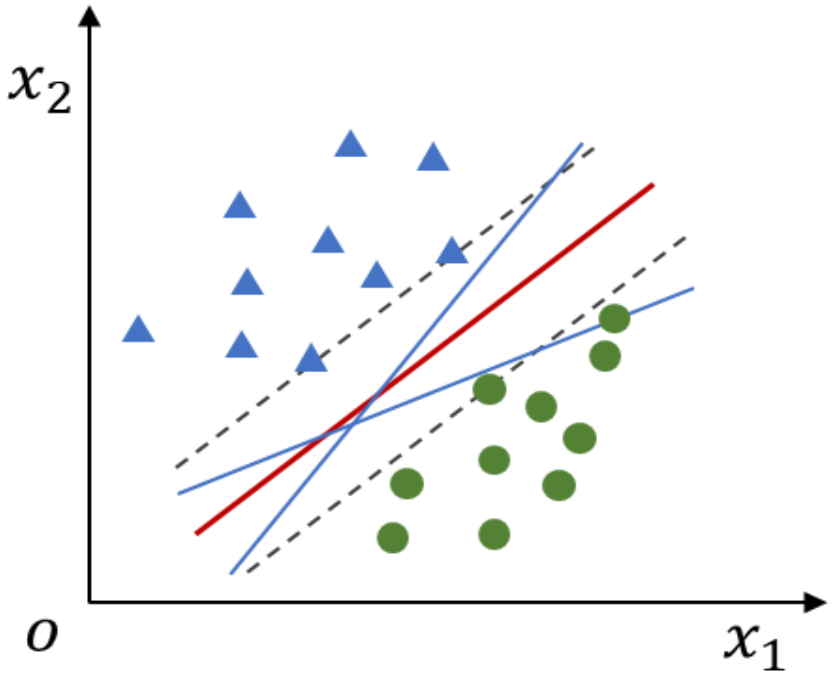




05 支持向量机

支持向量机

- 线性可分情况
  - 给定训练样本集 $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$，分类学习最基本的思想是基于训练数据集D在样本空间中找到一个划分超平面，将不同类别的样本分开



当能将训练样本分开的划分超平面可能有很多个，应该如何选择？

支持向量机

- 分类超平面

$$\mathbf{w}^{\mathrm{T}}\mathbf{x}+b=0,$$

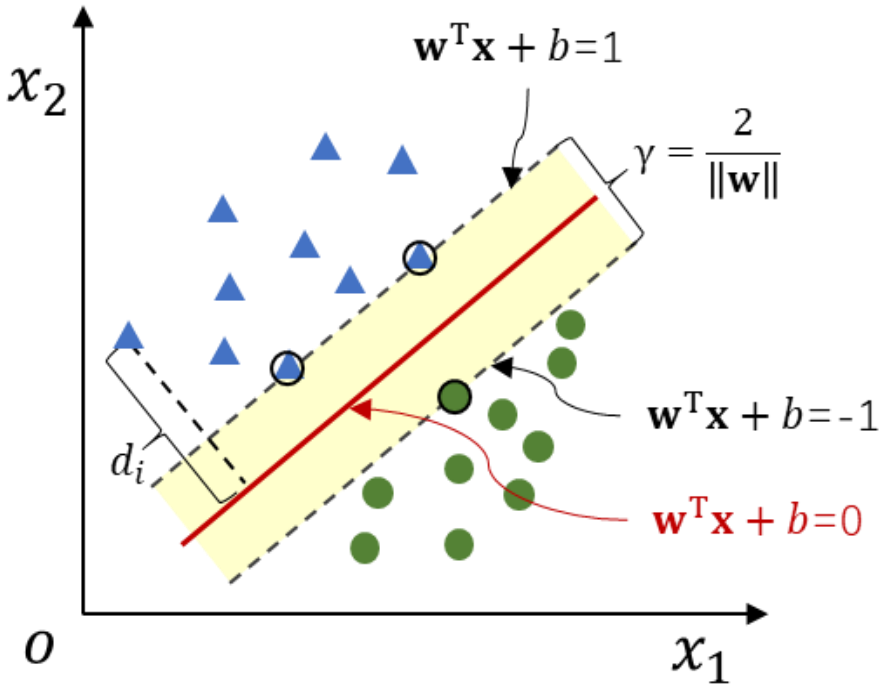
- 样本到超平面距离

$$d_i=\frac{\left|\mathbf{w}^{\mathrm{T}}\mathbf{x}_i+b\right|}{\|\mathbf{w}\|}.$$

