Deep Learning I

3. 线性神经网络

WU Xiaokun 吴晓堃

xkun.wu [at] gmail

2021/03/07

线性回归

什么是回归?

回归 regression

- Galton 发现父亲的身高和儿子的身高之间存在着某种给定的关系
  - “子辈的平均身高是其父辈及其所处族群平均身高的加权平均和”。
  - 差异性? 从整个人群上来看，父亲和孩子的身高分布是很相近的。

什么是回归?

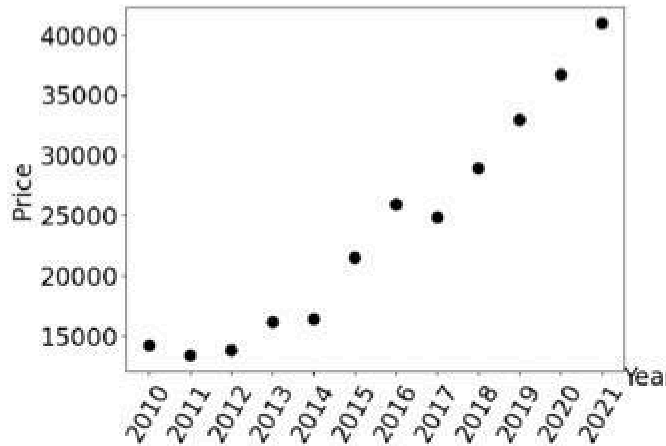
回归 regression

- Galton 发现父亲的身高和儿子的身高之间存在着某种给定的关系
  - “子辈的平均身高是其父辈及其所处族群平均身高的加权平均和”。
  - 差异性? 从整个人群上来看，父亲和孩子的身高分布是很相近的。

- 统计数据**驱向均值**的现象
  - `re-`：重复；`gress`：行走；`-sion`：名词化
  - 多次模拟考试成绩的均值可以作为真实水准的依据

为什么讨论回归?

回归分析：对自变量和因变量之间的关系建模



- 用于预测：源于**对趋势（主观判断）的信念**
  - 统计学信条："相关性并不意味着因果关系"
  - 推断：人脑对客观规律的简化描述

线性假设

购房时需要对其价值进行评估

- 找出几个**关键因素**，如面积，地段，存款等
- 做最简单的线性假设：固定比率变化
  - 人脑不善于做非线性计算

线性假设

购房时需要对其价值进行评估

- 找出几个**关键因素**，如面积，地段，存款等
- 做最简单的线性假设：固定比率变化
  - 人脑不善于做非线性计算

线性模型：$y = w_1x_1 + w_2x_2 + .. + w_dx_d + b$

- 任务：预测房价
  - 特征：自变量，相关的一些因素
  - 权重：对应自变量的影响力
  - 偏置：基准值，如当年均值

≡

线性回归问题

假如能够确定参数$\mathbf{w}, b$:

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

线性回归问题

假如能够确定参数$\mathbf{w}, b$:

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

问题转化：如何确定参数$\mathbf{w}, b$?

线性回归问题

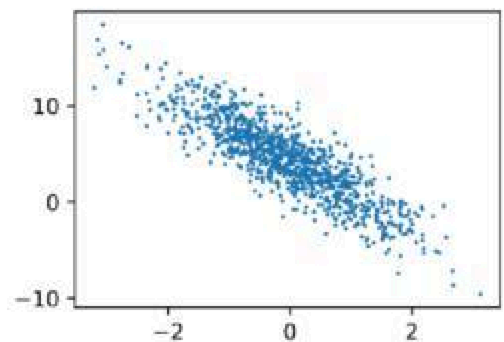
假如能够确定参数$\mathbf{w}, b$:

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

问题转化：如何确定参数$\mathbf{w}, b$?

