数据科学基础课程报告

王晨曦

2018 年 6 月 13 日

目 录

1 任务描述

Adult 数据集的目标是根据一些特征来预测一个人的收入高低 (年收入 $50K 为高)，完成对 Adult 数据集的分析报告。

- 探索数据集，包括各特征的分布和潜在相关性。
- 对数据集进行合适的划分。
- 构造多种分类模型。
- 评估各种模型的预测结果。
- 对数据集的分析结论。

2 数据集探索

本章首先介绍了数据集概况，并对数据集进行清洗，接着展示了各属性的分布情况，最后通过图像展示了属性间的潜在相关性。

2.1 Adult 数据集

Adult 数据集 [1] 源自政府真实数据，包含了十五个属性（年龄、工作类别、最终权重、教育程度、受教育时间、婚姻状况、职业、家庭关系、种族、性别、资本收益、资本亏损、每周工作小时数、祖国、收入），既有连续数值变量，也有离散类别变量。

Adult 是机器学习和数据挖掘领域的经典数据集，其目的在于预测人的收入是否高于 50K，详细属性介绍可参见附录 A。本次实验中用到的 Adult 数据集为官网提供版本[1]，共有 48842 个实例，部分实例有缺失值。

2.2 数据集清洗

原始数据集有 48842 个实例，但部分实例存在缺失值情况，详细统计见表 1。

表 1: 数据集缺失情况统计

属性			缺失数量	属性			缺失数量
编号	名称	类型		编号	名称	类型	
1	年龄	数值型	0	9	种族	字符串	0
2	工作类别	字符串	2799	10	性别	字符串	0
3	最终权重	数值型	0	11	资本收益	数值型	0
4	教育程度	字符串	0	12	资本亏损	数值型	0
5	教育时间	数值型	0	13	工作时间	数值型	0
6	婚姻状况	字符串	0	14	祖国	字符串	857
7	职业	字符串	2809	15	收入	字符串	0
8	家庭关系	字符串	0	总计			**3620**

[1] http://archive.ics.uci.edu/ml/datasets/Adult

经统计，数据集中有三个属性存在缺失值情况，共计 3620 个实例。为了后续模型能使用完整的数据，又不破坏原数据的分布，对于缺失值的情况，统一采取了删除操作，经删除后剩余 45222 个实例。

2.3 单属性统计

2.3.1 数值变量

Adult 数据集中共有 6 个数值型变量，其统计性参数见表 2。从四分位数可以发现，资本收益、资本亏损的分布极不均匀，大部分都为 0。

表 2: 数值变量统计量

	年龄	最终权重	教育时间	资本收益	资本亏损	工作时间
平均数	38.76	189436	10.07	1082.91	87.90	40.39
标准差	13.85	105715	2.57	7583.94	403.11	12.48
最小值	17.00	13492	1.00	0.00	0.00	1.00
下四分位数	28.00	116736	9.00	0.00	0.00	40.00
中位数	37.00	177831	10.00	0.00	0.00	40.00
上四分位数	48.00	238384	12.00	0.00	0.00	45.00
最大值	90.00	1490400	16.00	99999.00	3770.00	99.00

2.3.2 年龄

图 1 展示了数据集中的年龄分布情况，曲线为拟合结果，数据集中 30-40 岁的人群最多，周围分布最密集，这种与社会中的人员构成基本相符。

2.3.3 最终权重

图 2 展示了数据集中最终权重的分布情况，曲线为拟合结果。最终权重是根据一系列的社会指标给每个人分配的一个综合分数。可以看到，绝大部分人的权重处于中低端水平，极少数人群享有大量权重。

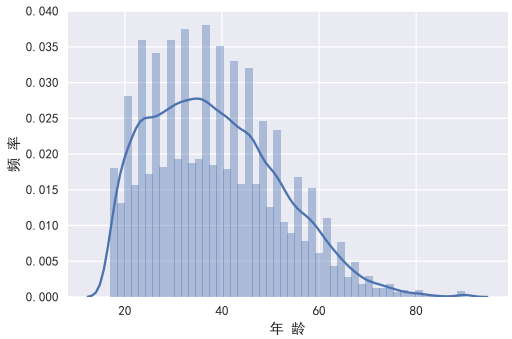

图 1: 年龄分布

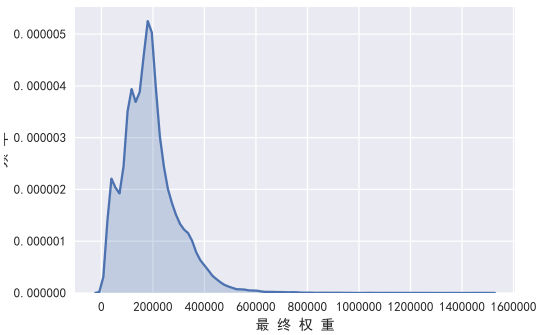

图 2: 最终权重

2.3.4 职业

图3展示了职业的频率分布直方图，可以看到，Adm-clerical、Exec-managerial、Prof-specialty、Sales 和 Craft-repair 等职位人数较多，而 Armed-Forces 和 Priv-house-serv 等职位人数很少。

