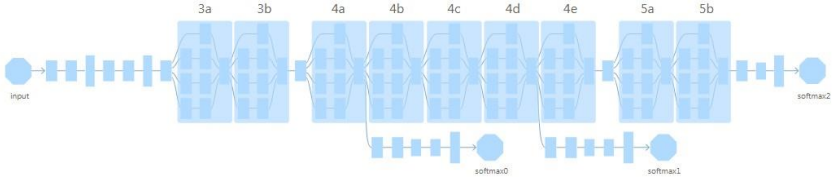
Google TensorFlow™

浙江大学城市学院 ZHEJIANG UNIVERSITY CITY COLLEGE



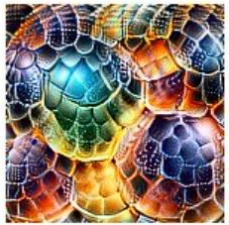
深度学习应用开发
基于TensorFlow的实践

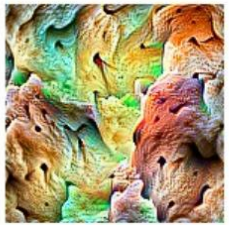
吴明晖 李卓蓉 金苍宏

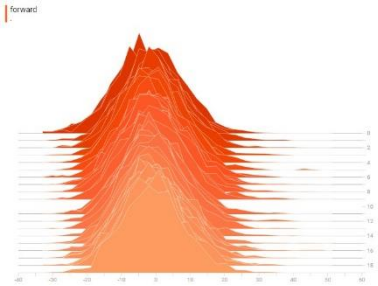
浙江大学城市学院

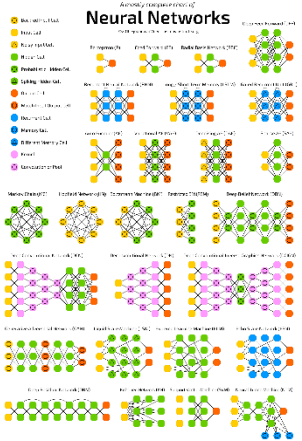
计算机与计算科学学院

Dept. of Computer Science
Zhejiang University City College









波士顿房价预测
多元线性回归问题TensorFlow实践

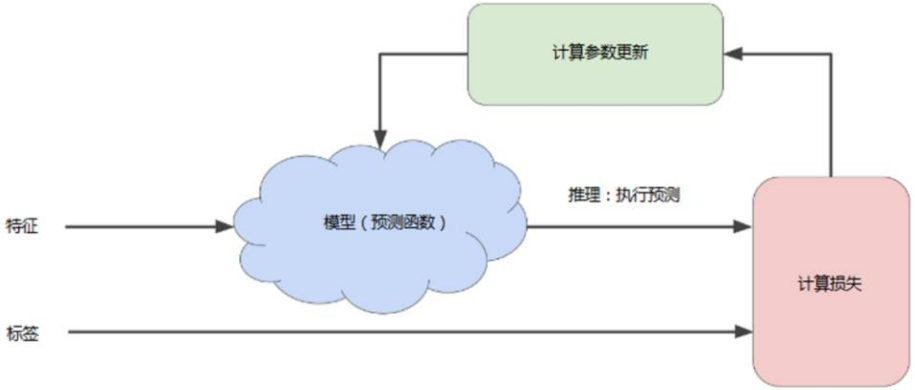
波士顿房价预测

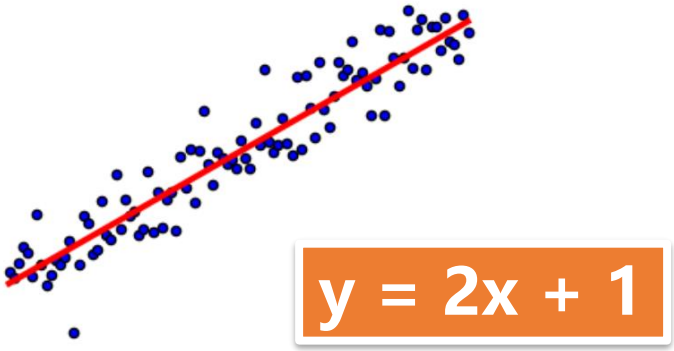
波士顿房价数据集包括**506**个样本，每个样本包括**12个特征变量**和该地区的**平均房价**