基本假设:

- 自变量和因变量线性相关
- 噪音符合正态分布



数据驱动

收集样本数据

- **样本**：可观测的数据对$(\mathbf{x}, y)$
  - **样本-特征存储（矩阵）**：每一行是一个样本

$$\mathbf{y} = \mathbf{X}\mathbf{w} + b, \mathbf{X} \in \mathbb{R}^{N \times d}$$

数据驱动

收集样本数据

- **样本**：可观测的数据对$(\mathbf{x}, y)$
  - **样本-特征存储（矩阵）**：每一行是一个样本

$$\mathbf{y} = \mathbf{X}\mathbf{w} + b, \mathbf{X} \in \mathbb{R}^{N \times d}$$

两种求解思路：

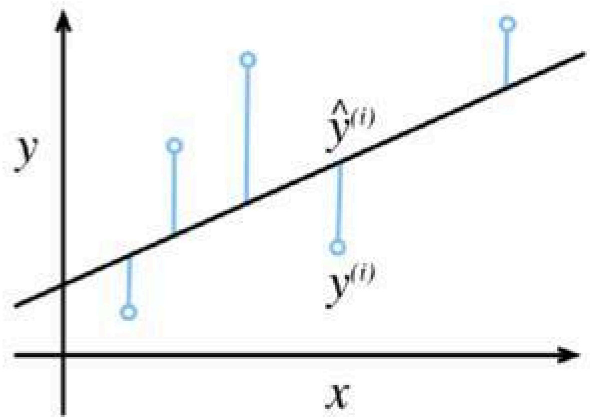
- **显式解**：$\bar{\mathbf{w}}^* = (\bar{\mathbf{X}}^T \bar{\mathbf{X}})^{-1} \bar{\mathbf{X}}^T \mathbf{y}$
  - 只适用于有解析解的简单问题
  - 满秩：数据之间不能有相关性
- **数值优化**

优化问题定义

定义损失函数，如**平方损失**：$l(y, \hat{y}) = \frac{1}{2}(y - \hat{y})^2$

- 但不可能每个点都最优："按下葫芦浮起瓢"

优化目标：$\mathcal{L}(\mathbf{X}, \mathbf{y}; \mathbf{w}, b) = \frac{1}{2N}\|\mathbf{y} - \mathbf{X}\mathbf{w} - b\|^2$

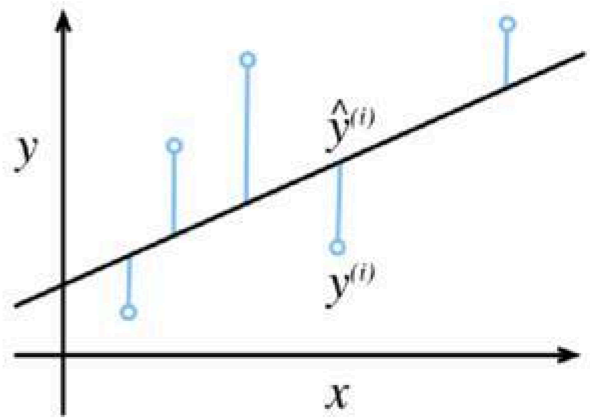


优化问题定义

定义损失函数，如**平方损失**：$l(y,\hat{y}) = \frac{1}{2}(y-\hat{y})^2$

- 但不可能每个点都最优："按下葫芦浮起瓢"

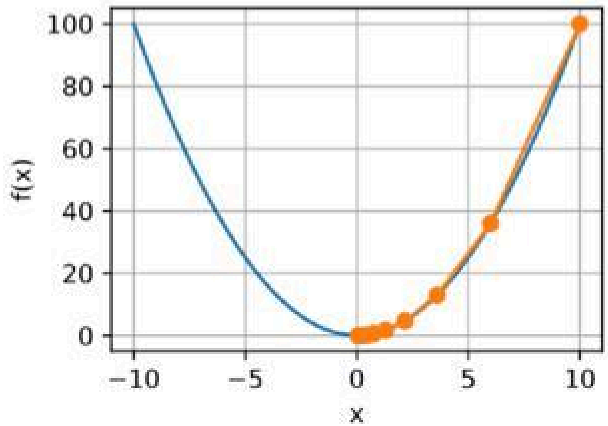
优化目标：$\mathcal{L}(\mathbf{X},\mathbf{y};\mathbf{w},b) = \frac{1}{2N}\|\mathbf{y}-\mathbf{X}\mathbf{w}-b\|^2$



