作业说明：

分类任务，根据一个人的年龄、职业、学历等信息预测此人的收入情况（大于 50000 美元或小于 50000 美元），因此为 2 分类问题。

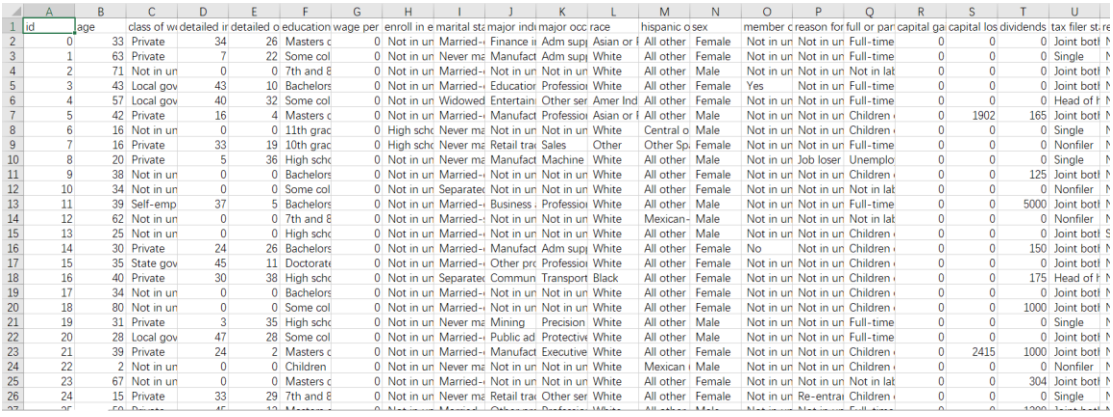
数据集说明：

数据存放于 data 文件夹中，共 6 个文件：

train.csv：原格式的训练数据，54256 行数据，代表 54256 个角色

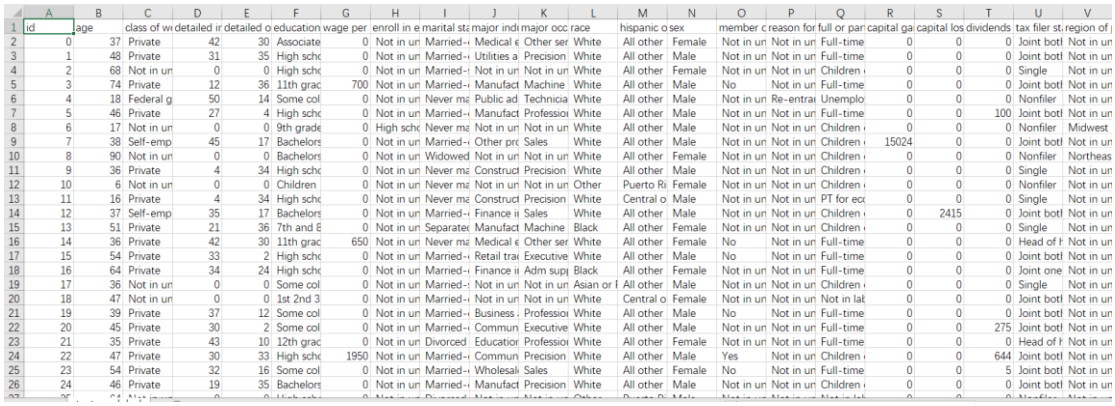
第一列为 id，最后一列为预测结果，即收入大于 50000 美元或小于 50000 美元；