- 最大几何间隔

$$\gamma=\min_i\left\{\frac{\left|\mathbf{w}^{\mathrm{T}}\mathbf{x}_i+b\right|}{\|\mathbf{w}\|}\right\}.$$

$$\gamma=\min_i\left\{\frac{y_i\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i+b\right)}{\|\mathbf{w}\|}\right\}.$$

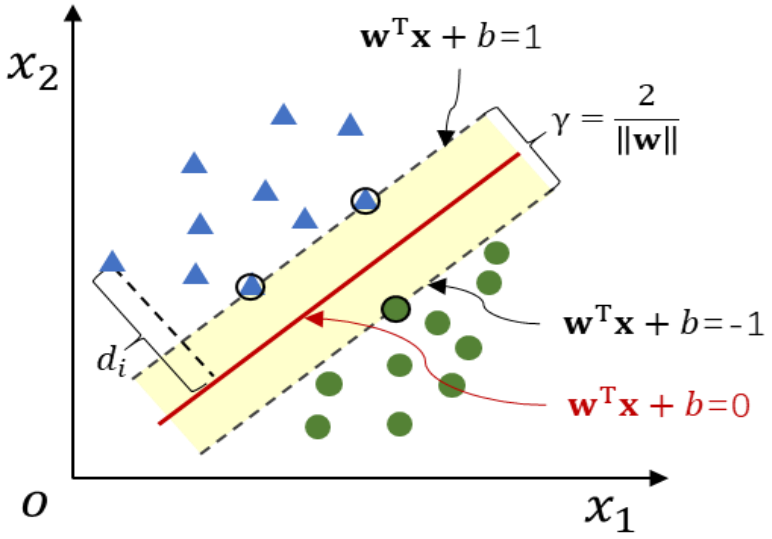


支持向量机

■最大化几何间隔

$$\max_{\mathbf{w},b}\left\{\min_{i}\left\{\frac{y_i(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i+b)}{\|\mathbf{w}\|}\right\}\right\}.$$

$$\begin{aligned}\max_{\mathbf{w},b}\quad & \frac{y_k(\mathbf{w}^{\mathrm{T}}\mathbf{x}_k+b)}{\|\mathbf{w}\|}\\ s.t.\quad & y_i(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i+b)\geq y_k(\mathbf{w}^{\mathrm{T}}\mathbf{x}_k+b),\end{aligned}$$



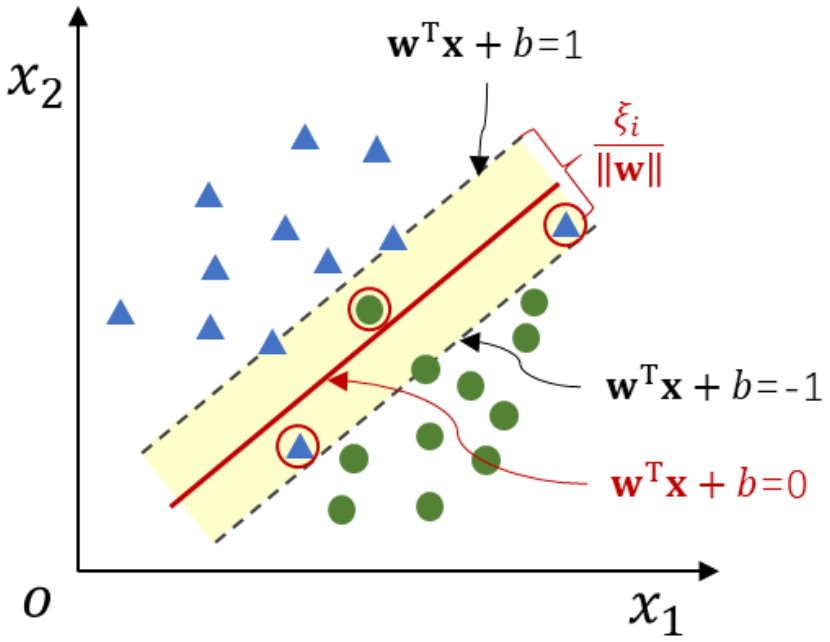
■变量代换，乘以2

$$\begin{aligned}\max_{\mathbf{w},b}\quad & \frac{2}{\|\mathbf{w}\|}\\ s.t.\quad & y_i(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i+b)\geq 1\end{aligned}$$

$$\Rightarrow\quad \begin{aligned}\min_{\mathbf{w},b}\quad & \frac{1}{2}\|\mathbf{w}\|^2\\ s.t.\quad & y_i(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i+b)\geq 1\end{aligned}$$

