Python 全栈 400 之机器学习练习

404 数据集（data set）

记录的集合，假如我们用 3 个特征，分别为性别、头衔、有无同行人来预测泰坦尼克号上船员的生死，并且拥有基于这 3 个特征的 892 条记录，其中一条记录的取值为：

复制

```
性别=female，头衔=Mrs，有无同行人=True
```

如果记录到 .csv 文件中，这个文件的结构可以记为：`train[892][3]`，这样一个二维数组，行数为 892，列数为 3.

405 示例（instance）

每条记录是关于一个事件或对象的描述，也称为样本，比如以上其中一条记录：

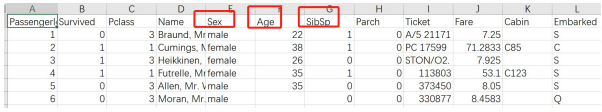
复制

```
性别=female，头衔=Mrs，有无同行人=True
```

可看做是一个实例

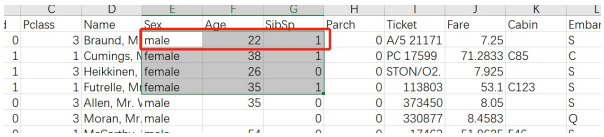
406 属性（attribute）

反映事件或对象在某方面的表现或性质的事项，例如色泽，根蒂，响声等，又称为特征 feature. 如下红框标出的便是 3 个特征：



PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr	male	22	1	0	A/5 21171	7.25		S
2	1	1	Cumings, M	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen,	female	26	0	0	STON/O2.	7.925		S
4	1	1	Futrelle, Mr	female	35	1	0	113803	53.1	C123	S
5	0	3	Allen, Mr. V	male	35	0	0	373450	8.05		S
6	0	3	Moran, Mr.	male		0	0	330877	8.4583		Q

属性上的取值如下红框所示，称为特征的取值。



Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embar
3	Braund, M	male	22	1	0	A/5 21171	7.25		S
1	Cumings, M	female	38	1	0	PC 17599	71.2833	C85	C
3	Heikkinen,	female	26	0	0	STON/O2.	7.925		S
1	Futrelle, Mr	female	35	1	0	113803	53.1	C123	S
3	Allen, Mr. V	male	35	0	0	373450	8.05		S
3	Moran, Mr.	male		0	0	330877	8.4583		Q

407 样本空间（sample space）

样本空间又称为属性空间，attribute space，或输入空间。

它可以理解为训练数据中实际出现的所有属性值构成的集合空间，如果仅考察数据集中的 Genre 列，Genre 列的样本空间为 27，因为 Genre 列一共有 27 种不同取值。

Rank	Title	Genre	Description	Director	Actors	Year	Runtime (M	Rating	Votes	Revenue (M	Metascore
1	Guardians	Action,Adventure,Sci-Fi	A group of	James Gun	Chris Pratt,	2014	121	8.1	757074	333.13	76
2	Prometheu	Adventure,Mystery,Sci-Fi	Following c	Ridley Scot	Noomi Rap	2012	124	7	485820	126.46	65
3	Split	Horror,Thriller	Three girls	M. Night Sl	James McA	2016	117	7.3	157606	138.12	62
4	Sing	Animation,Comedy,Family	In a city of	Christophe	Matthew M	2016	108	7.2	60545	270.32	59
5	Suicide Sq	Action,Adventure,Fantasy	A secret gc	David Ayer	Will Smith,	2016	123	6.2	393727	325.02	40
6	The Great \	Action,Adventure,Fantasy	European r	Yimou Zha	Matt Damc	2016	103	6.1	56036	45.13	42
7	La La Land	Comedy,Drama,Music	A jazz pian	Damien Ch	Ryan Goslir	2016	128	8.3	258682	151.06	93
8	Mindhorn	Comedy	A has-beer	Sean Foley	Essie Davis	2016	89	6.4	2490		71
9	The Lost Ci	Action,Adventure,Biography	A true-life	James Gray	Charlie Hur	2016	141	7.1	7188	8.01	78
10	Passengers	Adventure,Drama,Romance	A spacecra	Morten Tyl	Jennifer Lav	2016	116	7	192177	100.01	41
11	Fantastic Be	Adventure,Family,Fantasy	The advent	David Yate	Eddie Redr	2016	133	7.5	232072	234.02	66

和它有相似的一个概念叫做假设空间（hypothetical space），它是理论上的所有可能属性值构成的集合空间。

如果我们在购买某个股票时假定只考虑两个主要特征：股票经纪公司等级和股票最近3个月的涨幅情况，进而判断是否购买某只股票。

假定股票经纪公司等级取值为 4 种：A等，B等，C等，还要考虑到一种特殊取值 *，这个特征对于是否买这只股票是无关紧要的；

股票最近 3 个月的涨幅情况取值为 3 种：涨，降，`*`（同上面解释）

那么根据这 2 个特征和特征取值，并且股票的标签 y 取值为买或不买，因此理论上可以得到一个由 12 种不同取值组成的假设空间：$4 \times 3 = 12$