优化问题：$\mathbf{w}^*, b^* = \arg\min_{\mathbf{w},b} \mathcal{L}(\mathbf{X},\mathbf{y};\mathbf{w},b)$

梯度下降法求解

1. 初始化参数：$\mathbf{w} = \mathbf{w_0}$
2. 循环：
   1. 计算当前梯度：$\frac{\partial \mathcal{L}}{\partial \mathbf{w}}|_{\mathbf{w}}$
   2. 更新参数：$\mathbf{w} = \mathbf{w} - \eta \frac{\partial \mathcal{L}}{\partial \mathbf{w}}|_{\mathbf{w}}$



小批量随机梯度下降

数据量太大怎么办?

- 计算复杂度：损失值、梯度值都是$O(n)$
- 空间复杂度：内存、显存放不下

小批量随机梯度下降

数据量太大怎么办?

- 计算复杂度：损失值、梯度值都是$O(n)$
- 空间复杂度：内存、显存放不下

随机梯度下降（SGD）：每次迭代对训练数据随机均匀采样

- 近似计算梯度：$\widehat{\nabla f} = \frac{1}{B}\sum_{i=1}^{B}\nabla f_i$
- B称为批量大小

小批量随机梯度下降

数据量太大怎么办?

- 计算复杂度：损失值、梯度值都是$O(n)$
- 空间复杂度：内存、显存放不下

随机梯度下降（SGD）：每次迭代对训练数据随机均匀采样

- 近似计算梯度：$\widehat{\nabla f} = \frac{1}{B}\sum_{i=1}^{B} \nabla f_i$
- B称为批量大小

思考：随机梯度是对完整梯度的**无偏估计**

- 提示：计算数学期望

选择批量大小

不能太大

- 内存消耗增加
- 容易陷入**极小值**：批量之间的差异不足
  - 极小值附近地形比较平坦
- 最大可以和数据集大小相同

选择批量大小