支持向量机：软间隔

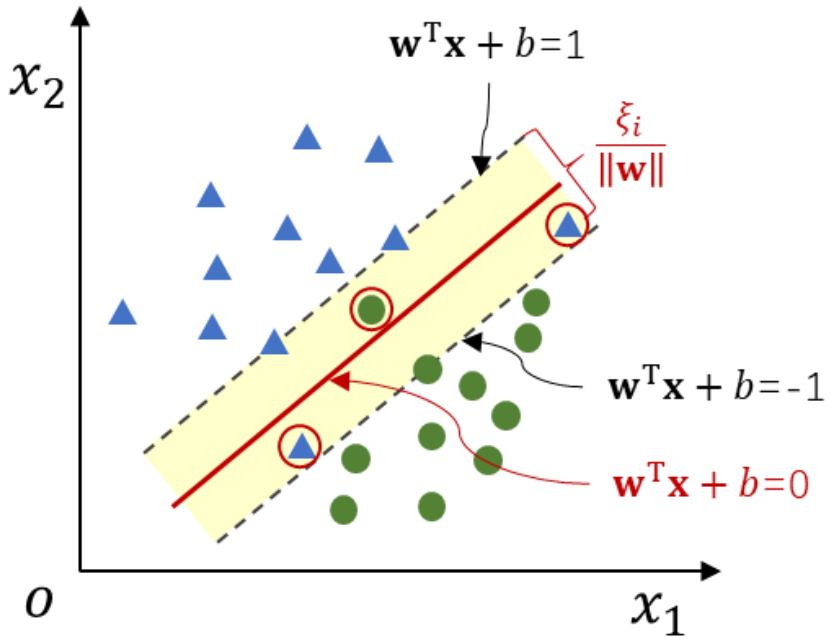
- 线性不可分怎么办？



支持向量机：软间隔

- 线性不可分怎么办？
  - 允许错误的出现

$$\begin{aligned} \min_{\mathbf{w},b} \quad & \frac{1}{2}\|\mathbf{w}\|^2 \\ s.t. \quad & y_i(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b) \geq 1 \end{aligned}$$



$$\begin{aligned} \min_{\mathbf{w},b,\xi_i} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i \\ \text{s.t.} \quad & y_i(\mathbf{w}'\mathbf{x}_i + b) \geq 1 - \xi_i \\ & \xi_i \geq 0, \end{aligned}$$

SVM的凸优化问题

- Primal

$$\begin{aligned} \min_{\mathbf{w},b,\xi_i} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i \\ \text{s.t.} \quad & y_i(\mathbf{w}'\mathbf{x}_i + b) \geq 1 - \xi_i \\ & \xi_i \geq 0, \end{aligned}$$

- Dual

$$\begin{aligned} \max_{\alpha_i} \quad & \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_i\sum_j y_i y_j \alpha_i \alpha_j \mathbf{x}_i^T \mathbf{x}_j \\ \text{s.t.} \quad & \sum_{i=1}^{n}\alpha_i y_i = 0, \\ & 0 \leq \alpha_i \leq C, \end{aligned}$$

约束优化问题：拉格朗日对偶问题

- 针对一个约束优化问题

$$\begin{aligned} \min_{\mathbf{x}} \quad & f(\mathbf{x}) \\ \text{s.t.} \quad & g_i(\mathbf{x}) \leq 0 \quad i = 1, \ldots, m \end{aligned}$$

- 拉格朗日函数（引入拉格朗日乘子）

$$\mathcal{L} = f(\mathbf{x}) + \sum_{i=1} \lambda_i g_i(\mathbf{x})$$

- 问题转化为非约束优化问题

$$\max_{\boldsymbol{\lambda} \geq \mathbf{0}} \min_{\mathbf{x}} f(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i g_i(\mathbf{x})$$

约束优化问题：拉格朗日对偶问题

- 针对一个约束优化问题

$$\begin{aligned} \min_{\mathbf{x}} \quad & f(\mathbf{x}) \\ \text{s.t.} \quad & g_i(\mathbf{x}) \le 0 \quad i = 1, \dots, m \end{aligned}$$

- 假设一个下界v使得下面方程无解

$$\begin{aligned} & f(\mathbf{x}) < v \\ & g_i(\mathbf{x}) \le 0, i = 1, 2, \dots, m \end{aligned} \quad (2) \quad \Longleftrightarrow \quad f(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i g_i(\mathbf{x}) < v \quad (3)$$