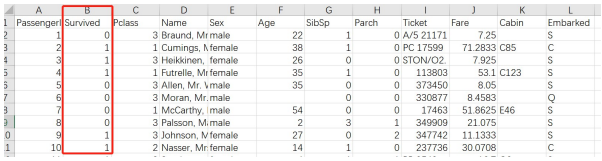
408 特征向量（feature vector）

假如将以下 11 个属性（注意：Survived 列为标签列，不算在内）作为 11 个坐标维度，其值就是一个坐标向量，被称为一个特征向量，记为 $(x1, x2, ..., x_11)$

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr	male	22	1	0	A/5 21171	7.25		S
2	1	1	Cumings, M	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen,	female	26	0	0	STON/O2.	7.925		S
4	1	1	Futrelle, Mr	female	35	1	0	113803	53.1	C123	S
5	0	3	Allen, Mr. V	male	35	0	0	373450	8.05		S
6	0	3	Moran, Mr.	male		0	0	330877	8.4583		Q
7	0	1	McCarthy,	male	54	0	0	17463	51.8625	E46	S
8	0	3	Palsson, M	male	2	3	1	349909	21.075		S
9	1	3	Johnson, W	female	27	0	2	347742	11.1333		S
10	1	2	Nasser, Mr	female	14	1	0	237736	30.0708		C

409 标记（label）

关于样本的标签信息，比如判断船员是否能被获救，那么这位船员便会拥有标记示例，一般用 (Xi , yi) 表示第i个样例，其中yi是样本Xi的标记。如下红框对应列就是样本的标记y_i$



PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr	male	22	1	0	A/5 21171	7.25		S
2	1	1	Cumings, M	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen,	female	26	0	0	STON/O2.	7.925		S
4	1	1	Futrelle, Mr	female	35	1	0	113803	53.1	C123	S
5	0	3	Allen, Mr. V	male	35	0	0	373450	8.05		S
6	0	3	Moran, Mr.	male		0	0	330877	8.4583		Q
7	0	1	McCarthy,	male	54	0	0	17463	51.8625	E46	S
8	0	3	Palsson, M	male	2	3	1	349909	21.075		S
9	1	3	Johnson, W	female	27	0	2	347742	11.1333		S
10	1	2	Nasser, Mr	female	14	1	0	237736	30.0708		C

410 维数（dimensionality）

每个样本包含的属性个数，泰坦尼克号源数据集共有 11 个特征如上图所示，那么它的维数便是 11，这是机器学习中需要理解的重要概念，同时要注意和线代中维数概念加以区分。

如下影评数据集的维数为 12 ：

Rank	Title	Genre	Description	Director	Actors	Year	Runtime (M	Rating	Votes	Revenue (M	Metascore
1	Guardians	Action,Adventure,Sci-Fi	A group of	James Gun	Chris Pratt,	2014	121	8.1	757074	333.13	76
2	Prometheu	Adventure,Mystery,Sci-Fi	Following c	Ridley Scot	Noomi Rap	2012	124	7	485820	126.46	65
3	Split	Horror,Thriller	Three girls	M. Night Sl	James McA	2016	117	7.3	157606	138.12	62
4	Sing	Animation,Comedy,Family	In a city of	Christophe	Matthew M	2016	108	7.2	60545	270.32	59
5	Suicide Sq	Action,Adventure,Fantasy	A secret gc	David Ayer	Will Smith,	2016	123	6.2	393727	325.02	40
6	The Great \	Action,Adventure,Fantasy	European r	Yimou Zha	Matt Damc	2016	103	6.1	56036	45.13	42
7	La La Land	Comedy,Drama,Music	A jazz pian	Damien Ch	Ryan Goslir	2016	128	8.3	258682	151.06	93
8	Mindhorn	Comedy	A has-beer	Sean Foley	Essie Davis	2016	89	6.4	2490		71
9	The Lost Cl	Action,Adventure,Biography	A true-life	James Gray	Charlie Hur	2016	141	7.1	7188	8.01	78
10	Passengers	Adventure,Drama,Romance	A spacecra	Morten Tyl	Jennifer Lat	2016	116	7	192177	100.01	41

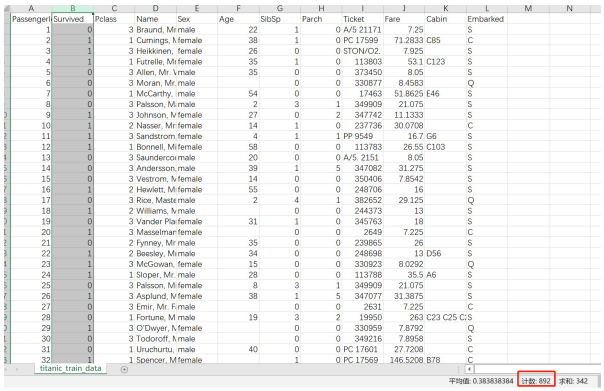
411 学习（learning）

从数据中学得模型的过程，又称为训练（training）。正如上文所示，892 条船员数据集，根据它的 11 个特征和每条特征对应的标记，经过计算最后得到了一个 f，通过这个 f 我们能预测第 893 位船员是否获救，这个过程被称为学习。

