Meets Specifications

恭喜你通过了项目, 继续你的征程吧! 勇敢的战士! 

**探索数据**

**学生正确地计算了下列数值:**

* **记录的数目**
* **收入大于50000美金的人数**
* **收入小于等于50000美金的人数**
* **收入大于50000美金的人数所占百分比**

**{拓展}**

* 从百分数可以看出这里的数据不均衡(imbalanced data), 正类比负类要少, 对于一些算法这可能会导致训练不均衡等问题, 这个博文介绍了解决不均衡数据的一些常用方法:  
  <http://machinelearningmastery.com/tactics-to-combat-imbalanced-classes-in-your-machine-learning-dataset/>
* 另外, 我也想引入一个叫[探索性数据分析(Exploratory Data Analysis)](https://en.wikipedia.org/wiki/Exploratory_data_analysis" \t "_blank)的概念. 有些人甚至认为机器学习中, 数据分析比数学原理更加重要(请看这里 : <http://sharpsightlabs.com/blog/machine-learning-prerequisite-isnt-math/>)

其中一个重要的手段是利用可视化来进行分析:  
除了matplotlib, 为你介绍一个友好的, 美丽优雅的可视化伙伴, [seaborn](https://seaborn.pydata.org/).  
譬如, 这里可以进行[factor plot](https://seaborn.pydata.org/generated/seaborn.factorplot.html?highlight=factorplot" \l "seaborn.factorplot" \t "_blank)来查看高收入阶层在各个特征上是如何分布的 (以婚姻状况为例).

import seaborn as sns

sns.set(style = 'darkgrid', color\_codes = True)

sns.factorplot("income", col = "marital-status", data = data, kind = "count", col\_wrap = 3)

**准备数据**

**学生正确地对特征和目标实现了独热编码。**

**{拓展}**

除了OHE, 也可以尝试一种更紧凑的编码方式，[Label Encoding](http://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.LabelEncoder.html" \t "_blank), 举个例子就是把 ['cat', 'dog', 'fish'] 编码成 [ 1, 2, 3 ], 但是这个方法会引入先验的大小的信息，一些算法，例如神经网络会学习到这中＂优劣＂的信息， 可能会有一定的影响，不过在实际操作中都可以尝试．

**评估模型表现**

**学生正确的计算了简单预测的准确率和F1分数。**

很棒! 你的计算过程很清晰!

**{拓展}**

F-score 其实是precision和recall的[加权调和平均(Weighted Harmonic Mean)](http://wiki.mbalib.com/wiki/%E5%8A%A0%E6%9D%83%E8%B0%83%E5%92%8C%E5%B9%B3%E5%9D%87%E6%95%B0" \t "_blank)

下面公式表示 x\_i 到 x\_n 的加权调和平均:

而Fbeta分数就是为precision赋权 1, 为recall赋权beta^2的调和平均．

[Fbeta.png](https://udacity-reviews-uploads.s3.amazonaws.com/_attachments/42601/1491295370/Fbeta.png)

特别地，当beta = 1时，将得到F1分数:

[F1.png](https://udacity-reviews-uploads.s3.amazonaws.com/_attachments/42601/1491295397/F1.png)

**学生解释了选择这几个模型的原因，并说明了每一个模型的优缺点。**

**{技巧}**

这个问题不好回答, 需要搜寻, 求证, 你做得很不错.

以下提供一些整理算法优缺点的资源以及搜集技巧, 希望对你日后的学习有帮助:

**优缺点**

关于常见模型的优缺点，以下这个页面给了超级简单的总结:  
<http://www.lauradhamilton.com/machine-learning-algorithm-cheat-sheet>  
中文的比较好的资料:  
<http://bigsec.com/bigsec-news/anan-20161111-jiqixuexi>  
其他一些复杂的模型, 比如集成方法的优缺点需要你去一些讨论热烈的地方去寻找, 比如随机森林在Quora就有很好的讨论:  
<https://www.quora.com/When-is-a-random-forest-a-poor-choice-relative-to-other-algorithms>

还有一个方法就是活用页面搜索，在对应算法的维基百科页或sklearn user guide中， ctrl + F. 搜索一些评价算法比较关注的词: overfit, accuracy, bias, time, speed, complexity, generalization等以及它们的不同词性．看看它们是怎么被描述的．

要求更高一点, 你需要对算法的原理流程有更深的理解, 这就需要论文和书籍的的阅读了, 中文书籍推荐李航的<<统计学习方法>>

**应用场景**

可以在百度学术搜索，关键词是相应模型的名称，这样做的好处是你可以在左侧边栏看到不同领域的文章有多少．

**学生成功的实现了一个监督学习算法的流程。**

现在你的流程非常正确!