2.3.5 教育程度

图4展示了数据集中人群教育程度的分布情况，从图中能够发现，大部分人的学历集中在高中、专科或本科，平均人口素质较高。

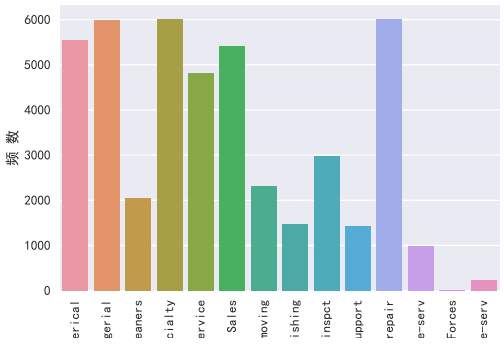

图 3: 职业

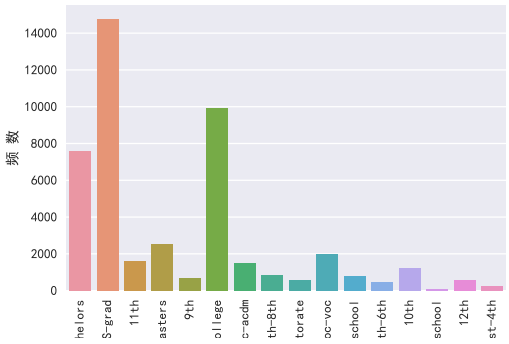

图 4: 教育程度

2.3.6 工作类别

图5展示了数据集中人们的工作类型分布情况，绝大部分人都从事着私人性质的工作，而无偿工作的人占的比例最小，几乎没有。

2.3.7 家庭关系

图6展示了数据集中人们家庭关系的分布状况，丈夫占到将近一半的比例，而不在家和未婚的人占的比例也相当高。

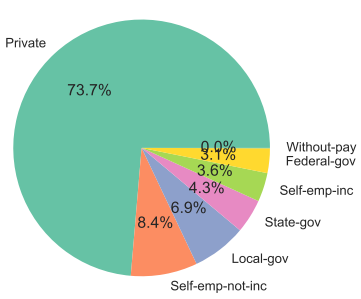

图 5: 工作类别

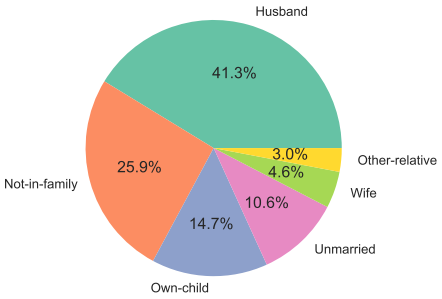

图 6: 家庭关系

2.3.8 受教育时间

图7展示了数据集中人们受教育时间的基本情况，曲线为拟合结果，大部分人的受教育时间集中在8-13年内，这与教育程度的分布相呼应。

2.3.9 每周工作小时数

图8展示了数据集中人们每周的工作时间分布拟合结果，大部分人的工作时间在40小时上下波动，也即5天8小时工作制。

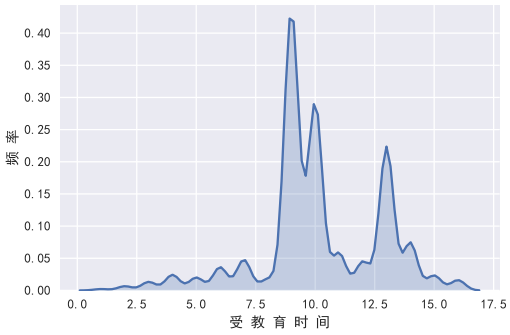

图7: 受教育时间

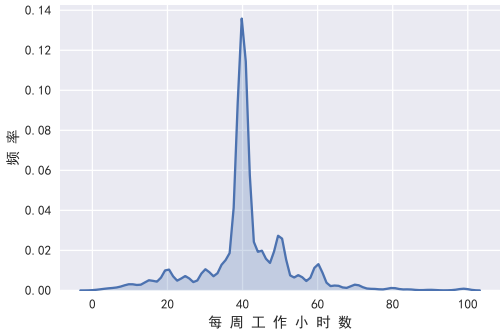

图8: 每周工作小时数

2.3.10 种族

图9展示了数据集中的种族分布情况，可以看到白人在这一地区占有绝对优势，黑人所占比例也不小，接下来是亚裔人口，其余族裔人群数量较少，这与美国社会的人口比例基本相符。

2.3.11 性别

图10展示了数据集中的男女人口比例，男性在该地区占到超过三分之二的人口，这与该地区的发展模式和职业供应有着密切关系。

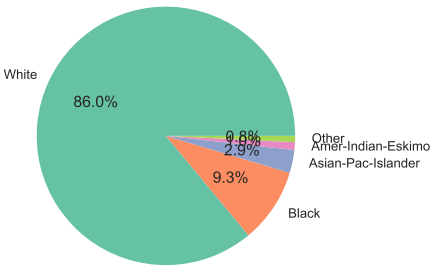

图9: 种族

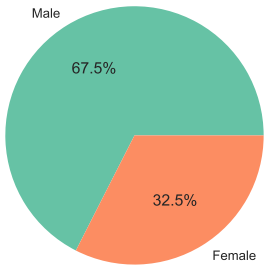

图10: 性别

2.3.12 资本收益

图 11 展示了数据集中的人群资本收益情况箱型图，可以看到，大部分人的资本收益都处在相近水平，异常值极少。

2.3.13 资本亏损

图 12 展示了数据集中的人群资本亏损情况箱型图，与资本收益分布情况相比，人们的资本亏损差异较大，表现为异常值很多且分布范围广。

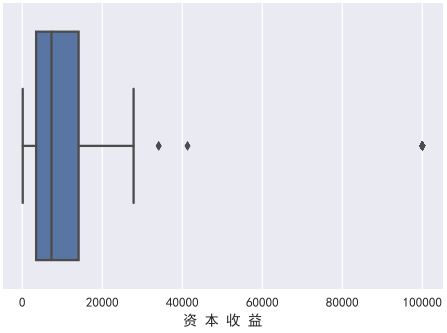

图 11: 资本收益

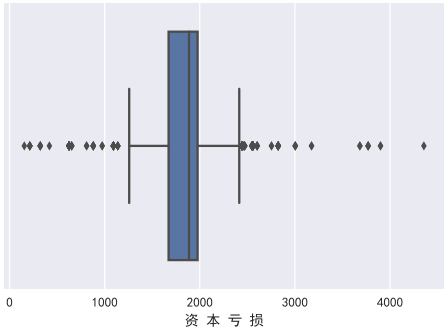

图 12: 资本亏损

2.3.14 祖国

图 13 展示了数据集中人口国籍情况，为了方便观看，纵坐标改为对数增长形式。可以看到美国国籍的人口比例占有绝对优势，外来人口十分稀少，基本都不超过一百人。

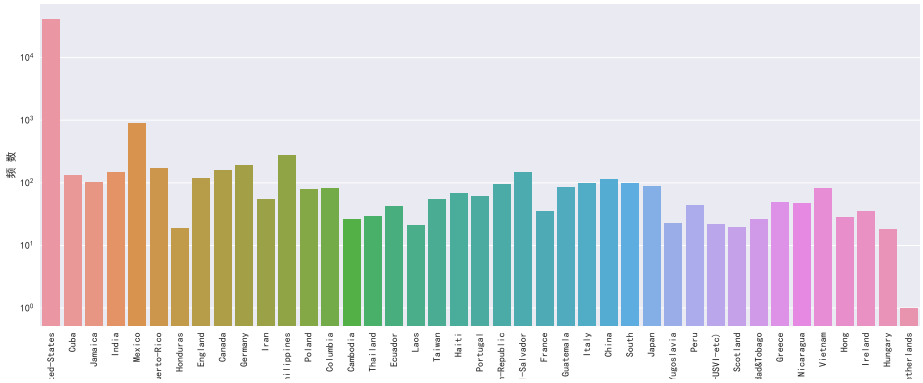

图 13: 祖国

2.3.15 婚姻状况

图 14 展示了数据集中的人群婚姻状况分布，其中已婚和单身人士最多，超过一半人至少有一次婚姻经历。

2.3.16 收入

图15展示了数据集中人群的收入分布状况，在这里被离散化表示成超过50K和低于50K两个类，可以看出，该地区四分之三的人收入都超过了50K，说明该地区为高收入地区。