412 训练数据 （training data）

训练过程中使用的数据，其中每个样本称为一个训练样本（training sample），训练样本组成的集合称为训练集（training set）。如下泰坦尼克号训练数据集的文件名称

共有 892 行，除去表头共有 891 个样本组成的训练数据，Survived 列为标签。



PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr	male	22	1	0	A/5 21171	7.25		S
2	1	1	Cumings, M	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen,	female	26	0	0	STON/O2.	7.925		S
4	1	1	Futrelle, Mr	female	35	1	0	113803	53.1	C123	S
5	0	3	Allen, Mr. V	male	35	0	0	373450	8.05		S
6	0	3	Moran, Mr.	male		0	0	330877	8.4583		Q
7	0	1	McCarthy,	male	54	0	0	17463	51.8625	E46	S
8	0	3	Palsson, Mi	male	2	3	1	349909	21.075		S
9	1	3	Johnson, M	female	27	0	2	347742	11.1333		S
10	1	2	Nasser, Mr	female	14	1	0	237736	30.0708		C
11	1	3	Sandstrom,	female	4	1	1	PP 9549	16.7	G6	S
12	1	1	Bonnell, Mi	female	58	0	0	113783	26.55	C103	S
13	0	3	Saundercor	male	20	0	0	A/5. 2151	8.05		S
14	0	3	Andersson,	male	39	1	5	347082	31.275		S
15	0	3	Vestrom, M	female	14	0	0	350406	7.8542		S
16	1	2	Hewlett, Mr	female	55	0	0	248706	16		S
17	0	3	Rice, Maste	male	2	4	1	382652	29.125		Q
18	1	2	Williams, M	male		0	0	244373	13		S
19	0	3	Vander Pla	female	31	1	0	345763	18		S
20	1	3	Masselmar	female		0	0	2649	7.225		C
21	0	2	Fynney, Mr	male	35	0	0	239865	26		S
22	1	2	Beesley, Mr	male	34	0	0	248698	13	D56	S
23	1	3	McGowan,	female	15	0	0	330923	8.0292		Q
24	1	1	Sloper, Mr.	male	28	0	0	113788	35.5	A6	S
25	0	3	Palsson, Mi	female	8	3	1	349909	21.075		S
26	1	3	Asplund, M	female	38	1	5	347077	31.3875		S
27	0	3	Emir, Mr. F	male		0	0	2631	7.225		C
28	0	1	Fortune, M	male	19	3	2	19950	263	C23 C25 C	S
29	1	3	O'Dwyer, M	female		0	0	330959	7.8792		Q
30	0	3	Todoroff, M	male		0	0	349216	7.8958		S
31	0	1	Uruchurtu,	male	40	0	0	PC 17601	27.7208		C
32	1	1	Spencer, M	female		1	0	PC 17569	146.5208	B78	C

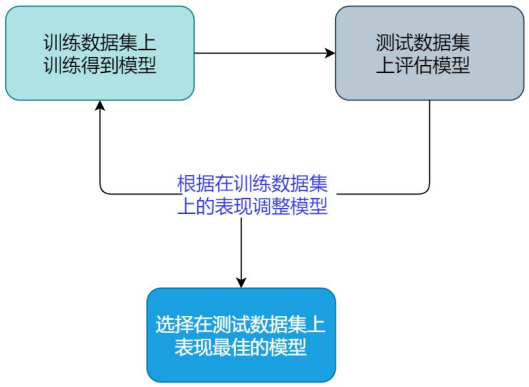
titanic_train_data

通过这些训练数据学习，最终得出一个 f，也就是我们学到的模型。与之相对应的是测试数据，测试数据中缺少标签列。例如，泰坦尼克号测试数据集中没有 Survived 列，是一个 418 行 11 列的数据集。

训练数据主要用于训练模型，训练后得到的模型对训练数据是可见的，那么再基于训练数据评估模型的好坏就完全失去意义，因此我们需要找到一些模型未知的新数据，以此来评估模型才具有价值，我们称这部分数据为测试数据。

通常训练数据占到整个数据集的 80%，测试数据占 20%，如下所示：

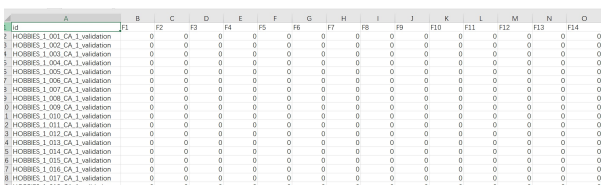
基于训练数据和测试数据模式的机器学习流程，主要就是先在训练数据集上得到一个模型，然后再在测试数据集上评估模型，根据在测试数据集上获得的效果调整模型，然后再训练，重复迭代。从中选出在测试数据集上表现最好的模型。