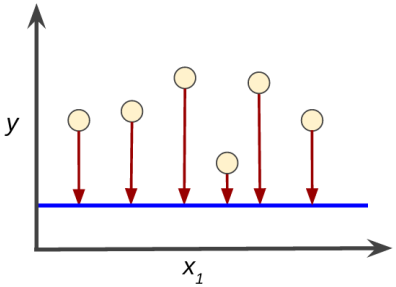
房价（单价）显然和多个特征变量相关，不是单变量线性回归（**一元线性回归**）问题

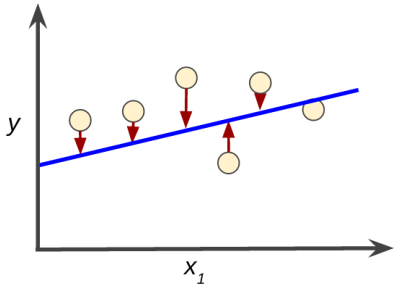
选择多个特征变量来建立线性方程，这就是多变量线性回归（**多元线性回归**）问题

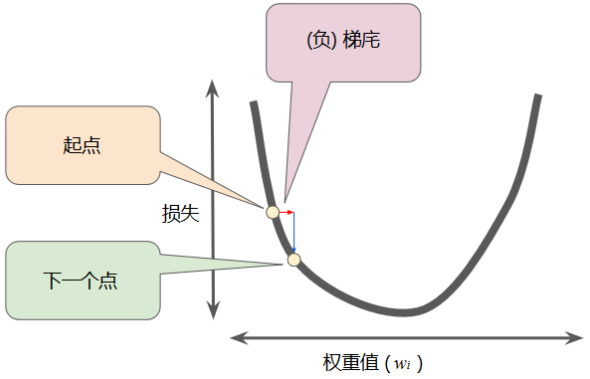
前情回顾：一元线性回归


计算参数更新
特征
模型（预测函数）
推理：执行预测
计算损失
标签


y = 2x + 1


y
x_1


y
x_1


(负) 梯度
起点
损失
下一个点
权重值（w_i）

前情回顾：机器学习的步骤

使用Tensorflow进行算法设计与训练的核心步骤

- （1）准备数据
- （2）构建模型
- （3）训练模型
- （4）进行预测

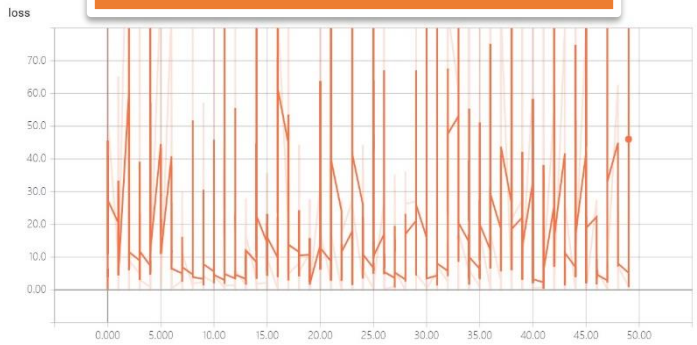
上述步骤是我们使用Tensorflow进行算法设计与训练的核心步骤，贯穿于后面介绍的具体实战中。本章用一个简单的例子来讲解这几个步骤。