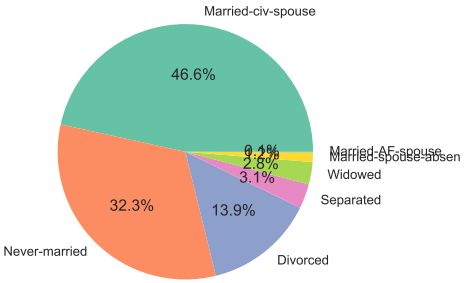

图 14: 婚姻状况

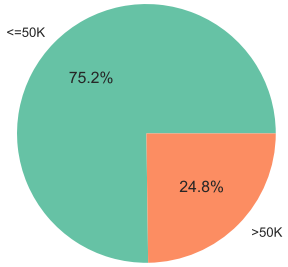

图 15: 收入

2.4 多属性分析

2.4.1 数值变量间的关系

对于数据集中的数值型变量，可以通过两两绘制散点图的方式探索其中的联系。由于图像太大，无法在正文展示，故把散点图放在了附录B的图34中。

实际上，仅从图中难以找到十分有用的信息，这说明要想发掘数据间潜在的联系，还要用更多方法进一步探索。

2.4.2 教育程度与受教育时间

通过前一部分的探究，我们已经知晓，教育程度和受教育时间的分布极为相似，图16直观的展示了教育程度和受教育时间的关系。可以发现两个变量实际上是同一个变量的两种表述，教育程度与受教育时间的16个值一一对应。因此，在Adult数据集中，由教育程度能直接推出受教育时间，反之亦然。

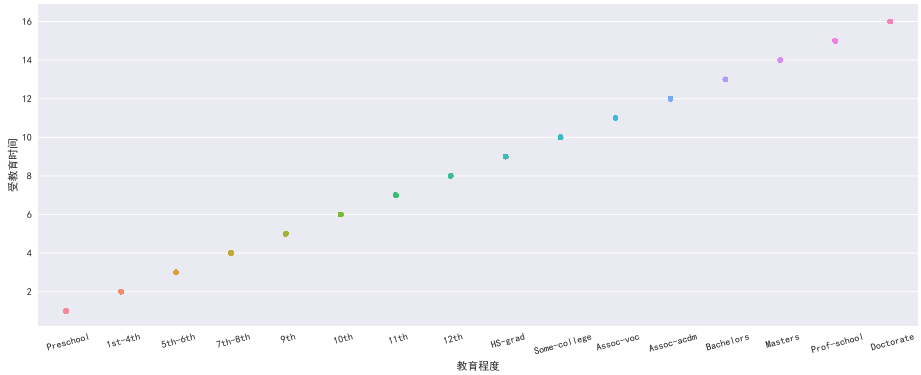

图 16: 教育程度与受教育时间的关系

2.4.3 受教育时间与资本盈亏

图 17 和图 18 分别展示了资本收益和资本亏损随受教育时间变化的误差棒状图，其中每条棒的中心点位置和长度分别代表该教育时间下所有人收益（或亏损）的均值和方差。

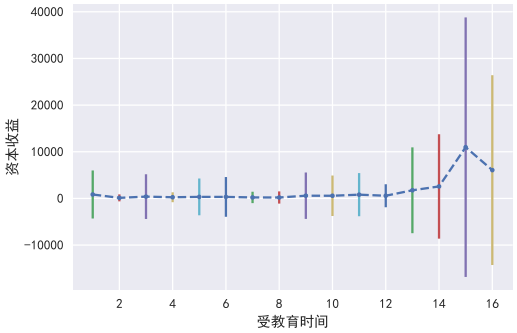

图 17: 受教育时间与资本收益的关系

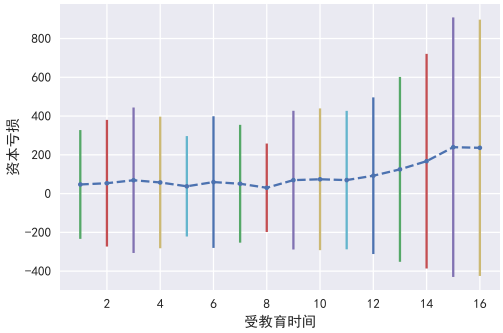

图 18: 受教育时间与资本亏损的关系

对比两图可以发现，不论是资本收益还是资本亏损，其均值与受教育时间均呈正相关，随着受教育时间的增长，资本盈亏的变化幅度也越来越大，这一点在左图中比较明显；不论是哪一个教育程度的人群，其资本亏损的差异都是巨大的。

2.4.4 家庭关系与工作时间

图 19 展示了家庭关系与工作时间的关系箱型图，仅从不同家庭关系的工作时长分布上就能看出很多不同。

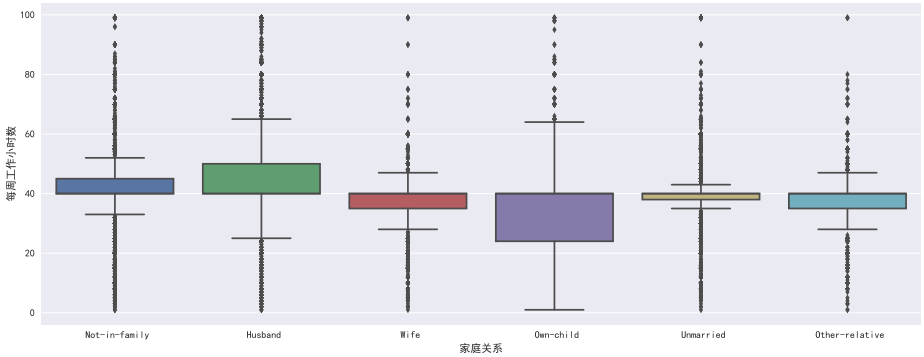

图 19: 家庭关系与工作时间的关系

从图中可以看到，外出离家的人和丈夫工作时间比较长，而妻子和照顾孩子的人工作时间比较短。这一点并不难理解，如果有人需要在家庭中花费更多精力，那么外出工作时间相应就会减少了。

2.4.5 职业与受教育时间

图 20 展示了不同职业的人群中受教育时间的分布情况箱型图，可以看出差异十分之大。

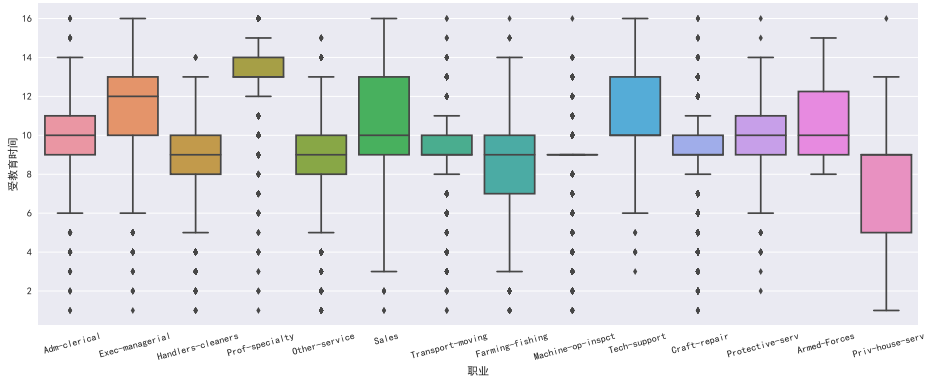

图 20: 职业与受教育时间的关系

根据图中反映的情况，我们能推断出不同职位对学历要求的影响：如 Exec-managerial、Prof-specialty、Sales 和 Tech-support 对学历要求较高，而 Handlers-cleaners、Farming-fishing 和 Priv-house-serv 对学历要求相对较低。