还有一种更加优秀的训练模式，就是在训练数据集上再拿出一部分数据作为验证集，这种模式就更加避免了模型对测试数据集地可见性，所以能更好的评估我们得到的模型。

413 回归（regression）

如果预测的可能取值不是有限个，例如预测商品在未来 14 天的销量 ，理论上销量可取值为不小于 0 的任意整数值，如下表格是要预测第一列商品在未来 14 天销量的，默认都为 0，所以需要根据在训练集上得到模型预测这些商品的销量。



id	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	F12	F13	F14
HOBBIES_1_001_CA_1_validation	0	0	0	0	0	0	0	0	0	0	0	0	0	0
HOBBIES_1_002_CA_1_validation	0	0	0	0	0	0	0	0	0	0	0	0	0	0
HOBBIES_1_003_CA_1_validation	0	0	0	0	0	0	0	0	0	0	0	0	0	0
HOBBIES_1_004_CA_1_validation	0	0	0	0	0	0	0	0	0	0	0	0	0	0
HOBBIES_1_005_CA_1_validation	0	0	0	0	0	0	0	0	0	0	0	0	0	0
HOBBIES_1_006_CA_1_validation	0	0	0	0	0	0	0	0	0	0	0	0	0	0
HOBBIES_1_007_CA_1_validation	0	0	0	0	0	0	0	0	0	0	0	0	0	0
HOBBIES_1_008_CA_1_validation	0	0	0	0	0	0	0	0	0	0	0	0	0	0
HOBBIES_1_009_CA_1_validation	0	0	0	0	0	0	0	0	0	0	0	0	0	0
HOBBIES_1_010_CA_1_validation	0	0	0	0	0	0	0	0	0	0	0	0	0	0
HOBBIES_1_011_CA_1_validation	0	0	0	0	0	0	0	0	0	0	0	0	0	0
HOBBIES_1_012_CA_1_validation	0	0	0	0	0	0	0	0	0	0	0	0	0	0
HOBBIES_1_013_CA_1_validation	0	0	0	0	0	0	0	0	0	0	0	0	0	0
HOBBIES_1_014_CA_1_validation	0	0	0	0	0	0	0	0	0	0	0	0	0	0
HOBBIES_1_015_CA_1_validation	0	0	0	0	0	0	0	0	0	0	0	0	0	0
HOBBIES_1_016_CA_1_validation	0	0	0	0	0	0	0	0	0	0	0	0	0	0
HOBBIES_1_017_CA_1_validation	0	0	0	0	0	0	0	0	0	0	0	0	0	0

或者是预测今年某地大樱桃的甜度，连续的任意多种可能值，根据今年的雨水量、光照情况等，这种预测称为回归。

回归任务通过学习带标签的数据，得到 f. 在预测时输入 X_m，带入到f中，得到y_m.

414 分类（classification）

如果要预测的是取值有限的离散值，信用交易是否存在欺诈，邮件是否为垃圾邮件，船员是否能获救等等，这些类学习任务被称为分类。

如果分类的结果为两类，又称此分类为二分类，通常称其中一个为正类（positive class），另一个为反类（negative class）。

415 聚类（clustering）

整个数据集都不带标签，但是数据集本身的分布具有一定规律。根据某些特征和聚类算法，可以将训练中的数据分成若干组，自动形成几簇，这些簇可能对应一些潜在的概念。

比如 A 簇中的用户群体都喜欢喜剧电影，或者都信仰某种文化等等这些概念都是我们事先不知道的。

416 有监督学习 (supervised learning）

如果数据集是带标签的数据，则成为有监督学习，比如泰坦尼克号训练数据集学习任务便是有监督学习。

417 无监督学习 (unsupervised learning)

无监督学习是指在无标签数据上的学习过程，常见的聚类任务便是无监督学习，使用聚类算法或神经网络模型最终输出一些潜在的概念。

当然世上很少是非红即白的事，有时给定的数据集中有的含有 y，有的缺失 y，基于此类数据集的学习问题被称为半监督学习任务。

418 泛化能力

泛化能力（**generalization**）是指学到的模型适用于新样本的能力，关乎最后的预测精度。

举个例子来说明什么是泛化能力。 上学那回小明爱动脑筋，老师讲的题目不光会做，还能举一反三；小红学习很努力，上课认真听讲，老师布置的作业完成的非常好，但是这仅限于老师讲过的知识范畴内，小红不怎么喜欢思考，主要是填鸭式地学习知识，老师讲什么她就学什么，并且对老师教授的知识总会一遍又一遍的反复温习。

不过一次数学竞赛，考的题目基本不是以前老师上课讲过的，考试的结果，小明 90，小红 50。

小明的变通能力更强，能根据老师所讲的东西变通解题。但是，小红这方面能力很弱，虽然对老师讲过的知识掌握的很好，但是当题目样式新颖后，她就会答错很多题。

引起泛化能力不足的一个重要原因是过拟合，过拟合导致在测试集上变现非常好，但是在新来的数据集上表现非常差。在以后的学习过程中**过拟合**将会是一个老生常谈的问题。