注意到方程组(2)有解可以推出对于任意的$\boldsymbol{\lambda} \ge \mathbf{0}$, 方程(3)有解；根据逆否命题，方程组(2)无解的充分条件是存在$\boldsymbol{\lambda} \ge \mathbf{0}$，让方程(3)无解。方程(3)无解的充要条件是

$$\min_{\mathbf{x}} f(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i g_i(\mathbf{x}) \ge v \quad (4)$$

因为要找最好的下界，所以$v, \boldsymbol{\lambda}$应该取：

$$v = \max_{\boldsymbol{\lambda} \ge \mathbf{0}} \min_{\mathbf{x}} f(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i g_i(\mathbf{x}) \quad (5)$$

SVM的凸优化问题

- 拉格朗日乘子法

$$\max_{\alpha_i} \min_{\mathbf{w},b,\xi_i} \quad \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i - \sum_i \alpha_i(y_i(\mathbf{w}'\mathbf{x}_i + b) - 1 + \xi_i) - \sum_i \eta_i\xi_i$$

$$\begin{aligned} \min_{\mathbf{w},b,\xi_i} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i \\ \text{s.t.} \quad & y_i(\mathbf{w}'\mathbf{x}_i + b) \geq 1 - \xi_i \\ & \xi_i \geq 0, \end{aligned}$$

SVM的凸优化问题

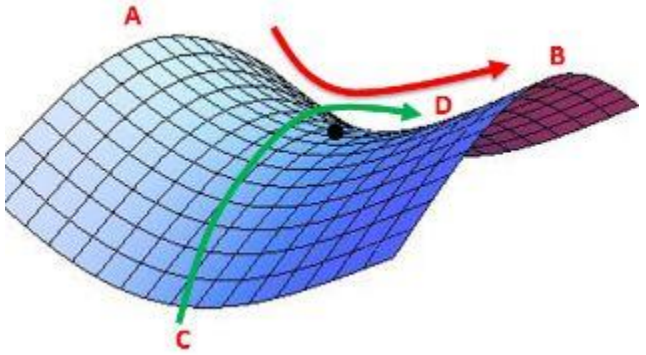
■拉格朗日乘子法

$$\max_{\alpha_i} \min_{\mathbf{w},b,\xi_i} \quad \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i - \sum_i \alpha_i(y_i(\mathbf{w}'\mathbf{x}_i + b) - 1 + \xi_i) - \sum_i \eta_i\xi_i$$

$$\mathbf{0} = \frac{\partial \mathcal{L}}{\partial \mathbf{w}} = \mathbf{w} - \sum_i \alpha_i y_i \mathbf{x}_i$$