**学生正确的实现了三个监督学习模型，得出了模型表现可视化的图表。**

**优化结果**

**在考虑了计算成本、模型表现和数据特点之后，学生选出了最好的模型并给出了充足的理由。**

**学生能够用清晰简洁的话来向一个没有机器学习或任何其他技术背景的人来解释最优模型的工作原理。**

这里其实你已经做得非常好了, 要点都已经涵盖了. 不过你的局部最优local optima) 是不正确的, 梯度下降的目的都是要找到全局最优 (global optima). 尽管一般都找不到.

参考阅读:  
<https://cn.mathworks.com/help/optim/ug/local-vs-global-optima.html>

机器学习不仅仅需要建立模型, 在现实应用还需要将它呈现给客户或者其他领域的合作者看, 让他们对这个结果充分了解.

以下是一个Kaggler关于这方面的见解.  
<http://blog.kaggle.com/2016/06/13/communicating-data-science-an-interview-with-a-storytelling-expert-tyler-byers/>

最后, 你可以多使用 markdown 技巧, 帮助你良好地分段. 你可以模仿题目书写的markdown或者花20来分钟看看这个帖子快速学习:  
<http://discussions.youdaxue.com/t/markdown/24241>

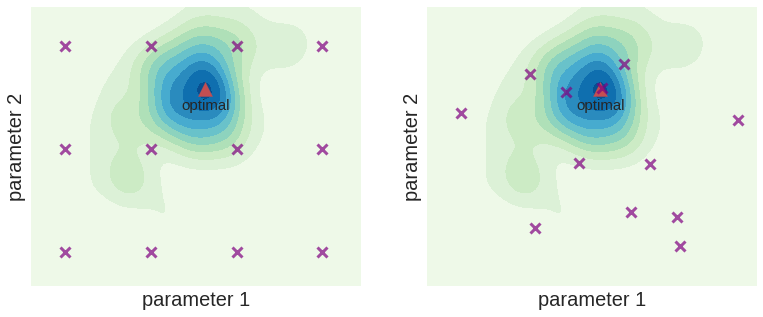
**最终模型利用了网格搜索进行参数调优，至少挑战了一个参数，并且至少有三个可选值。如果模型参数不需要任何调整，学生需要给出明确的理由。**

这边C参数选择10的幂来调参会更有效率.

{'C' : [1e-1, 1e-2, 10, 1e2, 1e3, 1e4]

**{拓展}**

在这里为你介绍另一种比较流行的调参方法, 它叫[随机搜索(Randomized Search](http://scikit-learn.org/stable/modules/generated/sklearn.model_selection.RandomizedSearchCV.html" \t "_blank). 就是在给定的参数空间中随机的进行参数组合, 搜索. 这个调参方法往往比网格搜索更有效. 下图小山顶表示最优, 参数组合被表示成了平面上的点, 可以看出随机化的选取参数是让参数组合散落到最优附近的一个更有效策略.

[](https://udacity-reviews-uploads.s3.amazonaws.com/_attachments/42601/1493364960/grid_rand.png)  
以下这篇文章图文并茂地解释了两种算法. 鼓励你去看一下.  
<https://medium.com/rants-on-machine-learning/smarter-parameter-sweeps-or-why-grid-search-is-plain-stupid-c17d97a0e881>  
还有sklearn比较两者的专页 (很多即改即用的代码):  
<http://scikit-learn.org/stable/auto_examples/model_selection/randomized_search.html#sphx-glr-auto-examples-model-selection-randomized-search-py>

**学生在表格中正确汇报了调优过后、调优之前以及基准模型的准确率和 F1 分数。学生把最终模型的结果与之前得到的结果进行了对比。**

**特征重要性**

**学生列出了他们认为对预测个人收入最重要的5个特征，同时给出了选择这些特征的理由。**

**学生调用了一个监督学习模型的 feature\_importances\_ 属性。此外，学生列出了这些重要的特征并讨论了这些特征的相同点和不同点。**

**学生用最重要的5个特征建模并分析了和对比了改模型与问题五中的最优模型的表现。**