当然，不能因为不同职业中学历分布不同就能完全推断出职业的学历需求，因为人们接受了不同程度的教育之后，其自身的求职兴趣也可能会发生改变。

2.4.6 教育和工作中的性别影响

上一节中已经知道，该地区的男女比例十分不平衡，这可能是因为该地区职业文化和生活方式的影响。这种影响不仅涉及人口比例，还可能波及到男女差异化的方方面面。

图 21 展示了该地区男女人群受教育时间的分布情况箱型图，从图中可以看出，女性受教育时间的平均水平与男性相当，但是最长受教育时间不如男性。这说明该地区尖端人才中男女数量不平衡。

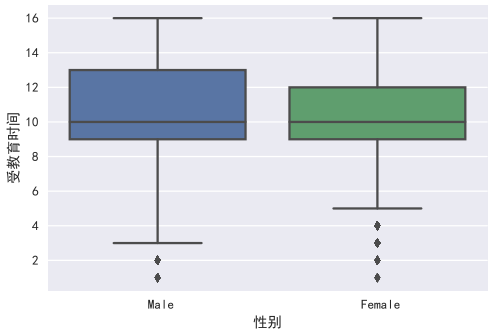

图 21: 男女受教育时间对比

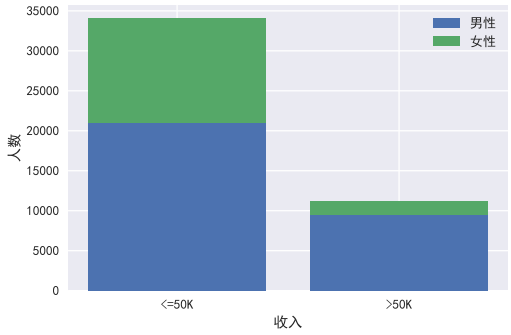

图 22: 男女收入对比

图 22 展示了该地区男女人群的收入情况分布，尽管女性只占了该地区四分之一的人口，但收入超过 50K 的人群中女性比例远远小于这个数字，大部分女性的收入都在 50K 以下。当然，这并不一定是职场的性别歧视，很可能是因为该地区提供的职业更为适合男性，从总体男女比例就可见一斑。

2.4.7 婚姻状况与工作时间

图 23 展示了不同婚姻状况下的人群每周工作时间分布箱型图。整体上来说，已婚人群的平均工作时间较长，而未婚和丧偶人群的工作时间相对短些，这说明一个正常稳定的家庭结构对提升工作积极性有很大帮助。

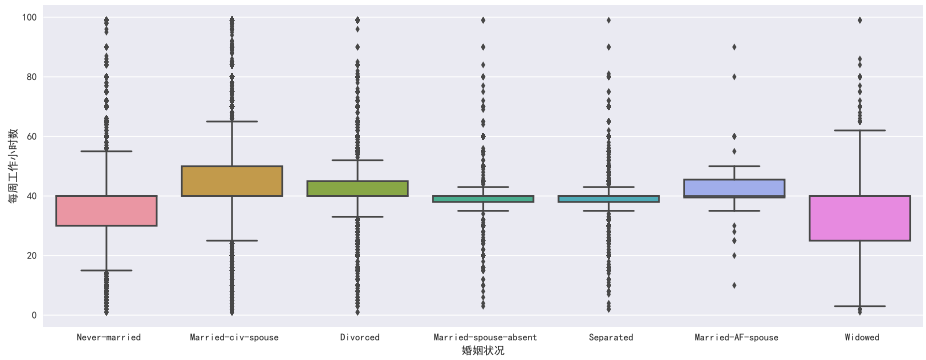

图 23: 婚姻与工作时间的关系

3 实验设置

本章先介绍实验中会用到的两种数据表示方法以及数据集的划分方式，接着对本次实验使用的评价指标做作简要介绍，最后列出所有使用的模型。

3.1 数据表示

由于 Adult 数据集中大量属性为字符串表示的离散类别值，所以需要用合适的表示方式将其数字化，才能供进一步的模型训练和预测。

本次实验使用了两种表示方式：**离散数值编码**和**独热向量编码**。离散数值编码是将每个字符串属性中的所有值取出，将每种字符串用一个数值代替，一般从 0 开始计数；独热向量编码（one hot vector）则是在离散数值编码的基础上，将每种属性转为一个稀疏向量，只有对应位的值为 1，其余全为 0。

我们可以对两种数据表示方式做一个初步的认识，对 KNN 和 SVM 模型分别训练两种数据，比较他们的不同。经离散数值编码的数据有 14 维，独热向量编码的数据有 104 维，表 3 展示了两个模型对不同数据表示的预测结果。

表 3: 数据表示方式对比

模型	KNN			SVM		
	准确率	F1 分数	运行时间	准确率	F1 分数	运行时间
离散数值编码	**0.8199**	**0.82**	**4.06**	0.8295	0.81	**33.50**
独热向量编码	0.8186	0.81	39.07	**0.8339**	**0.82**	102.71

可以看到，两种表示方式在准确率上的优势因模型而异，但独热向量编码后的数据比离散向量编码后的数据运行时间要长得多，这是因为独热编码会使数据的维数大幅增加，给

训练造成了困难。

在接下来的实验中，为了时间便利性，没有特殊说明的部分均采用离散数值编码，独热向量编码会单独说明。

3.2 数据集划分

实验中采用了官方数据集的划分方式（三分之一测试数据），将 48842 个实例中的 32561 项作为训练集，其余 16281 项作为测试集。经过数据清洗，训练集和测试集的实例个数分别为 30162 和 15060，共计 45222 个实例。

3.3 评价指标

本次实验用三种指标来衡量模型的性能：准确率、F1 分数和运行时间，前两者反映了模型的效果，第三个反映了模型的效率，详细介绍如下。

3.3.1 准确率

模型的准确率是指在所有样本中分类正确的样本所占的比例，对于一个分类模型来说，准确率是其最基础的性能度量，但它有很多不足之处，无法全面反映一个模型的性能。因此很多时候还需要有其他指标去评价模型的优劣性。