419 机器学习评价指标之准确率

表面上看这是一个简单的问题，如果分类的**准确率**越高，就断言分类模型越好。

据此评价方法，对于二分类问题，评价分类算法准确率的计算公式为：

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

其中，P 全称 Positive; N 全称 Negative; T 全称 True, 表示预测正确；F 全称 False, 表示预测错误。

如果正负样本个数较为均衡，使用以上评价公式是没有问题的。

实际中，我们要分类的问题大都满足正负样本个数均衡吗？

如果一下能举出很多反例，大概率就可以说正负样本不均衡的情况还是很多。银行卡信贷欺诈判断、交通违规判断、考试作弊判断、垃圾邮件检测、涉黄电影判断、恶性肿瘤检测...

并且下意识告诉我们，这些分类任务的数据集中正负样本个数往往是不均衡的，欺诈的交易总归占据少数，交通违规、考试作弊大概率也如此...

如果正负样本个数比例真是这样不均衡，使用以上公式评价问题就出现了。比如 100 个肿瘤检测报告中，只有 1 个是正类别(确定为肿瘤)，对于这类数据集，我们只要写一行代码，预测所有都为负类别(即确定不是肿瘤)，则：

$$Accuracy = \frac{0+99}{0+99+0+1} = 99\%$$

你看，我们什么都没做，仅靠投机取巧，模型预测的准确率就达到 `99%`，这太匪夷所思！

420 机器学习评价指标之精确率和召回率

显然，仅仅使用准确率评价模型好坏，失败了。原因在于正负样本个数的不均衡，导致评价出现问题。

所以，需要设计出更加科学健全的评价指标。于是就有了**精确率+召回率**的评价体系。

其中，**精确率** 的计算公式为：

$$Precision = \frac{TP}{TP+FP}$$

公式意义：被预测为正类别的样本中，确实为正类别的比率。

召回率的计算公式为：

$$Recall = \frac{TP}{TP+FN}$$

公式意义：在所有正类别样本中，能够正确的识别为正类别的比率。

按照此评价体系，如果还是纯碎靠猜测，即预测 100 个肿瘤全为负类别，则：

$$Precision = \frac{0}{0+0}$$

这种极端情况，我们没有预测出正样本，所以精确率公式失去意义。下面考察召回率：

$$Recall = \frac{0}{0+1} = 0$$

等于 0，所以判定纯碎靠猜是不可取的，所以**精确率+召回率**的评价体系更优于仅凭准确率的方法。

数学知识应用于各个领域。在机器学习中，我们也需要具备一些基本的数学知识，像前面一章介绍的概率论与数理统计。除此之外，我们还需要知道一些最基本的高等数学，线性代数方面的知识。不用太担心，都是最基本的，也都很容易理解。

很多高等数学的知识点被广泛应用在机器学习中，比如， 导数，偏导数， 方向梯度，偏微分方程，拉普拉斯算子，等等。机器学习中到处都可以见到线性代数知识的应用. 比如，主成分分析（ PCA ），奇异值分解，特征值特征向量，向量空间和范数，等等。

今天我们主要介绍最经常用到的知识，那就是求导数和偏导数，以及线性代数中的矩阵特征值和特征向量以及矩阵的分解。

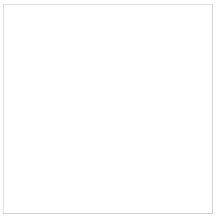
421 导数以及偏导数

首先我们回想一下，什么是导数？简单的讲导数就是一种数学上对于改变率的一种连续的描述。

> 导数就是对于一个或者多个自变量发生小的变化的时候，函数值是怎么变化的这种规律的一个连续描述。

如果我们想把上面的话用数学式子表示出来，应该怎么做呢？也就是说你怎么描述自变量改变很小，函数值变化，连续呢？

这就是数学语言的魅力所在，它可以精确的把这些描述性的语言，准确的用方程式给我们量化出来。如果我们以一元函数 f(x) 为例，那么导数的定义式就是：



而对于多元函数，由于有多个变量，这时候对于某一个变量的导数就改称为偏导数了。偏导数的定义也是一样的, 以二元函数f(x,y)为例，偏导数的可以定义为：

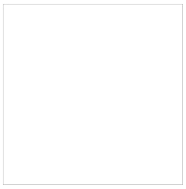
$$\frac{\partial f}{\partial x} = \lim_{h \to 0} \frac{f(x+h,y) - f(x,y)}{h}$$

$$\frac{\partial f}{\partial y} = \lim_{h \to 0} \frac{f(x,y+h) - f(x,y)}{h}$$

通过这些方程式，我们就可以回到上面的问题，比如什么是自变量改变很小，我们用的是 h->0 来定义, 自变量从x变到 x+h 就是变化很小。当然，这些概念是建立在极限的概念之上。如果不知道极限的话，需要先学习极限的概念。

