

墨尔本大学计算机和信息系统学院  
COMP90049，机器学习简介，第一学期 2023年

作业3：工作薪金预测

**发布：** 2023年4月17日，星期一。

**到期：第一阶段：** 5月19日星期五下午5点  
**第二阶段：** 5月24日星期三下午5点

**分数：** 该项目满分为30分，将占总分的30%。

## 1 概述

在这项任务中，你将开发并批判性地分析预测**工作薪金**的模型。也就是说，给定一个职位描述，你的模型将预测为该职位提供的薪水。你将得到一个标有工资的职位描述的数据集。此外，每个职位描述都标有其所属职业类别的性别平衡：*男女平衡 (0)* 或 *以男性为主 (1)* 或 *以女性为主 (2)*。例如，"工程"类职业往往以男性为主，而与教育有关的职业往往是女性多于男性。你可以利用这些额外的信息来研究你的模型在不同性别的职业中是否同样有效。该评估为你提供了一个机会，在一个开放式研究问题的背景下思考机器学习的概念，并加强你的数据分析和解决问题的技能。

这项任务的目标是**批判性地评估**各种机器学习算法在确定工作薪水问题上的**有效性**，并在**一份技术报告中表达你所获得的知识**。这个项目的技术层面将涉及将适当的机器学习算法应用于数据，以解决这个任务。将会有有一个Kaggle课内竞赛，你可以和你的同学比较你的算法的性能。

项目的重点是报告，格式为简短的研究论文。在报告中，你将以合理的知情读者能够理解的方式，展示你所获得的知识。

## 2 可交付的成果

**第一阶段：** 模型开发和测试以及报告撰写（5月19日到期）：

1. 一个或多个用Python编写的程序，包括重现你报告中的结果所需的所有代码（包括模型实现、标签预测和评估）。你还应该包括一个README文件，简要地介绍你的实现。通过Canvas提交。

2. 一份匿名的书面报告，2000字（ $\pm 10\%$ ），**不包括**参考文献列表。你的名字和学生证**不应该**出现在报告的任何地方，包括元数据（文件名，等等）。通过Canvas/Turnitin提交。**你必须将报告作为一个单独的PDF文件上传。**不要把它作为压缩档案（zip，tar，...）文件的一部分或以其他格式上传。
3. 对提交给Kaggle的工作描述测试集的预测<sup>1</sup>第7节中描述的课内竞赛。

**第二阶段：同行评审（5月24日到期）：**

1. 评论你的同学写的两份报告，每份200-400字。通过Canvas提交。

### 3 数据集

您将获得

- 一个由13,902个职位描述组成的训练集。前8,000份描述被贴上了工作的薪水（目标标签）和工作的职业组的性别平衡（人口标签）。其余的5,902个描述是**没有标签的**。你可以将这些用于半监督或无监督的学习方法。
- 一个由1,738个有标签的工作描述组成的开发集，有目标和人口标签，你应该用这些标签来选择和调整模型；
- 一个由1,737个职位描述组成的测试集，没有目标（但有人口统计学）标签，这将用于Kaggle课内竞赛的最终评估

#### 3.1 目标标签

这些是你的模型应该预测的标签（ $y$ ）。我们以两种形式提供这个标签：

- 平均预期工资（浮动；在原始数据\*.csv文件中名为mean\_salary的列中）；以及
- 表示薪资区间的分类标签，我们将平均薪资分成10个等频区间（在原始数据\*.csv文件中名为salary\_bin的列）。

你可以在你的实验中使用任一标签表示，但不同的表示可能需要不同的机器学习方法。

#### 3.2 人口统计学标签

人口学标签提供了关于招聘广告的职业类别中的性别倾斜的额外元信息（在原始数据\*.csv文件中名为gender\_code的列）。它们**只能**用于评估特定的雇员子群（男性对女性）的模型，但**不能被预测**（可能也不能作为特征使用，尽管你可以在报告中讨论这个问题）。在所提供的数据集中，每个招聘广告都被贴上了三种可能的人口统计学标签之一，表明该工作的职业类别的性别平衡：