3.3.2 F1 分数

用 TP、FP 和 FN 分别代表真阳性、假阳性和假阴性样本的个数，则模型的精确率 P 和召回率 R 可分别表示为

$$P = \frac{TP}{TP + FP}$$
$$R = \frac{TP}{TP + FN}$$

对于复杂问题来说精确率和召回率有着不可调和的矛盾，精确率越高，往往召回率就越低，反之亦如此。为了能综合反映模型的性能，定义 F1 分数如下：

$$F1 = 2 \times \frac{PR}{P + R}$$

可以看出，F1 分数是精确率和召回率的调和平均，因此只有当二者均很高时 F1 分数才会较高；若其中有一方明显低于另一方，F1 分数就会大打折扣。

3.3.3 运行时间

一个模型的训练和测试需要花费时间，有的模型训练花的时间长，有的模型测试花的时间长，通过比较模型的运行时间，我们能估计出使用各种模型的性价比，这对于大数据集或者大量重复任务有很大的帮助。

在本次实验中，所有时间的单位均为秒（s）。

3.4 分类模型

本次实验总共使用了六种模型：SVM 分类器 [2]、MLP 分类器 [3]、KNN 分类器 [4]、决策树 [5]、朴素贝叶斯分类器 [6] 和随机森林 [7]，在下一章中会对六种模型的性能依次展开讨论。

4 模型性能探究

4.1 SVM 分类器

SVM 是分类任务中的经典模型，在很多问题上有着广泛的应用。我们可以调节 SVM 中的惩罚系数和核函数类别，观察 SVM 的性能。

4.1.1 惩罚系数对 SVM 的影响

惩罚系数是影响 SVM 泛化能力的重要参数，实验中我将惩罚系数从 0.2 逐渐增大到 2.0，观察 SVM 的准确率和训练时间，结果记录在图 24 中。

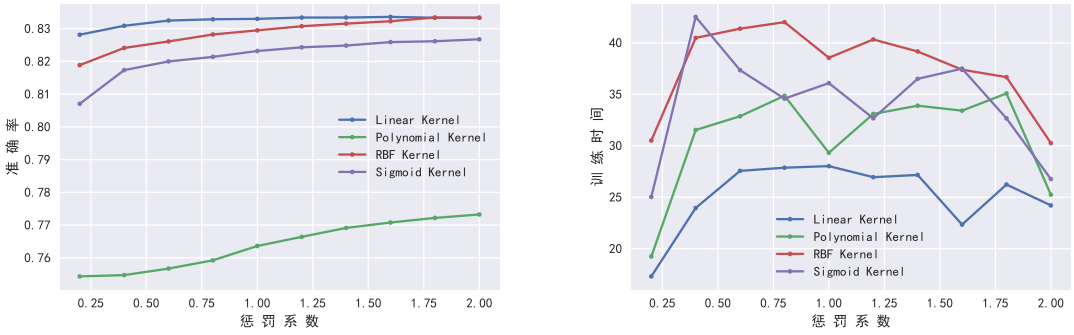

图 24: 惩罚系数对 SVM 性能的影响

从左图中可以看到，在实验限定的搜索空间内，随着惩罚系数的增加，各类模型的准确率都在逐渐提升，其中线性核表现最好，但随着系数的增加，RBF 和也赶了上来，追平了线性核，多项式核准确率最低，只有不到 0.8。

从右图可以看到，随着惩罚系数的增加，各种核函数的训练时间先增后减，但相对排名基本不变，线性核最低，RBF 核最高。由此可见，虽然线性核和 RBF 核表现都比较好，但考虑时间因素后，线性核的优势就大了很多。

4.2 神经网络

神经网络是除了 SVM 之外另一个著名的模型，在深度学习领域尤为重要，下面的实验中的神经网络均为简单的多层感知机（MLP）结构。

4.2.1 网络深度对 MLP 的影响

神经网络的一大优势在于其深度的拓展会提升模型的性能但也可能会适得其反。在这里我们用一种统一的结构 $(32, 64, \cdots, 64, 16, 2)$，仅改变模型的深度，测试 MLP 的性能。

图 25 左图展示了 MLP 训练和测试准确率随深度的变化趋势，在 Adult 数据集上，过深的 MLP 并没有起到帮助作用，反而降低了性能，这可能是因为太深的网络难以训练，一点点微笑的误差都会引起巨大的变化。

图25 右图展示了 MLP 训练和测试时间随深度的变化趋势，虽然预测时间几乎不变，但训练时间随着深度增加会大幅增长，对比准确率的变化，更能说明 MLP 并不是越深越好，要根据不同的数据集灵活的做出选择。

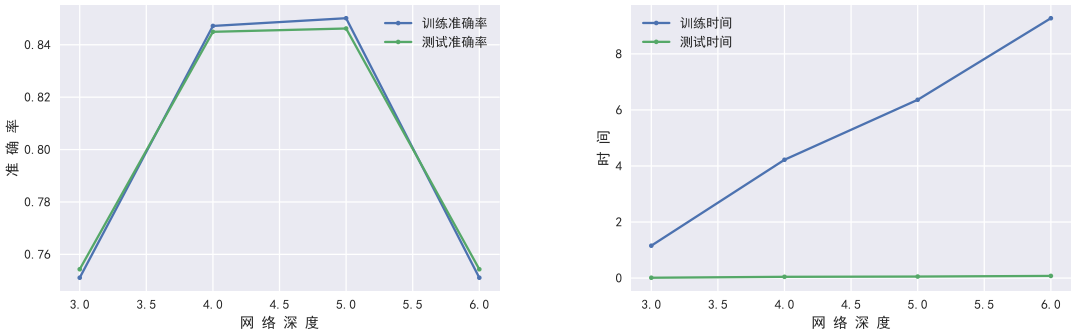

图 25: 网络深度对 MLP 性能的影响

4.3 KNN 分类器

KNN 分类器与其他模型不同，几乎不需要训练，但在测试的时候会消耗大量时间。KNN 中距离的定义会影响整个模型的性能，下面将展开讨论。

4.3.1 距离指数对效果的影响

KNN 中默认的距离为闵可夫斯基距离，其距离的指数默认为 2，也即欧氏距离的定义，下面我们改变指数的取值，测试 KNN 的效果。

图 26 展示了不同距离指数的模型准确率随选取近邻数 K 的变化趋势。可以看到，随着 K 的增加，五种模型的准确率均呈上升趋势，指数为 1 的模型效果最突出。当 K 变得更大的时候，实际上模型准确率不会一直上升，经过测试，当 $K = 1000$ 时，模型的准确率只有不到 0.82。