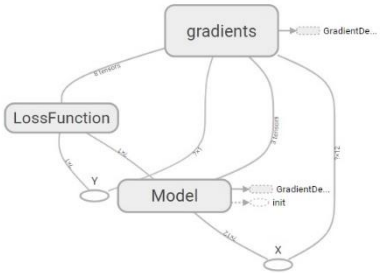
房价预测问题：多元线性回归及TensorFlow编程进阶

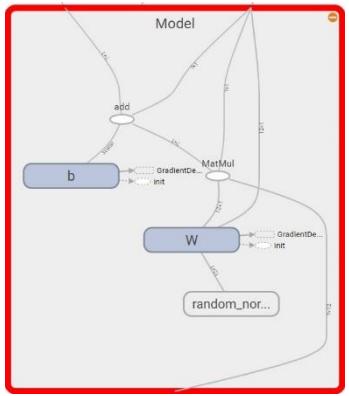
	CRIM	ZN	INDUS	CHAS	NOX	RM	\
count	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000	
mean	3.613524	11.363636	11.136779	0.069170	0.554695	6.284634	
std	8.601545	23.322453	6.860353	0.253994	0.115878	0.702617	
min	0.006320	0.000000	0.460000	0.000000	0.385000	3.561000	
25%	0.082045	0.000000	5.190000	0.000000	0.449000	5.885500	
50%	0.256510	0.000000	9.690000	0.000000	0.538000	6.208500	
75%	3.677082	12.500000	18.100000	0.000000	0.624000	6.623500	
max	88.976200	100.000000	27.740000	1.000000	0.871000	8.780000	

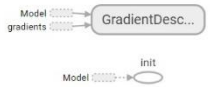
	AGE	DIS	RAD	TAX	PTRATIO	LSTAT	\
count	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000	
mean	68.574901	3.795043	9.549407	408.237154	18.455534	12.653063	
std	28.148861	2.105710	8.707259	168.537116	2.164946	7.141062	
min	2.900000	1.129600	1.000000	187.000000	12.600000	1.730000	
25%	45.025000	2.100175	4.000000	279.000000	17.400000	6.950000	
50%	77.5					11.360000	
75%	94.0					16.955000	
max	100.0					37.970000	

Boston房价预测










```
epoch= 1 loss= 44.3441744805 b= 3.60815 w= [[-0.60289431]
 [ 1.38135946]
 [-0.78309298]
 [ 0.49668884]
 [ 2.51221323]
 [ 7.16771173]
 [-0.05658912]
 [ 0.80084318]
 [ 0.38297549]
 [ 0.33776706]
 [ 2.31645942]
 [-4.39285994]]
epoch= 2 loss= 32.0449512219 b= 3.99625 w= [[ -1.15317142]
 [  1.96271288]
 [ -1.51236773]
 [  0.85116291]
 [  2.88758063]
 [ 10.60607815]
 [ -0.82200134]
 [  0.35972106]
 [  0.63090378]
 [ -0.2541407 ]
 [  1.15306044]
 [ -8.10907555]]
epoch= 3 loss= 27
```

```
x_test = x_test.reshape(1,12)
predict = sess.run(pred, feed_dict={x: x_test})
print("预测值：%f" % predict)

target = y_data[n]
print("标签值：%f" % target)
```

预测值：23.972641
标签值：24.500000

数据读取

数据集解读

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	LSTAT	MEDV
2	0.00632	18	2.31	0	0.538	6.575	65.2	4.09	1	296	15.3	4.98	24
3	0.02731	0	7.07	0	0.469	6.421	78.9	4.9671	2	242	17.8	9.14	21.6
4	0.02729	0	7.07	0	0.469	7.185	61.1	4.9671	2	242	17.8	4.03	34.7
5	0.03237	0	2.18	0	0.458	6.998	45.8	6.0622	3	222	18.7	2.94	33.4
6	0.06905	0	2.18	0	0.458	7.147	54.2	6.0622	3	222	18.7	5.33	36.2
7	0.02985	0	2.18	0	0.458	6.43	58.7	6.0622	3	222	18.7	5.21	28.7