不能太大

- 内存消耗增加
- 容易陷入**极小值**：批量之间的差异不足
  - 极小值附近地形比较平坦
- 最大可以和数据集大小相同

不能太小

- 不利于**并行计算**
- 但在线应用时可能为1

选择批量大小

不能太大

- 内存消耗增加
- 容易陷入**极小值**：批量之间的差异不足
  - 极小值附近地形比较平坦
- 最大可以和数据集大小相同

不能太小

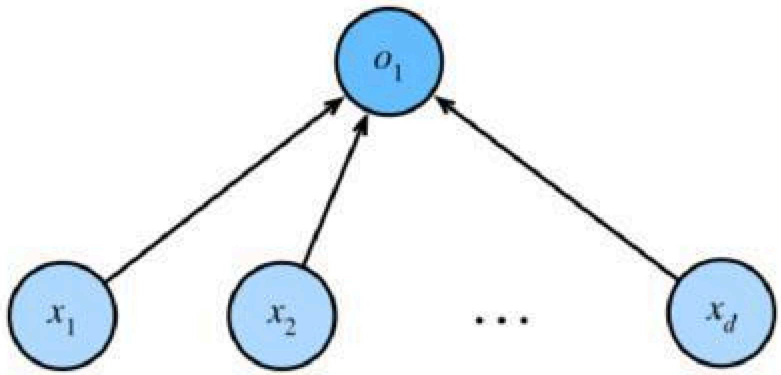
- 不利于**并行计算**
- 但在线应用时可能为1

确定批量大小：主要靠**经验估计**

- 显存能够容纳的最大值
- 二分搜索算法

线性回归可以看成单层神经网络

$$y = w_1x_1 + w_2x_2 + .. + w_dx_d + b$$



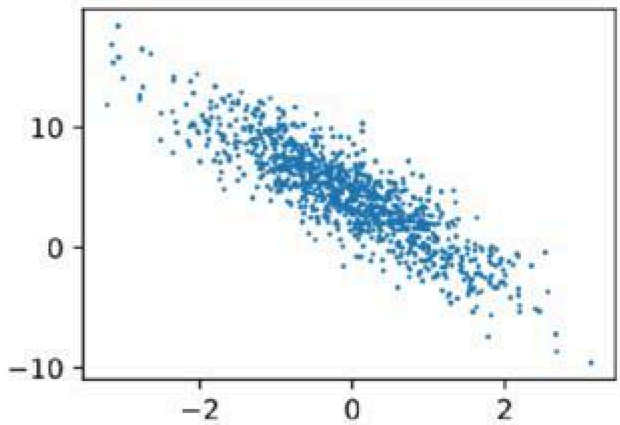
注意：不存在线性的多层神经网络！

实验：线性回归的从零开始实现

人造数据集

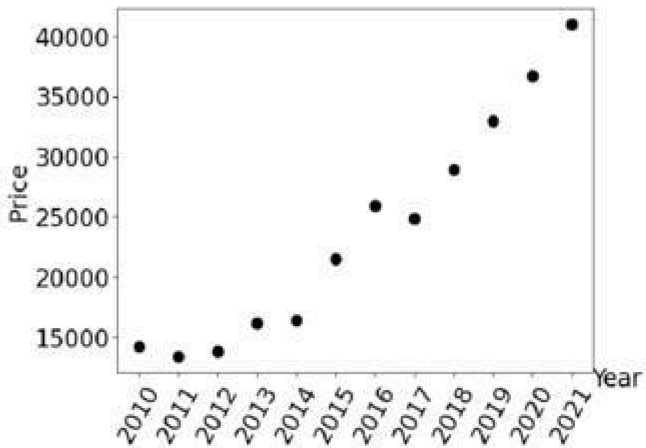
带噪音的人造数据

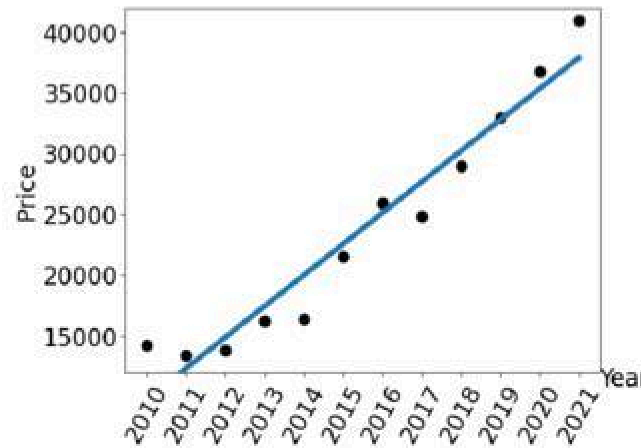
$$\mathbf{y} = \mathbf{X}\mathbf{w} + b + \epsilon$$



实验：线性回归的简洁实现

作业：线性回归模型


Price
40000
35000
30000
25000
20000
15000
Year
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021

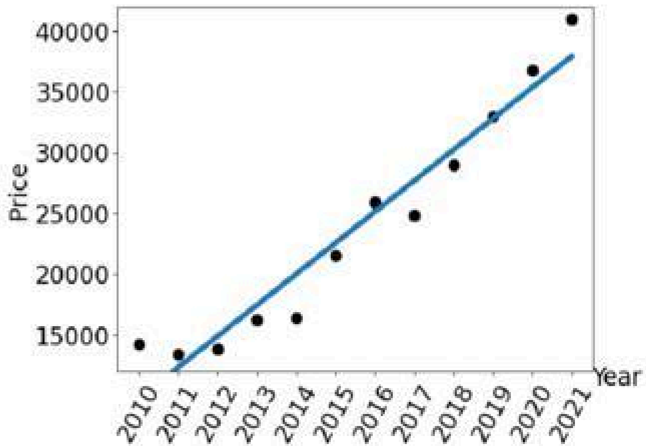

Price
40000
35000
30000
25000
20000
15000
Year
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021