我们现在列出一些经常用到的求导公式。

求导公式在机器学习的梯度下降中经常使用，因为梯度就意味着要求导，所以将使用频率最高的几个公式罗列在下面，方便查阅。



其中，第三个公式是第二个求导公式的特列

我们都知道在高等数学中，导数的一个很重要的应用就是求极值和判断函数的单调性。这也是考试中经常或者必考的题目。这种题目，相信大家都知道，就是求一阶或者二阶导数，然后解方程求解，等等。

那么，在机器学习中，导数的应用在哪里呢？

在机器学习中，我们用导数来寻求极值，也就是最优解。下面我们会以线性回归为例子来仔细理解导数的应用。

线性回归(linear regression)是我们经常用到的模型，线性回归模型是描述的是自变量和因变量之间是一种线性关系。这个咱们在中学就知道的线性关系：如果我们用X表示自变量，用Y表示因变量，那么两者之间的线性关系就可以定义为下面的方程：

```
Y = kX + b
```

复制

在这个方程中，X-自变量， Y-因变量， k-斜率，b-截距。这些都是咱们早就知道的。而在线性回归模型中，就是X和Y已经给定了，我们要做的是我们要训练一个线性模型。那么也就是我么要决定在上面的方程中的系数：k 和 b的值。而且我们要使得我们训练的模型和真实的值之间拥有最小的误差。

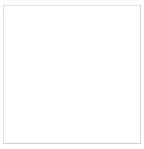
那么怎么做到让我们的误差最小，也就是说咱们建立一个最好的线性回归模型呢？为了达到这个目的，我们用一个叫做所示函数的东西来量化这个误差，从而找到最小误差，也就是找到最好模型。

422 链式求导法则

求导比较重要的一条性质便是链式求导法则，求导数意味着由外及内，一层一层地将变化传递到最里头。

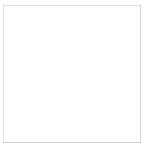
例如，对 J 函数求导，自变量为 θ ，(x^i, y^i) 为已知参数：

对 J 函数，先设定如下 g 函数：

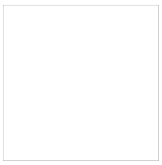


先对 $g(\theta)$ 函数的平方求导数，

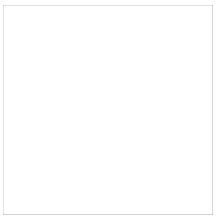
然后再对 $g(\theta)$ 求导数，结果为：



由链式求导法则，再联立以上式子得到：



将 g 函数带入得到最终求导公式：



423 手推特征值和特征向量

求下面矩阵 A 的特征值和特征向量

$$A = \begin{pmatrix} 1.2 & 0.8 \\ 0.8 & 1.2 \end{pmatrix}$$

下面我们手动推导求出特征值，特征值满足如下等式：

$$Ax = \lambda x$$
$$(A - \lambda E)x = 0$$
$$|A - \lambda E| = 0$$

带入矩阵 A 到上式，求得两个特征值 2 或 0.4

当特征值等于 2 时，计算对应的特征向量为 (1,1)

$$\lambda = 2$$

$$\begin{pmatrix}1.2 & 0.8\\ 0.8 & 1.2\end{pmatrix}\begin{pmatrix}a\\ b\end{pmatrix}=2\begin{pmatrix}a\\ b\end{pmatrix}$$

$$1.2a+0.8b=2a$$

$$0.8b=0.8a$$

$$a=b$$

$$\begin{pmatrix}1\\ 1\end{pmatrix}$$

当特征值等于 0.4 时，计算对应的特征向量为 (1,-1)

$$\lambda=0.4$$

$$\begin{pmatrix}1.2 & 0.8\\ 0.8 & 1.2\end{pmatrix}\begin{pmatrix}a\\ b\end{pmatrix}=0.4\begin{pmatrix}a\\ b\end{pmatrix}$$

$$1.2a+0.8b=0.4a$$

$$0.8a=-0.8b$$

$$a=-b$$

$$\begin{pmatrix}1\\ -1\end{pmatrix}$$

特征值对于矩阵是重要的，能够反映矩阵的重要取值特性，后面的矩阵特征值分解，都会用到求矩阵的特征值和特征向量。

424 损失函数(loss function)

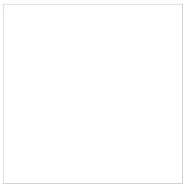
损失函数描述了你建立的模型的预测值和真实值之间的差别。在我们这个线性回归的例子中，就是我们预测的 k 和 b 的预测值给我们带来的误差，自然我们的目标是要让这个误差最小。

首先怎么定义这个误差？

其实有很多的误差定义方式，我们这里将会使用一个最常用的标准：均方误差(mean square error or MSE). 得到这个均方误差我们可以想成有两步：

- 找出真实值和预测值之间的差别：也就是真实的 y 和预测得到的 y；
- 然后把这个差别值先取平方，再对每一个自变量值 X 取均值。

这样我们就得到了均方误差的公式如下, 其中 yi 是真实值， 而 是预测值。



我们有了这个均方误差方程之后，我们还记得咱们的目标是要使得这个误差最小，也就是咱们要找出这个误差最小情况下的系数们：k 和 b的值。

那么怎么来找这个误差的最小值呢？我们将会使用梯度下降算法来找出误差的最小值，这就是为什么我们还使用了优化算法在们建立的线性回归模型中。当然，别的回归模型也是一样的道理。