CRIM: 城镇人均犯罪率

ZN：住宅用地超过 25000 sq.ft. 的比例

INDUS: 城镇非零售商用土地的比例

CHAS: 边界是河流为1，否则0

NOX: 一氧化氮浓度

RM: 住宅平均房间数

AGE: 1940年之前建成的自用房屋比例

DIS：到波士顿5个中心区域的加权距离

RAD: 辐射性公路的靠近指数

TAX: 每10000美元的全值财产税率

PTRATIO: 城镇师生比例

LSTAT: 人口中地位低下者的比例

MEDV: 自住房的平均房价，单位：千美元

读取数据

```
%matplotlib notebook

import tensorflow as tf
import matplotlib.pyplot as plt
import numpy as np
import pandas as pd
from sklearn.utils import shuffle
```

```
# 读取数据文件
df = pd.read_csv("data/boston.csv", header=0)

#显示数据摘要描述信息
print (df.describe())
```

读取数据

```
%matplotlib notebook

import tensorflow as tf
import matplotlib.pyplot as plt
import numpy as np
import pandas as pd
from sklearn.utils import shuffle

# 读取数据文件
df = pd.read_csv("data/boston.csv",

#显示数据摘要描述信息
print (df.describe())
```

	CRIM	ZN	INDUS	CHAS	NOX	RM	\
count	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000	
mean	3.613524	11.363636	11.136779	0.069170	0.554695	6.284634	
std	8.601545	23.322453	6.860353	0.253994	0.115878	0.702617	
min	0.006320	0.000000	0.460000	0.000000	0.385000	3.561000	
25%	0.082045	0.000000	5.190000	0.000000	0.449000	5.885500	
50%	0.256510	0.000000	9.690000	0.000000	0.538000	6.208500	
75%	3.677082	12.500000	18.100000	0.000000	0.624000	6.623500	
max	88.976200	100.000000	27.740000	1.000000	0.871000	8.780000	

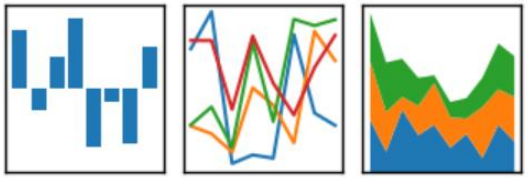
	AGE	DIS	RAD	TAX	PTRATIO	LSTAT	\
count	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000	
mean	68.574901	3.795043	9.549407	408.237154	18.455534	12.653063	
std	28.148861	2.105710	8.707259	168.537116	2.164946	7.141062	
min	2.900000	1.129600	1.000000	187.000000	12.600000	1.730000	
25%	45.025000	2.100175	4.000000	279.000000	17.400000	6.950000	
50%	77.500000	3.207450	5.000000	330.000000	19.050000	11.360000	
75%	94.075000	5.188425	24.000000	666.000000	20.200000	16.955000	
max	100.000000	12.126500	24.000000	711.000000	22.000000	37.970000	

	MEDV
count	506.000000
mean	22.532806
std	9.197104
min	5.000000
25%	17.025000
50%	21.200000
75%	25.000000
max	50.000000