Halftime

softmax回归

回归、分类

回归：输出连续数值



分类：输出离散类别

回归用于分类

根据**取值范围**分类

- 例如$[0, 10)$的坐标区间：$[0, 1)$代表第一个类

回归用于分类

根据**取值范围**分类

- 例如$[0,10)$的坐标区间：$[0,1)$代表第一个类

问题：不容易度量

- 目标：同类尽量聚在一起；**异类尽量分散开**
  - 边界附近很难区分

回归用于分类

根据**取值范围**分类

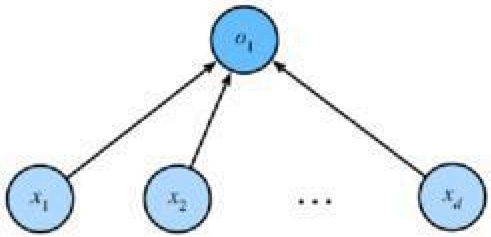
- 例如$[0, 10)$的坐标区间：$[0, 1)$代表第一个类

问题：不容易度量

- 目标：同类尽量聚在一起；**异类尽量分散开**
  - 边界附近很难区分
- 如何构造目标（损失）函数?
  - 连续值，却聚在几个点附近

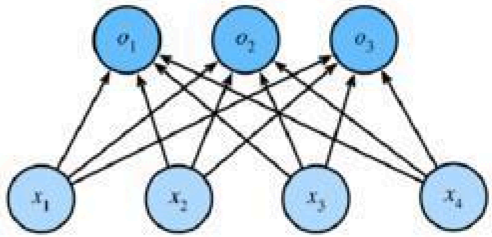
从回归到多类分类

回归



- 输出单个连续值
  - 输出：预测值
- 误差：输出与真实值的差别

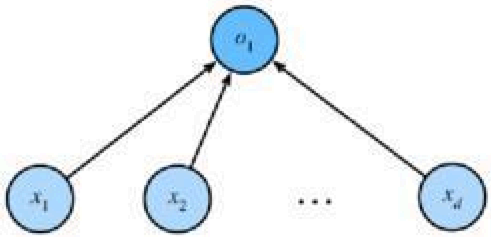
分类



- 输出多个不同的连续值
  - 每个输出：预测类别的**可信程度**
- 误差：输出与真实概率的差别

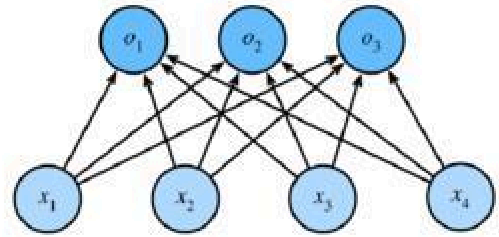
从回归到多类分类

回归



- 输出单个连续值
  - 输出：预测值
- 误差：输出与真实值的差别

分类



- 输出多个不同的连续值
  - 每个输出：预测类别的**可信程度**
- 误差：输出与真实概率的差别

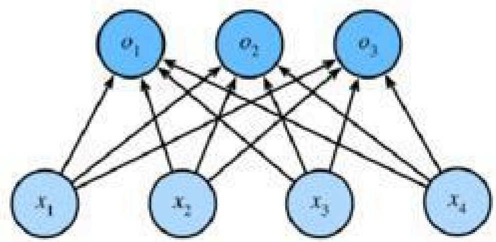
问题：

- 如何从输出的多个实数值构造概率？
- 如何表达“真实概率”？
- 如何计算概率之间的差别？

独热编码

独热编码 one-hot encoding：单有效位编码的向量

- $y = [y_1, y_2, .., y_C]$
- C个类；$y_c = 1$表示正确类别是c



例如数字3的编码：

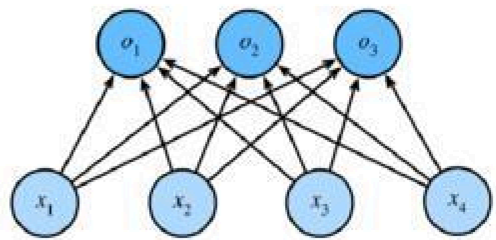
- $[0, 0, 0, 1, 0, 0, 0, 0, 0, 0]$
- $C = 10, c = 3$