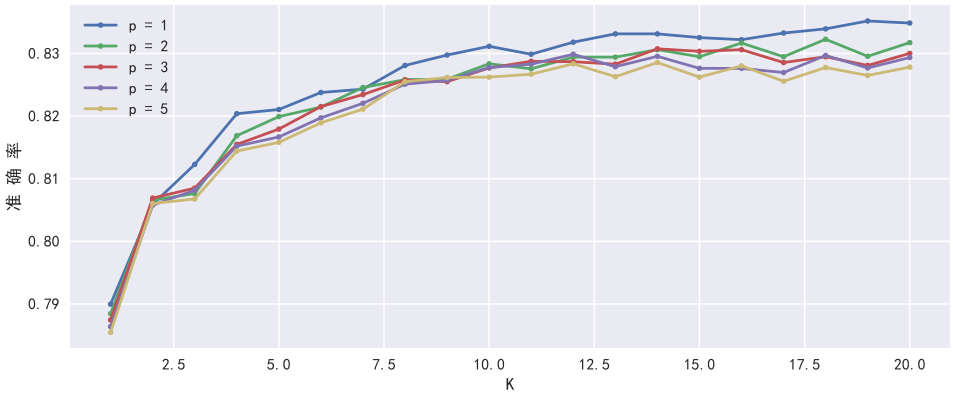

图 26: 距离指数对 KNN 准确率的影响

图 27 展示了训练和测试时间随距离指数的变化趋势。虽然训练时间很相似，但不同指数下的测试时间相差很大，指数不超过 2 时测试时间变化较为缓慢，一旦超过 2，就会快速增长。

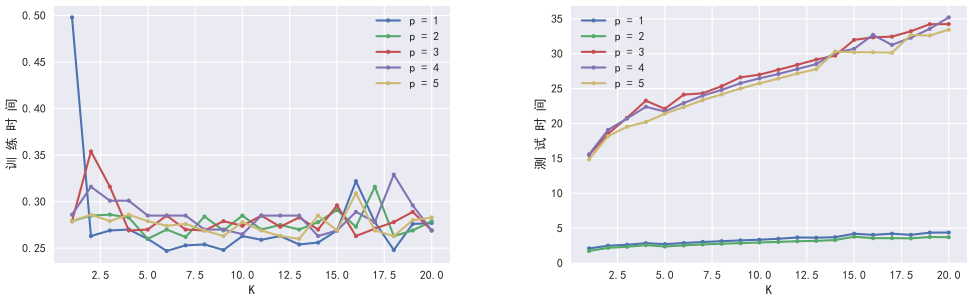

图 27: 距离指数对 KNN 运行时间的影响

4.3.2 算法对运行时间的影响

改变 KNN 的实现算法，测试其运行时间，得到结果如图 28，可以看出，虽然最后的准确率不变，但 BallTree 算法比 KDTree 算法要快很多。

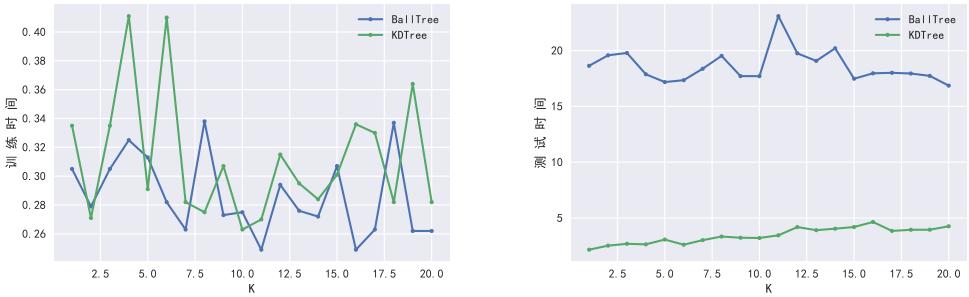

图 28: 算法对 KNN 运行时间的影响

4.4 决策树

决策树是一种运行很快的分类模型，对小数据集很试用，其深度决定了模型的拟合能力。

4.4.1 模型深度对决策树的影响

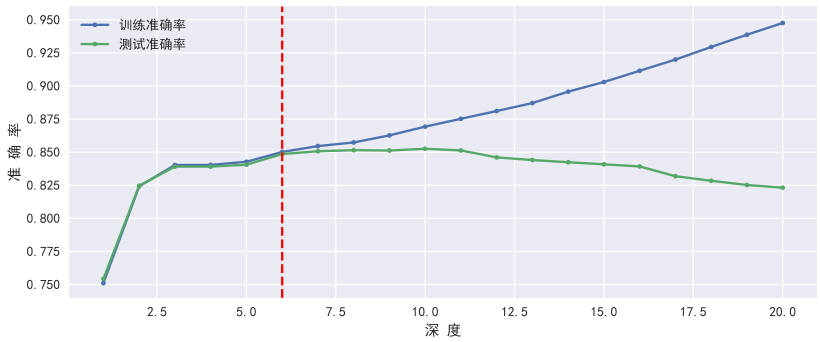

图 29: 模型深度对决策树准确率的影响

图 29 展示了模型准确率随深度的变化，可以看到，当深度超过 6 以后，训练准确率持续走高，而测试准确率缓慢下降，出现了 1 过拟合现象。图 30 展示了模型的运行时间，随着深度的增加，训练的时间成本也逐渐加大。因此，决策树的深度也要适当设置，不宜太深。

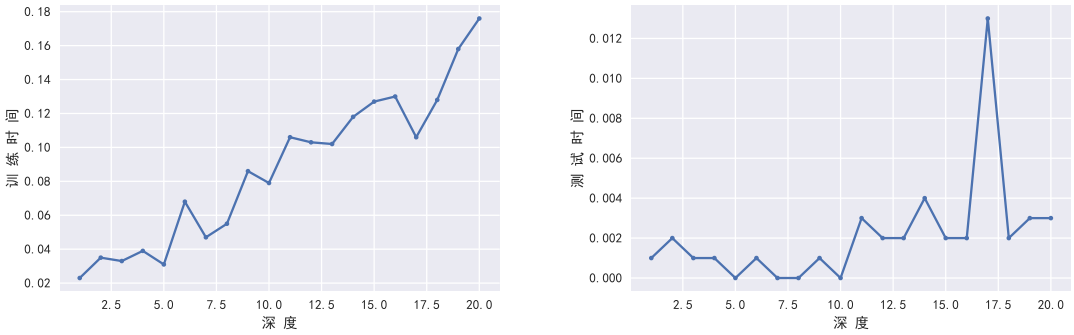

图 30: 模型深度对决策树运行时间的影响

4.5 朴素贝叶斯分类器

表 4 展示了朴素贝叶斯分类器在 Adult 数据集上的表现，可以看出其运算时间极短，训练时没有出现过拟合现象。

表 4: 朴素贝叶斯分类器的性能

	准确率	精确率	召回率	F1 分数	运行时间
训练	0.8031	0.86	0.80	0.82	0.021
测试	0.8025	0.86	0.80	0.82	0.011