- 0: 性别平衡的职业类别（例如，顾问、会计师.....）。

---

<sup>1</sup><https://www.kaggle.com/>

- 1: 以女性为主的职业类别（例如，护士、社会工作者.....）。
- 2: 以男性为主的职业类别（例如，工程师、建筑工人...）。

### 3.3 特点

为了帮助你进行初始实验，我们从原始作业描述中创建了不同的**特征表示**。你可以在你的实验中使用下面描述的任何子集，如果你愿意，你也可以从原始描述中设计自己的特征。所提供的表示方法是


I. 原始描述以单个字符串表示。我们把所有的词都用小写字母表示，并去掉结尾部分。例如、

*"制定实施评估健康营养方案，制定互动健康方案内容活动"*


原始数据中的requirements\_and\_roles列中提供了纯文本的工作描述。  
\*.csv文件）。

II. **TFIDF** 我们对招聘广告进行了术语频率-反文档频率的预处理，以进行特征选择。特别是，我们(1)删除了所有的停顿词，(2)只保留了完整的原始职位描述数据集中具有最高TFIDF值的500个词。因此，每个招聘广告现在被表示为一个500维的特征向量，每个维度对应于500个词中的一个。如果该词没有出现在职位描述中，则其值为0；如果该词出现在描述中，则为该词的TFIDF得分。请注意，由于工作描述很短，大多数值都是0.0。例如、


[0.0, 0.0, 0.0, 0.0, ... 0.998, ..... 0.0]



一个500维的数字  
列表



职务说明中单词的  
TFIDF得分



工作描述中没有的词

特征选择讲座和相关代码（第5周）提供了更多关于TFIDF的信息，以及Schu"tze等人（2008）。

文件tfidf\_words.txt包含500个TFIDF值最高的词，以及它们在向量表示中的索引。你可以使用这些信息进行模型/错误分析。

III. **嵌入** 我们将每个职位描述映射到一个384维的嵌入，该嵌入是用预先训练好的语言模型计算出来的，称为句子转化器（Reimers and Gurevych, 2019）。<sup>2</sup>这些向量涵盖了每个招聘广告的"意义"，因此类似的招聘广告将在384维空间中被紧密联系在一起。例如：、

[2.05549970e-02 8.67250003e-02 8.83460036e-02 -1.26217505e-01 1.31394998e-02 . .]

384维的数字列表

<sup>2</sup><https://pypi.org/project/sentence-transformers/>

**数据格式** 原始数据以csv格式提供（train.csv, valid.csv和test.csv）。标签数据集也包含目标标签格式（实值和bin）和人口学标签，有以下几列：

job\_id 每个实例的唯一标识符 requirements\_and\_role 输入特征（原始工作描述文本） salary\_bin 目标标签（分类分档）。  
平均工资目标标签（浮动）。  
gender\_code 人口学标签（分类）。

**TFIDF和Embeddings**的表示方法是以密集numpy矩阵的形式提供的（文件结尾为\*.npy）。<sup>3</sup>同一数据集类型的行号指的是同一实例，例如，train.csv中的第5行、train-embeddings.npy和train-tfidf.npy是同一招聘广告的不同表示。

## 4 项目阶段一

你应该制定一个研究问题（下面提供两个例子），并开发机器学习算法和适当的评估策略来解决研究问题。

你应该在你的报告中**最少实现和分析一个基线**，以及**至少两个不同的机器学习模型**。**注意：我们更感兴趣的是你对方法和结果的批判性分析，而不是你模型的原始性能**。你可能无法对你的研究问题得出一个明确的答案，这完全没有问题。但是，你应该对你的（可能是负面的）结果进行深入的分析 and 讨论。

### 4.1 研究问题

你应该在你的项目中解决**一个研究问题**。我们在下面提出了两个研究问题，供大家参考，但你也可以提出你自己的问题。你的报告应明确说明你的研究问题。解决一个以上的研究问题**并不能**获得高分。我们更感兴趣的是你对方法和结果的**批判性分析**，而不是对更多内容或材料的覆盖。