其余各列为用于分类的属性值，例如：年龄、职业等；



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	id	age	class of w	detailed in	detailed o	education	wage per	enroll in e	marital sta	major indu	major occ	race	hispanic o	sex	member o	reason for	full or par	capital ga	capital los	dividends	tax filer st
2	0	33	Private	34	26	Masters d	0	Not in un	Married-s	Finance i	Adm supp	Asian or I	All other	Female	Not in un	Not in un	Full-time	0	0	0	Joint both
3	1	63	Private	7	22	Some col	0	Not in un	Never ma	Manufact	Adm supp	White	All other	Female	Not in un	Not in un	Full-time	0	0	0	Single
4	2	71	Not in un	0	0	7th and 8	0	Not in un	Married-s	Not in un	Not in un	White	All other	Male	Not in un	Not in un	Not in lat	0	0	0	Joint both
5	3	43	Local gov	43	10	Bachelors	0	Not in un	Married-s	Educatior	Professio	White	All other	Female	Yes	Not in un	Full-time	0	0	0	Joint both
6	4	57	Local gov	40	32	Some col	0	Not in un	Widowed	Entertain	Other ser	Amer Ind	All other	Female	Not in un	Not in un	Full-time	0	0	0	Head of h
7	5	42	Private	16	4	Masters d	0	Not in un	Married-s	Manufact	Professio	Asian or I	All other	Male	Not in un	Not in un	Children	0	1902	165	Joint both
8	6	16	Not in un	0	0	11th grac	0	High scho	Never ma	Not in un	Not in un	White	Central o	Male	Not in un	Not in un	Children	0	0	0	Single
9	7	16	Private	33	19	10th grac	0	High scho	Never ma	Retail trac	Sales	Other	Other Sp	Female	Not in un	Not in un	Full-time	0	0	0	Nonfiler
10	8	20	Private	5	36	High scho	0	Not in un	Never ma	Manufact	Machine	White	All other	Male	Not in un	Job loser	Unemplo	0	0	0	Single
11	9	38	Not in un	0	0	Bachelors	0	Not in un	Married-s	Not in un	Not in un	White	All other	Female	Not in un	Not in un	Children	0	0	125	Joint both
12	10	34	Not in un	0	0	Some col	0	Not in un	Separatec	Not in un	Not in un	White	All other	Female	Not in un	Not in un	Not in lat	0	0	0	Nonfiler
13	11	39	Self-emp	37	5	Bachelors	0	Not in un	Married-s	Business	Professio	White	All other	Male	Not in un	Not in un	Full-time	0	0	5000	Joint both
14	12	62	Not in un	0	0	7th and 8	0	Not in un	Married-s	Not in un	Not in un	White	Mexican-	Male	Not in un	Not in un	Not in lat	0	0	0	Nonfiler
15	13	25	Not in un	0	0	High scho	0	Not in un	Married-s	Not in un	Not in un	White	All other	Male	Not in un	Not in un	Children	0	0	0	Joint both
16	14	30	Private	24	26	Bachelors	0	Not in un	Married-s	Manufact	Adm supp	White	All other	Female	No	Not in un	Children	0	0	150	Joint both
17	15	35	State gov	45	11	Doctorate	0	Not in un	Married-s	Other pro	Professio	White	All other	Female	Not in un	Not in un	Children	0	0	0	Joint both
18	16	40	Private	30	38	High scho	0	Not in un	Separatec	Commun	Transport	Black	All other	Female	Not in un	Not in un	Children	0	0	175	Head of h
19	17	34	Not in un	0	0	Bachelors	0	Not in un	Married-s	Not in un	Not in un	White	All other	Female	Not in un	Not in un	Children	0	0	0	Joint both
20	18	80	Not in un	0	0	Some col	0	Not in un	Married-s	Not in un	Not in un	White	All other	Female	Not in un	Not in un	Children	0	0	1000	Joint both
21	19	31	Private	3	35	High scho	0	Not in un	Never ma	Mining	Precision	White	All other	Male	Not in un	Not in un	Full-time	0	0	0	Single
22	20	28	Local gov	47	28	Some col	0	Not in un	Married-s	Public ad	Protective	White	All other	Male	Not in un	Not in un	Full-time	0	0	0	Joint both
23	21	39	Private	24	2	Masters d	0	Not in un	Married-s	Manufact	Executive	White	All other	Female	Not in un	Not in un	Children	0	2415	1000	Joint both
24	22	2	Not in un	0	0	Children	0	Not in un	Never ma	Not in un	Not in un	White	Mexican	Male	Not in un	Not in un	Children	0	0	0	Nonfiler
25	23	67	Not in un	0	0	Masters d	0	Not in un	Married-s	Not in un	Not in un	White	All other	Female	Not in un	Not in un	Not in lat	0	0	304	Joint both
26	24	15	Private	33	29	7th and 8	0	Not in un	Never ma	Retail trac	Other ser	White	All other	Female	Not in un	Re-entrar	Children	0	0	0	Single