softmax运算

输出向量：$\mathbf{o} = [o_1, o_2, .., o_C]^T$

- 需要计算每个分量的比重

$$\hat{\mathbf{y}} = \mathrm{softmax}(\mathbf{o}), \hat{y}_j = \frac{\exp(o_j)}{\sum_k \exp(o_k)}$$

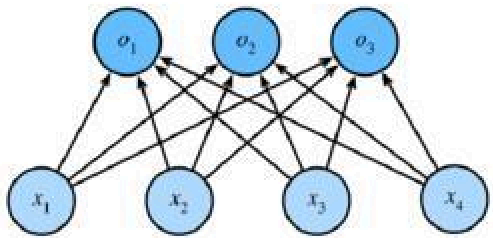


softmax运算

输出向量：$\mathbf{o} = [o_1, o_2, .., o_C]^T$

- 需要计算每个分量的比重

$$\hat{\mathbf{y}} = \text{softmax}(\mathbf{o}), \hat{y}_j = \frac{\exp(o_j)}{\sum_k \exp(o_k)}$$



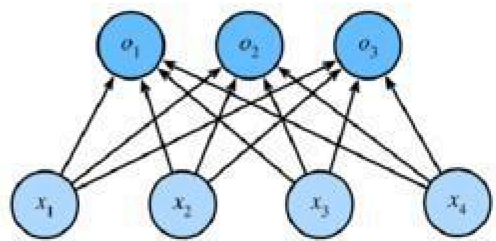
- softmax本质上是归一化
  - $0 \leq \hat{y}_j \leq 1$，并且 $\sum_j \hat{y}_j = 1$

小批量样本

假设批量大小为B，特征维度为d

$$\mathbf{O} = \mathbf{X}\mathbf{W} + \mathbf{b}$$
$$\hat{\mathbf{Y}} = \mathrm{softmax}(\mathbf{O})$$

- $\mathbf{X} \in \mathbb{R}^{B\times d}, \mathbf{W} \in \mathbb{R}^{d\times C}, \mathbf{b} \in \mathbb{R}^{1\times C}$

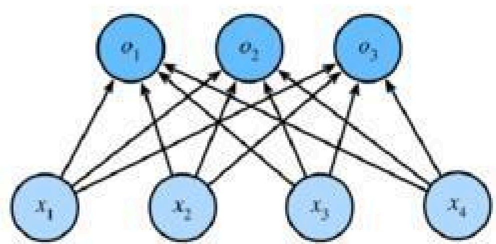


小批量样本

假设批量大小为B，特征维度为d

$$\mathbf{O} = \mathbf{X}\mathbf{W} + \mathbf{b}$$
$$\hat{\mathbf{Y}} = \mathrm{softmax}(\mathbf{O})$$

- $\mathbf{X} \in \mathbb{R}^{B\times d}, \mathbf{W} \in \mathbb{R}^{d\times C}, \mathbf{b} \in \mathbb{R}^{1\times C}$