4.6 随机森林

将随机森林运用到数据集上，得到了如表 5 所示的结果。与朴素贝叶斯分类器不同，随机森林的分类效果更好了，但过拟合现象严重，训练时间也有所增加。

表 5: 随机森林的性能

	准确率	精确率	召回率	F1 分数	运行时间
训练	0.9995	1.00	1.00	1.00	1.464
测试	0.8495	0.86	0.85	0.85	0.149

5 模型综合对比

本章在上一章的基础上对六种模型进行综合对比，首先对比模型的基本指标，接着比较维度缺失时模型的性能，最后利用交叉验证对比模型的稳定性。

5.1 基本指标对比

表 6 中记录了六种模型的各项性能指标，为了方便对比，将模型的准确率和运行时间作为两个维度绘制成性能散点图，如图 31 所示，越靠左上角表示模型性能越好。

表 6: 模型性能对比

	训练			测试		
	准确率	F1 分数	运行时间	准确率	F1 分数	运行时间
SVM	0.8301	0.85	24.27	0.8295	0.84	6.67
MLP	0.8457	0.85	2.76	0.8434	0.85	0.03
KNN	0.8503	0.86	0.27	0.8349	0.84	6.11
决策树	0.8693	0.09	0.87	**0.8525**	**0.85**	**0.001**
朴素贝叶斯	0.8031	0.82	**0.014**	0.8025	0.82	0.009
随机森林	**0.9993**	**1.00**	1.38	0.8476	0.85	0.14

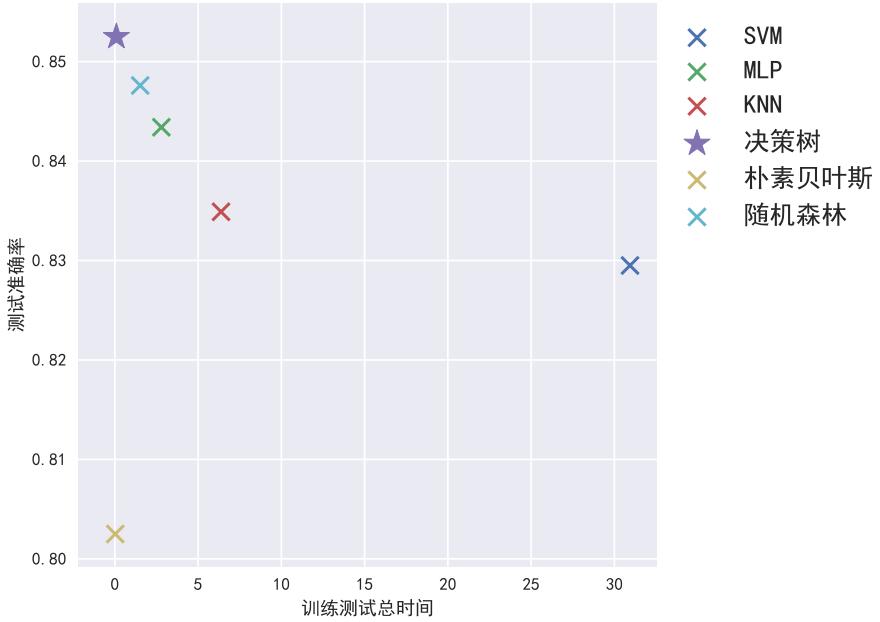

图 31: 模型基本指标对比图

结合图表可以看出，决策树在时间和准确率上都取得了第一；随机森林的过拟合效果明显；SVM 运行时间远远长于其他模型，朴素贝叶斯准确率不如其他几个模型。

5.2 维度缺失对模型性能的影响

之前的数据都是完整的信息，现在我们利用 PCA 对独热向量编码的数据进行降维，测试模型在维度缺失时的性能，结果见图 32。

可以看出，MLP 表现很不稳定，KNN 在极少维度下表现惊人，SVM 上升较快，但训练时间最长，决策树和朴素贝叶斯在低维数据中表现不好。

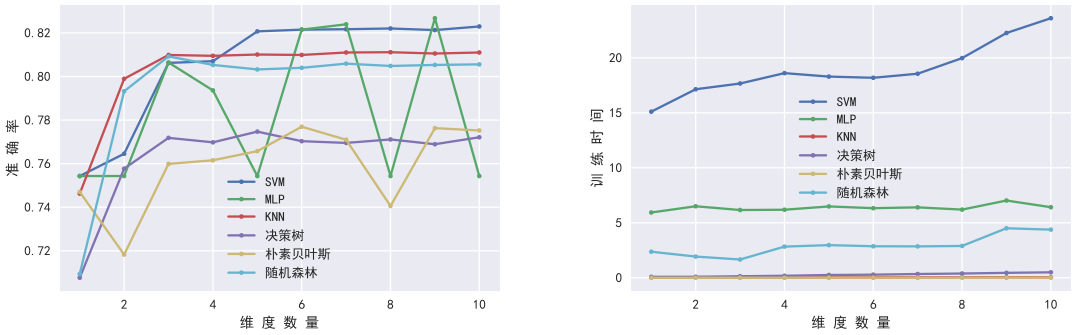

图 32: 维度缺失情况下模型性能对比

5.3 模型稳定性对比

利用训练集对各模型进行十折交叉验证，测得各模型准确率的均值和方差，以比较模型的稳定性，结果见图 33，棒的中心点位置和长度分别代表了准确率的均值和方差。

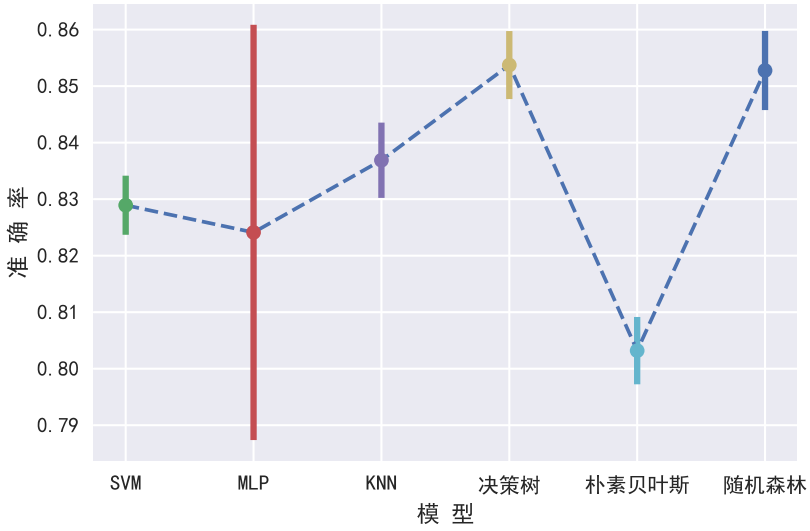

图 33: 模型稳定性对比

可以看出，MLP 的稳定性太弱，而其他模型均比较稳定，不会轻易随着数据的变换而发生巨大改变。

6 结论

本次实验中，我全面探索了 Adult 数据集，首先对数据进行了清洗，接着结合图像讨论了各个属性的分布情况，然后对属性间的隐含关联进行了探究，发现了很多难以想到但又符合常规的关联，比如受教育时间与资本盈亏关联紧密，婚姻状况影响着工作时间，受教育时间决定着职业选择等等。