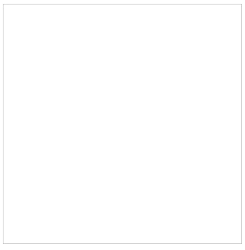
425 梯度下降算法的3大求解步骤

梯度下降法是一种经常使用的找最小值的优化算法。在这里，我们使用梯度下降法来找均方误差的最小值。下面我们会一步一步的很简单的展现我们是怎么使用这个梯度下降法找到这个误差最小值，也就是最优的预测值 k 和 b的值的。

于是问题就变成了怎么求最小值的问题，也就是求导数或者求偏导数问题。所以我们下面就对上面的损失函数求偏导数。

Step 1 求偏导数

先求对 k 的偏导数：



同样的，下面我们再求对 b 的偏导数：

上面就完成了求偏导数的部分。那么这个偏导数有什么用呢？我们知道在迭代过程中，我们需要知道两个部分：一个是前进方向，一个就是步长。而我们上面求的偏导数就是迭代的前进方向。而迭代的步长，我们可以假设步长是固定的，比如我们取 0.01.

Step 2 初始化

我们假设初始化的 k 和 b的值都为0: k=0, b=0;

**Step 3 使用下面的方程式来更新当前的 k 和 b **

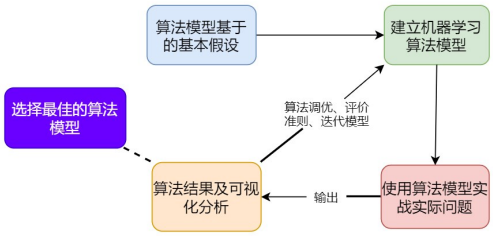
重复上面的迭代过程，知道我们的损失函数很小，或者足够小于我们事先设置的预知（比如 <0.00001), 我们就可以停止。

这样我们就得到了最优的 k,b 的值，也就是说我们建立了我们的线性回归模型。我们就可以使用我们所建立好的模型进行预测了。

通过上面线性回归的例子我们就理解了导数的应用， 导数或者偏导数确实是应用在机器学习中的。

426 机器学习模型迭代示意图

正式进入机器学习的理论和实践阶段，机器学习开展的基本思路：



427 线性回归模型的三个假定

为了保证使用的线性数学模型能够取得较好的拟合效果，那么有三个前提假定就非常重要。

1）假设真实值与预测值的误差项 ϵ 服从正态分布；

2）假定每个样本之间都是相互独立的；

3）预测的数据分布和训练时用到的数据分布是相同的（至于为什么在 Day 49 有解释）

428 阐述建立线性回归模型的主要推导过程

因为第 j 个样本的误差项 $\epsilon^{(j)}$ 服从高斯分布，因此可得：

$$f(\epsilon^{(j)}) = \frac{1}{\sqrt{2\pi}\sigma} e^{(-\frac{\epsilon^{(j)^2}}{2\sigma^2})}$$

因为建立的是线性数学模型，因此第 j 个样本根据模型预测值为

$$f(x^{(j)}|\theta) = \theta_0 + \sum_{i=1}^{n} \theta_i x^{(j)}$$

损失函数为：

$$\epsilon^{(j)} = f(x^{(j)}|\theta) - y^{(j)}$$

以上式子中：

$x^{(j)}$，$y^{(j)}$分别表示第 j 个样本的实际取值；

n 表示特征的个数

综上可得：

$$f(\varepsilon^j) = \frac{1}{\sigma\sqrt{2\pi}} e^{\left(-\frac{(\Sigma_{i=1}^{n} x^{(j)}\theta_i + \theta_0 - y^{(j)})^2}{2\sigma^2}\right)}$$

至此，我们得到一个含有 $n+1$ 个特征参数的等式，f 表示事件 ϵ^j ，也就是第 j 个样本的误差项的概率密度值。

参数估计中，使用最大似然估计(Maximum Likelihood Estimation，简称为 MLE) 求权重参数，接下来介绍。

429 最大似然估计和梯度下降求参数的过程

最大似然估计会使得已经发生的所有事件联合概率取值最大，上面说到的第 3 个假定样本每个特征间是相互独立的，所以 m 个样本误差概率密度 $f(\epsilon^j)$ 同时都发生的概率转化为累乘积：

$$\prod_{j=1}^{m} f(\epsilon^{(j)})$$

样本个数 m 通常会很很大，所以相乘的结果会很小。

通常做法转化为求对数，因此又称最大对数似然估计，可得如下公式：

$$\max\left(\sum_{j=1}^{m} log f(\varepsilon^{(j)})\right)$$

结合已经得出的公式：

$$f(\varepsilon^{j}) = \frac{1}{\sigma\sqrt{2\pi}} e^{\left(-\frac{(\sum_{i=1}^{n} x^{(j)}\theta_i + \theta_0 - y^{(j)})^2}{2\sigma^2}\right)}$$