问题：如何计算误差（损失值）？

度量概率相似程度

度量概率分布相似程度：采样点上的**重合度**

$$\|\mathbf{y}, \hat{\mathbf{y}}\| \propto \prod_c \hat{y}_c^{y_c}$$

- 唯一的$y_c = 1$对应的$\hat{y}_c$被选出来

度量概率相似程度

度量概率分布相似程度：采样点上的**重合度**

$$\|\mathbf{y}, \hat{\mathbf{y}}\| \propto \prod_c \hat{y}_c^{y_c}$$

- 唯一的$y_c = 1$对应的$\hat{y}_c$被选出来
- 数值上与内积相同

度量概率相似程度

度量概率分布相似程度：采样点上的**重合度**

$$\|\mathbf{y},\hat{\mathbf{y}}\| \propto \prod_c \hat{y}_c^{y_c}$$

- 唯一的$y_c = 1$对应的$\hat{y}_c$被选出来
- 数值上与**内积**相同

问题：大量概率值的乘积导致下溢?

对数似然

取负对数得到损失函数：

$$\begin{aligned} l(\mathbf{y}, \hat{\mathbf{y}}) &= -\log \prod_c \hat{y}_c^{y_c} \\ &= -\sum_c y_c \log \hat{y}_c \end{aligned}$$

- 常用数学技巧。
  - 注意：取对数**不改变序关系**
- 数值计算：加法总是比乘法好

信息论解释

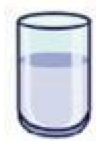
在信息论中，**熵 entropy**用于度量数据中的信息量

$$H(\mathbf{P}) = -\sum_i \mathbf{P}_i \log \mathbf{P}_i$$

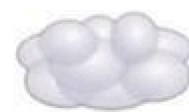
对从分布 $\mathbf{P}$ 中随机抽样的数据编码，至少需要 $H[\mathbf{P}]$ “**纳特 nat**”。

- $1 \text{ nat} = \log(2)^{-1} \approx 1.44 \text{ bit}$

Low



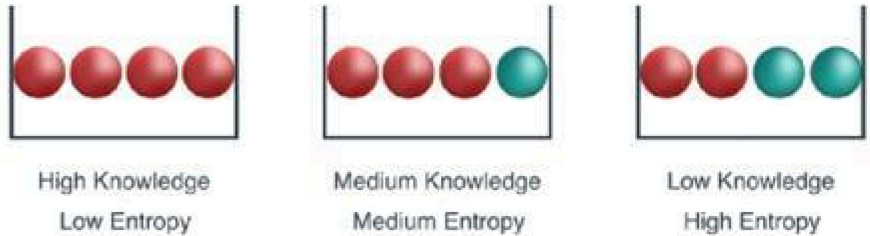
Medium



High

熵的理解

想象从盒子里抽奖



- 熵可以理解为：对数据流进行连续预测的**预期“惊异”**

交叉熵

交叉熵 cross entropy用于度量两个概率分布的差异

$$H(\mathbf{P},\mathbf{Q}) = -\sum_i P_i \log Q_i$$

- 即前面损失函数的定义
  - 损失函数：$-\sum_c y_c \log \hat{y}_c$
  - 预测是$\mathbf{Q}$，但实际是$\mathbf{P}$

交叉熵

交叉熵 cross entropy用于度量两个概率分布的差异

$$H(\mathbf{P}, \mathbf{Q}) = -\sum_i P_i \log Q_i$$

- 即前面损失函数的定义
  - 损失函数： $-\sum_c y_c \log \hat{y}_c$
  - 预测是$\mathbf{Q}$，但实际是$\mathbf{P}$

数学上漂亮：梯度是预测与真实概率的差值

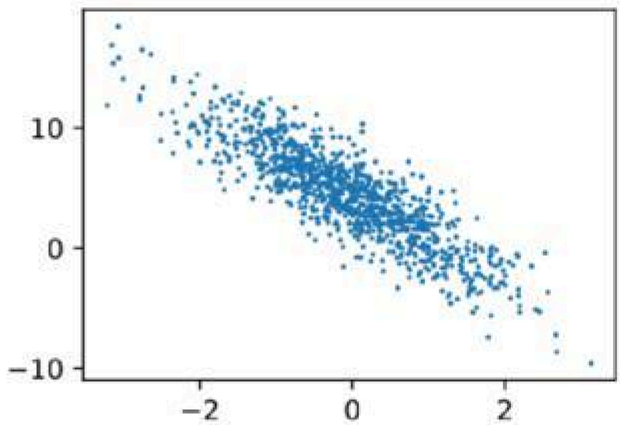
- $\frac{\partial}{\partial o_i} l(\mathbf{y}, \hat{\mathbf{y}}) = \mathrm{softmax}(\mathbf{o})_i - y_i = \hat{y}_i - y_i$