通过pandas读取数据文件，列出统计概述

pandas

$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$



浙江大学城市学院
ZHEJIANG UNIVERSITY CITY COLLEGE

想快速读取常规大小的数据文件时，通过创建读缓存区和其他的机制可能会造成额外的开销。此时建议采用Pandas库来处理。

Pandas官网（http://pandas.pydata.org)这样介绍Pandas：

“Pandas是一款开源的、基于BSD协议的Python库，能够提供高性能、易用的数据结构和数据分析工具。”他具有以下特点：

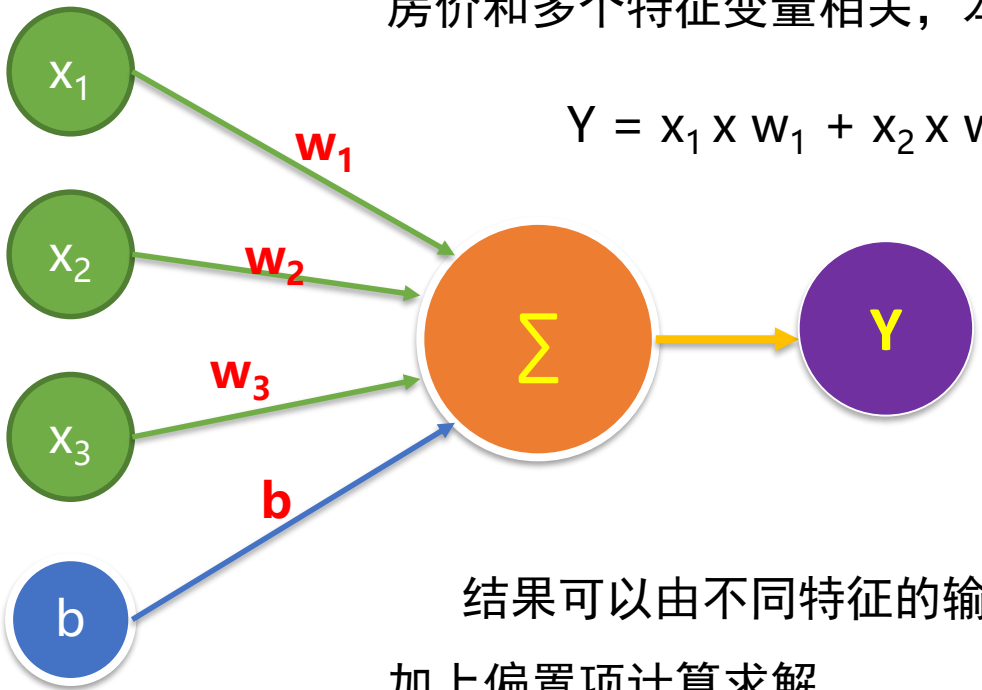
- 能够从CSV文件、文本文件、MS Excel、SQL数据库，甚至是用于科学用途的HDF5格式
- CSV文件加载能够自动识别列头，支持列的直接寻址
- 数据结构自动转换为Numpy的多维数组

准备建模

多元线性回归模型

房价和多个特征变量相关，本讲尝试使用多元线性回归建模

$$Y = x_1 \times w_1 + x_2 \times w_2 + \ldots + x_{12} \times w_{12} + b$$

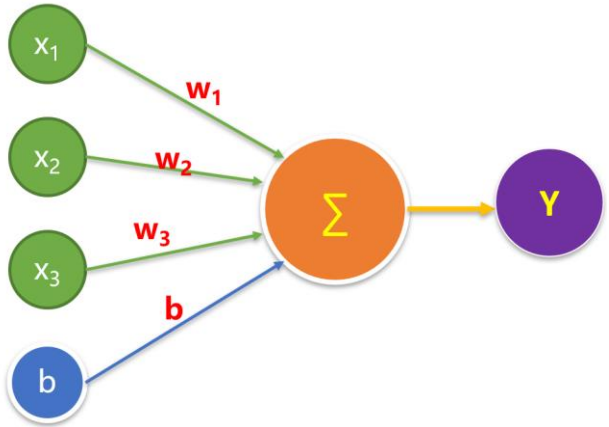


结果可以由不同特征的输入值和对应的权重相乘求和，加上偏置项计算求解

多变量线性方程的矩阵运算表示

$$Y = x_1 \times w_1 + x_2 \times w_2 + \ldots + x_n \times w_n + b$$

$$Y = \sum_{k=0}^{n} x_k * w_k + b$$



$$[x_1 \quad x_2 \quad x_3] * \begin{bmatrix} w_1 \\ w_2 \\ w_3 \end{bmatrix} + b = [x_1 * w_1 + x_2 * w_2 + x_3 * w_3] + b$$

矩阵运算是机器学习的基本手段，必须掌握！