test_no_label.csv：原格式的测试数据，与 tain.csv 类似，没有给出预测结果，需要模型输出；



id	age	class of wo	detailed in	detailed o	education	wage per	enroll in e	marital sta	major indu	major occ	race	hispanic o	sex	member c	reason for	full or part	capital gai	capital los	dividends	tax filer st	region of
0	37	Private	42	30	Associate	0	Not in un	Married-(	Medical e	Other ser	White	All other	Female	Not in un	Not in un	Full-time	0	0	0	Joint both	Not in un
1	48	Private	31	35	High scho	0	Not in un	Married-(	Utilities a	Precision	White	All other	Male	Not in un	Not in un	Full-time	0	0	0	Joint both	Not in un
2	68	Not in un	0	0	High scho	0	Not in un	Married-(	Not in un	Not in un	White	All other	Female	Not in un	Not in un	Children	0	0	0	Single	Not in un
3	74	Private	12	36	11th grac	700	Not in un	Married-(	Manufact	Machine	White	All other	Male	No	Not in un	Full-time	0	0	0	Joint both	Not in un
4	18	Federal g	50	14	Some col	0	Not in un	Never ma	Public ad	Technicia	White	All other	Male	Not in un	Re-entrai	Unemploy	0	0	0	Nonfiler	Not in un
5	46	Private	27	4	High scho	0	Not in un	Married-(	Manufact	Professio	White	All other	Male	Not in un	Not in un	Full-time	0	0	100	Joint both	Not in un
6	17	Not in un	0	0	9th grade	0	High scho	Never ma	Not in un	Not in un	White	All other	Male	Not in un	Not in un	Children	0	0	0	Nonfiler	Midwest
7	38	Self-emp	45	17	Bachelors	0	Not in un	Married-(	Other prc	Sales	White	All other	Male	Not in un	Not in un	Children	15024	0	0	Joint both	Not in un
8	90	Not in un	0	0	Bachelors	0	Not in un	Widowed	Not in un	Not in un	White	All other	Female	Not in un	Not in un	Children	0	0	0	Nonfiler	Northeas
9	36	Private	4	34	High scho	0	Not in un	Never ma	Construct	Precision	White	All other	Male	Not in un	Not in un	Children	0	0	0	Single	Not in un
10	6	Not in un	0	0	Children	0	Not in un	Never ma	Not in un	Not in un	Other	Puerto Ri	Female	Not in un	Not in un	Children	0	0	0	Nonfiler	Not in un
11	16	Private	4	34	High scho	0	Not in un	Never ma	Construct	Precision	White	Central o	Male	Not in un	Not in un	PT for ec	0	0	0	Single	Not in un
12	37	Self-emp	35	17	Bachelors	0	Not in un	Married-(	Finance a	Sales	White	All other	Male	Not in un	Not in un	Children	0	2415	0	Joint both	Not in un
13	51	Private	21	36	7th and 8	0	Not in un	Separatec	Manufact	Machine	Black	All other	Female	Not in un	Not in un	Children	0	0	0	Single	Not in un
14	36	Private	42	30	11th grac	650	Not in un	Never ma	Medical e	Other ser	White	All other	Female	No	Not in un	Full-time	0	0	0	Head of h	Not in un
15	54	Private	33	2	High scho	0	Not in un	Married-(	Retail trac	Executive	White	All other	Male	No	Not in un	Full-time	0	0	0	Joint both	Not in un
16	64	Private	34	24	High scho	0	Not in un	Married-(	Finance a	Adm supp	Black	All other	Female	Not in un	Not in un	Full-time	0	0	0	Joint one	Not in un
17	36	Not in un	0	0	Some col	0	Not in un	Married-S	Not in un	Not in un	Asian or F	All other	Male	Not in un	Not in un	Children	0	0	0	Single	Not in un
18	47	Not in un	0	0	1st 2nd 3	0	Not in un	Married-(	Not in un	Not in un	White	Central o	Female	Not in un	Not in un	Not in lab	0	0	0	Joint both	Not in un
19	39	Private	37	12	Some col	0	Not in un	Married-(	Business	Professio	White	All other	Male	No	Not in un	Full-time	0	0	0	Joint both	Not in un
20	45	Private	30	2	Some col	0	Not in un	Married-(	Commun	Executive	White	All other	Male	Not in un	Not in un	Full-time	0	0	275	Joint both	Not in un
21	35	Private	43	10	12th grac	0	Not in un	Divorced	Educatior	Professio	White	All other	Female	Not in un	Not in un	Full-time	0	0	0	Head of h	Not in un
22	47	Private	30	33	High scho	1950	Not in un	Married-(	Commun	Precision	White	All other	Male	Yes	Not in un	Children	0	0	644	Joint both	Not in un
23	54	Private	32	16	Some col	0	Not in un	Married-(	Wholesali	Sales	White	All other	Female	No	Not in un	Full-time	0	0	5	Joint both	Not in un
24	46	Private	19	35	Bachelors	0	Not in un	Married-(	Manufact	Precision	White	All other	Male	Not in un	Not in un	Children	0	0	0	Joint both	Not in un

sample_submission.csv：输出样例

两列数据，第一列为 id，第二列为预测结果

id	label
1	0
2	0
3	0
4	0
5	0
6	0
7	0
8	0
9	0
10	0
11	0
12	0
13	0
14	0
15	0
16	0
17	0
18	0
19	0
20	0
21	0
22	0
23	0
24	0
25	0
26	0

X_train、Y_train：处理后的训练数据，其中 Y_train 为输出部分，X_train 为输入部分。

建议使用 notepad 打开；

相较于原格式的训练数据，对其中的非数值属性进行 one-hot 编码，例如将每一种职业扩展为一个属性，若从事该职业，对应属性为 1，否则为 0；方便模型处理。

```
id,age, Private, Self-employed-incorporated, State governm
0,33,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0
1,63,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
2,71,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
3,43,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
4,57,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
5,42,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0
6,16,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
7,16,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0
8,20,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
9,38,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
```