实验：图像分类数据集

实际工程问题不是数学题

解决工程问题的思路需要考虑实际情况

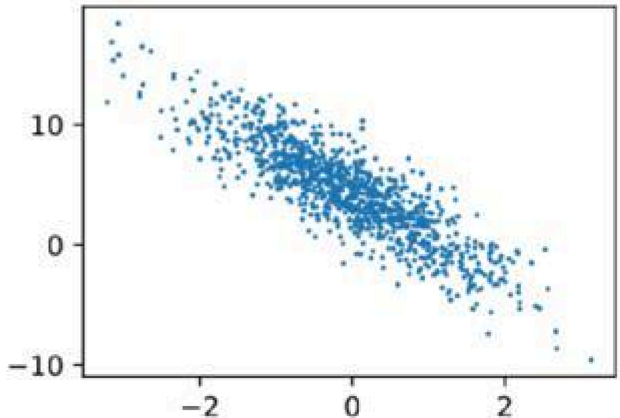
- 人造数据：训练出来的模型只能拟合给定的数据



实际工程问题不是数学题

解决工程问题的思路需要考虑实际情况

- 人造数据：训练出来的模型**只能拟合给定的数据**



优化目标：保证理论上参数对训练数据是最优的

- 实际应用：通常是**完全不同**的数据

训练集、测试集

测试集：实际应用时才能拿到的数据

- 对于训练过程来说是**未知**的

类比：训练集=日常习题；测试集=考试题

- 训练参数：通过习题掌握**解题思路**

训练集、测试集

测试集：实际应用时才能拿到的数据

- 对于训练过程来说是**未知**的

类比：训练集=日常习题；测试集=考试题

- 训练参数：通过习题掌握**解题思路**

模拟实际应用：将数据划分成两部分

- 对比：全部数据当成训练集

实验：softmax回归的从零开始实现

实验：softmax回归的简洁实现

课程项目：手写数字分类

MNIST 数据集

Review

本章内容

线性回归模型。实验：线性回归的实现。softmax回归模型。实验：softmax回归的实现

重点：线性模型；线性回归的实现；softmax模型；softmax回归的实现

难点：交叉熵。

学习目标

- 理解线性回归模型的一般形式、损失函数、优化算法。
- 掌握使用线性回归解决预测问题的实现方法。
- 理解选取批量大小的注意事项。
- 理解分类问题中softmax运算的作用。
- 理解softmax回归模型的一般形式、损失函数、优化算法。
- 掌握使用softmax回归解决分类问题的实现方法。
- 了解信息论中交叉熵的作用。

问题

(*) 假设$\mathbf{y} = \mathbf{X}\mathbf{w}, \mathbf{X} \in \mathbb{R}^{N\times d}, \mathbf{w} \in \mathbb{R}^{d\times 1}$，求参数$\mathbf{w}$的解析解。

假设$\mathbf{y} = \mathbf{X}\mathbf{w}, \mathbf{X} \in \mathbb{R}^{N\times d}, \mathbf{w} \in \mathbb{R}^{d\times 1}$，简述随机梯度下降法对参数$\mathbf{w}$的求解算法。

简述选取批量大小的注意事项。

简述分类问题中softmax运算的作用。

简述独热 one-hot 编码在分类问题中的应用。

简述分类问题中损失函数的构造方法，及其计算方法。

(*) 简述信息论中交叉熵的作用，及其在分类问题中的应用。