接着我用六种模型对数据集尝试了分类任务，探究了各种模型参数对性能的影响，最后

通过综合对比，发现决策树在 Adult 数据集上表现最好，达到了 85.25% 的准确率。通过其他对比实验，我还发现了 KNN 和 SVM 在低维数据中表现优秀，MLP 在本次任务中很不稳定。

通过这次实验，我完整的体验了数据分析的全过程，从读数据到清理数据，再到认识变量之间的关系，定义优化的目标，以及最后大量的实验探索和报告总结，这对我是一次难忘的收获。Adult 数据集中有很多值得挖掘的地方，只有不断探索，反复回味，才能发现有趣的结论。感谢老师和助教这学期的帮助，让我对数据科学有了更加深刻的认识。

参考文献

[1] Ron Kohavi. Scaling up the accuracy of naive-bayes classifiers: a decision-tree hybrid. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, page to appear, 1996.

[2] Bernhard E Boser, Isabelle M Guyon, and Vladimir N Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152. ACM, 1992.

[3] Dennis W Ruck, Steven K Rogers, Matthew Kabrisky, Mark E Oxley, and Bruce W Suter. The multilayer perceptron as an approximation to a bayes optimal discriminant function. *IEEE Transactions on Neural Networks*, 1(4):296–298, 1990.

[4] Thomas Cover and Peter Hart. Nearest neighbor pattern classification. *IEEE transactions on information theory*, 13(1):21–27, 1967.

[5] David M Magerman. Statistical decision-tree models for parsing. In *Proceedings of the 33rd annual meeting on Association for Computational Linguistics*, pages 276–283. Association for Computational Linguistics, 1995.

[6] Andrew McCallum, Kamal Nigam, et al. A comparison of event models for naive bayes text classification. In *AAAI-98 workshop on learning for text categorization*, volume 752, pages 41–48. Citeseer, 1998.

[7] Andy Liaw, Matthew Wiener, et al. Classification and regression by randomforest. *R news*, 2(3):18–22, 2002.

A Adult 数据集属性详细介绍

Adult 数据集共有 15 个属性，每个属性的详细介绍如下：

- 年龄：数值
- 工作类别：Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked
- 最终权重：数值
- 教育程度：Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool
- 受教育时间: 数值
- 婚姻状况:Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse
- 职业:Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces
- 家庭关系：Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried
- 种族：White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black
- 性别：Female, Male
- 资本收益：连续数值
- 资本亏损：连续数值
- 每周工作小时数：连续数值
- 祖国：United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands
- 收入：>50K, <=50K

B 数值变量相关性可视化

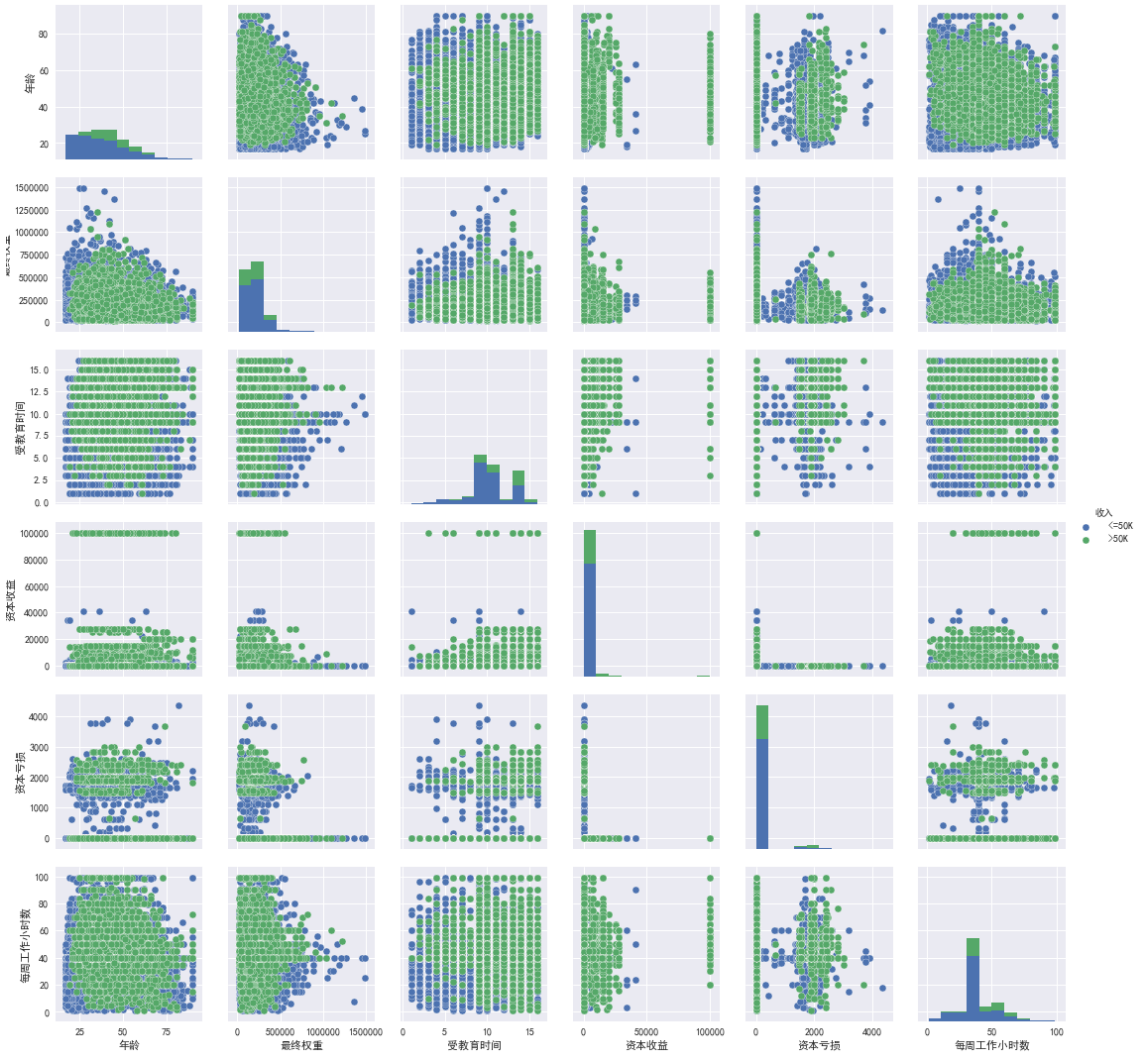

图 34: pairplot