最终得到：

$$\max\left(\sum_{j=1}^{m} log \frac{1}{\sigma\sqrt{2\pi}} e^{\left(-\frac{(\sum_{i=1}^{n} x^{(j)}\theta_i + \theta_0 - y^{(j)})^2}{2\sigma^2}\right)}\right)$$

上式含有 n 个未知权重参数，如何求解当上式取得最大值时各个参数的取值，使用梯度下降方法。

梯度下降法是一种经常使用的找最小值的优化算法（梯度下降的详细实施步骤大家参考 Day 47 ）。在这里，我们使用梯度下降法来找最小值，因此需要对上节式子取反后求最小值，故：

$$J(\theta) = \min\left(-\sum_{j=1}^{m} log \frac{1}{\sigma\sqrt{2\pi}} e^{\left(-\frac{(\sum_{i=1}^{n} x^{(j)}\theta_i + \theta_0 - y^{(j)})^2}{2\sigma^2}\right)}\right)$$

接下来，求出 $J(\theta)$ 对权重参数 $\thetai的偏导数每次迭代时步， \theta i$ 的更新公式：

$$\theta_i^{l} = \theta_i - \eta \frac{\partial J(\theta_i)}{\partial \theta_i}$$

其中 η 是学习率

至此公式推导全部结束。

430 正则化项的感性解读

有多少特征，就有多少参数需要学习。机器学习学习过程常见的问题之一便是过拟合，过拟合的重要表现就是训练数据集上表现会很好，因为它会试图去满足很多个性化的分布，进而失去泛化的能力，因此在训练数据集上的表现就会变糟糕。

例子解释一下，特征个数 $n = 5$，假如未添加正则项时，学习到 5 个参数分别为：

$ w0 = 0.4, w1 = -0.5, w2 = 1.5, w3 = -0.4, w_4 = 0.6$

每个参数的绝对值权重都相差不是很大，通俗理解就是每个参数都发挥差不多的作用。但是我们想惩罚某几个参数的作用，削弱它们，增强某些参数的作用。

添加常用的 $L2$ 正则项后，也就是添加一项： λW_i^2 ，假如 $\lambda = 1$，惩罚后各个参数的值：

$ w0 = 0.16, w1 = 0.25, w2 = 2.25, w3 = 0.16, w_4 = 0.36$

实施 $L2$ 正则化后，w_3 的相对权重变得更加突出，并且弱化了其他参数的权重，起到惩罚的作用。

以上就是正则化项的感性认识。

431 L1 和 L2 正则化的稀疏性比较解读

$L1$ 和 $L2$ 正则的一个主要不同：相比 $L2$， $L1$正则更容易使模型变稀疏，下面通俗易懂解释为什么。

$L1$ 是对模型中每个特征参数取绝对值，$L2$ 正则对特征参数取平方

如果施加 $L1$，则新的损失函数 f 为 ：

$$Loss() + C|w|$$

要想消除此特征的作用，只需要令 $w = 0$ 时，使 f 取得极小值。因为当 f 取得最小值时，必然保证参数 w 变为 0.

且容易证明，添加 $L1$正则后 ，只要满足：系数 C 大于原函数 $Loss()$在 0 点处导数的绝对值。

证明过程如下，要想在 0 点处取得极小值，根据高等数学知识得到：

1) w 小于 0 时，$\frac{\sigma(Loss)}{\sigma(w)} - C < 0$

2) 且 w 大于 0 时，$\frac{\sigma(Loss)}{\sigma(w)} + C > 0$

上面两个式子同时满足时，可以简化为：$|\frac{\sigma(Loss)}{\sigma(w)}| < C\ (x = 0)$

但是如果施加 $L2$ 正则，则新的函数为：$Loss() + Cw^2$ ，求导可得：

$\frac{\sigma(Loss)}{\sigma(w)} + 2Cw$

要想在 $w = 0$ 点处取得极小值，必须得满足:

$\frac{\sigma(Loss)}{\sigma(w)} = 0$

如果原函数 $Loss()$ 在 0 点处的导数不为 0，那么施加 $L2$ 正则后偏导数不会为 0，也就不会在0 点处取得极小值。这种概率很明显小于 $L1$ 正则在 0 点处取得极小值的概率值。

由此可得，$L1$ 更容易使得原来的特征变弱或消除，换句话说就是更容易使参数变稀疏。

下一章

互动评论

说点什么

评论

The Scrapper 1 个月前

又理解多一点

鼓掌

朱泊宇 3 个月前

后面跟60天内容一样了啊？？

鼓掌

zglg（作者） 3 个月前

题目主要整体到 Day50 左右，Day50 - Day 60 主要偏实践，没有整理到450题里。

鼓掌

朱泊宇 3 个月前

请问下431道后面的还有吗？？？后面看是之前讲的案例了啊。

鼓掌

存