X_train

```
id,label
0,1
1,0
2,0
3,0
4,0
5,1
6,0
7,0
8,0
```

Y_train

X_test：类似 X_train，对应的输出 Y_test 未给出，需要模型预测。

输入:

直接使用处理过的数据即可

训练数据：X_train、Y_train

测试数据：X_test

输出:

输出一个.csv 文件，参照 sample_submission.csv 的格式，第一列为测试数据中所有人的 id，第二列为收入预测结果。

示例程序:

1. Logistic_regression，逻辑回归方法，运行后会创建 output_logistic.csv，内容为对测试数据的预测；具体细节见代码及注释。

   其余输出内容:

   训练测试数据参数:

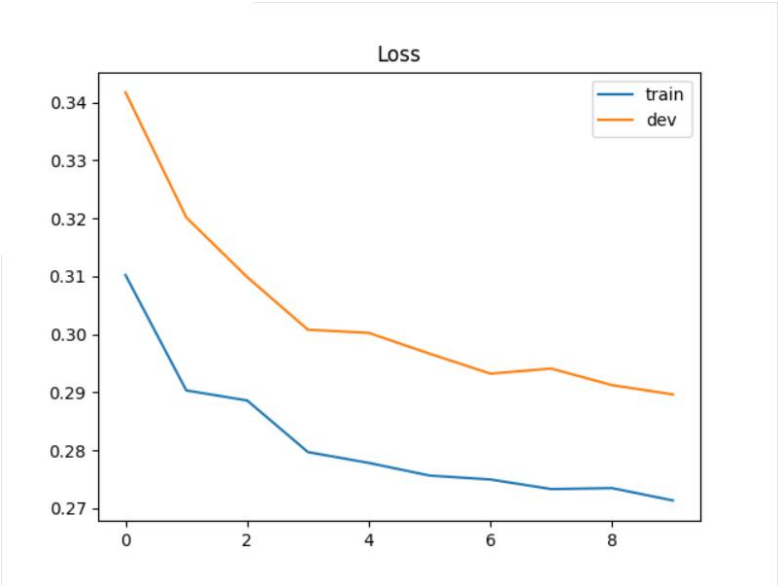
```
Size of training set: 48830
Size of development set: 5426
Size of testing set: 27622
Dimension of data: 510
```

   训练集、验证集效果:

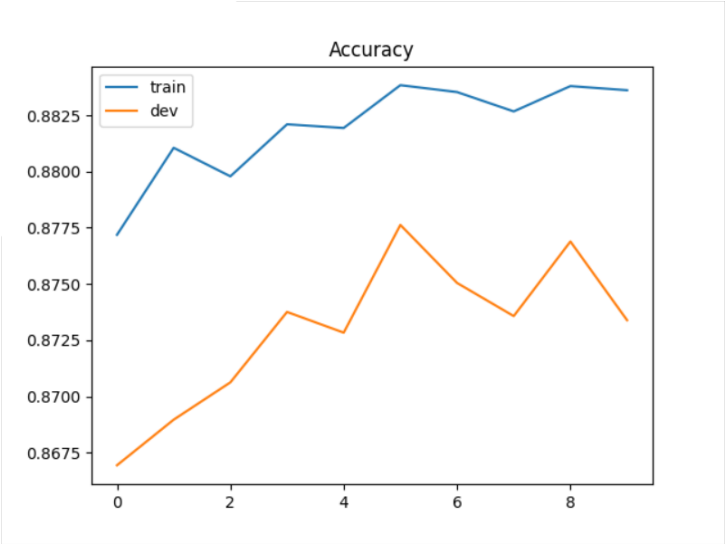
```
Training loss: 0.271355435246406
Development loss: 0.2896359675026287
Training accuracy: 0.8836166291214418
Development accuracy: 0.8733873940287504
```

训练效果可视化:

损失函数变化图:



准确率变化图:



数据前 10 项特征对应的权重：

```
Other Rel <18 never married RP of subfamily -1.4195759775765406
Child 18+ ever marr Not in a subfamily -1.2958572076664743
Unemployed full-time 1.171255828588591
Other Rel <18 ever marr RP of subfamily -1.167791807296237
Italy -1.093458143800618
Vietnam -1.0630365633146408
num persons worked for employer 0.9389922773566495
1 0.8226614922117187
```

2. Generative_model 方法和 Logistic regression 方法类似，不同之处在于 Generative model 可以直接计算出 w 和 b 的最佳解，而 Logistic regression 是将 w 和 b 进行初始化，通过迭代训练来更新 w 和 b，代码除了求解 w 和 b 地方不一样，其他地方类似。