$$0 = \frac{\partial \mathcal{L}}{\partial b} = \sum_i \alpha_i y_i$$

$$0 = \frac{\partial \mathcal{L}}{\partial \xi_i} = C - \alpha_i - \eta_i$$



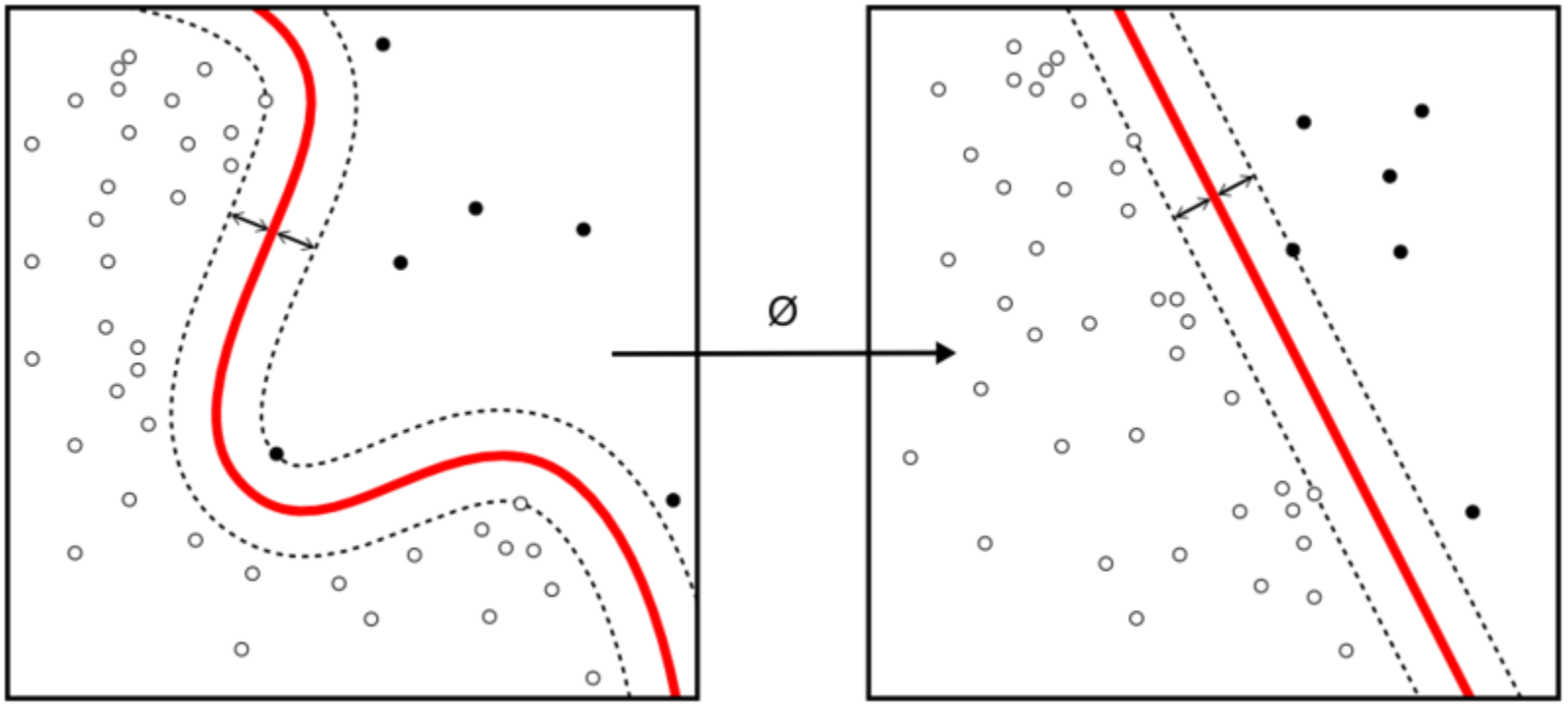
$$\max_{\alpha_i} \quad \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_i\sum_j y_i y_j \alpha_i \alpha_j \mathbf{x}_i^T \mathbf{x}_j$$

$$\text{s.t.} \quad \sum_{i=1}^{n}\alpha_i y_i = 0,$$

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

支持向量机

- 非线性情况



支持向量机

- 非线性情况

$$\begin{aligned} \min_{\mathbf{w},b,\xi_i} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i \\ \text{s.t.} \quad & y_i(\mathbf{w}'\phi(\mathbf{x}_i) + b) \geq 1 - \xi_i \\ & \xi_i \geq 0, \end{aligned}$$

- Dual

支持向量机

- 非线性情况

$$\begin{aligned} \min_{\mathbf{w},b,\xi_i} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i \\ \text{s.t.} \quad & y_i(\mathbf{w}'\phi(\mathbf{x}_i) + b) \geq 1 - \xi_i \\ & \xi_i \geq 0, \end{aligned}$$

- Dual

$$\begin{aligned} \max_{\boldsymbol{\alpha}} \quad & \boldsymbol{\alpha}'\mathbf{1} - \frac{1}{2}(\boldsymbol{\alpha}\cdot\mathbf{y})'\mathbf{K}(\boldsymbol{\alpha}\cdot\mathbf{y}) \\ \text{s.t.} \quad & \boldsymbol{\alpha}'\mathbf{y} = 0, \\ & \mathbf{0} \leq \boldsymbol{\alpha} \leq C\mathbf{1}, \end{aligned}$$

支持向量机

- 正则项⇔最大化间隔

$$\begin{aligned} \min_{\mathbf{w},b,\xi_i} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i \\ \text{s.t.} \quad & y_i(\mathbf{w}'\phi(\mathbf{x}_i) + b) \geq 1 - \xi_i \\ & \xi_i \geq 0, \end{aligned}$$

- Probably Approximated Correct 泛化误差
  - with probability of $1-\delta$

$$\mathcal{R}(f) \leq \mathcal{R}^{emp}(f) + \Delta(\delta, \|\mathbf{w}\|)$$

J. Shawe-Taylor, and N. Cristianini (2004). Kernel Methods for Patter Recognition. Cambriidge.

V. Vapnik (2000). The Nature of Statistical Learning Theory. Springer.

回顾

- 机器学习
  - 模拟人“学习”的能力
  - 从经验（数据）来学习，提高任务上的性能
- 三要素
  - 数据、模型、损失函数
- 主要内容
  - 经验误差最小化
  - 过拟合（模型复杂度）
  - 正则化