#### 研究问题1：无标签数据是否能改善工作薪金的预测？

各种机器学习技术已经（或将要）在这个主题中讨论过了（Naive Bayes, 0-R, clustering, semi-supervised learning）；还有很多。这些算法需要不同程度的监督：有些是监督的，有些是无监督的，有些则结合了两种策略。开发机器学习模型，利用不同数量的监督，使用训练数据集的有标记和无标记部分。你可能还想尝试不同的特征表示（第3.3节）。另外，你可能想开发有意义的方法来预测平均工资与工资档次（不同的Y标签代表）。我们强烈建议你在这个项目的尝试中使用机器学习软件和/或现有的库（如sklearn或scipy）。不同的机器学习范式的优势和劣势是什么？你能有效地结合有标签和无标签的训练数据吗？

---

<sup>3</sup>在这里了解如何阅读和处理这些文件：<https://numpy.org/doc/stable/reference/generated/numpy.load.html>。

## 研究问题2：探索工作薪金预测的偏见

分别比较不同的模型和/或特征表示在数据集中不同人口群体（平衡工作与女性为主的工作与男性为主的工作）上的表现。有些模型对所有这些群体的效果并不一样。批判性地分析这一差距，并尝试在本课题所涉及的概念范围内解释这一差距。你能调整你的模型以缩小业绩差距吗？如何做到？**注意：**你的成绩并不取决于你在缩小成绩差距方面的成功。有趣的是，深入分析的失败尝试是完全可以接受的。

### 4.2 特征工程（可选）

我们在本课题中讨论了三种类型的属性：分类、顺序和数字。这三种类型都可以为给定的数据构建。有些机器学习架构更喜欢数字属性（如k-NN）；有些机器学习架构在分类属性方面效果更好（如多变量奈何贝叶）--你可能会通过实验观察到这一点。

**你可以选择根据**原始职位描述数据集设计一些属性（并可能使用它们来代替--或与我们提供的特征表示一起使用）。或者，你可以简单地从我们为你生成的特征中选择特征（tfidf，和嵌入）。

### 4.3 评价

你们学习者的目标将是预测未见过的数据的标签。我们将使用一个**保持策略**。数据收集被分成三部分：一个训练集、一个开发集和一个测试集。这些数据可以在LMS上找到。

为了让你有可能在测试集上评估你的模型，我们将设立一个**Kaggle课内竞赛**。你可以在那里提交测试集的结果，并获得关于你的系统性能的即时反馈。这里有一个排行榜，可以让你看到与其他在线参赛的同学相比，你的表现如何。

### 4.4 报告

你将提交一份长度为2000字（ $\pm 10\%$ ）的**匿名**报告，**不包括**参考文献列表。报告应遵循短篇研究论文的结构，正如在学术写作的客座讲座中讨论的那样。它应该在你选择的研究问题的背景下描述你的方法和观察，包括工程（可选）特征，以及你尝试的机器学习算法。它的主要目的是为读者提供有关该问题的**知识**，特别是对**你的结果和发现的批判性分析**。众所周知的机器学习模型的内部结构，只有当它对连接理论和你的实际观察很重要时，才应该被讨论。

- 引言：对问题和数据集的简短描述，以及所涉及的研究问题
- 文献综述：对一些相关文献的简短总结，包括数据集参考文献和至少两篇你选择的其他相关研究论文。你可以在本文件的参考文献列表中找到灵感。我们鼓励你搜索其他的参考文献，例如在本文件所引用的论文中引用的文章。

- 方法：确定新设计的特征，以及包括这些特征的理由（可选）。解释你所使用的ML模型和评估指标（以及你为什么使用它们）。
- 结果：以评价指标的形式介绍结果，最好还有说明性的例子。强烈建议使用表格和图示。
- 讨论/批判性分析：根据对主题材料的理解以及研究问题的背景，在\*\*，系统的行为。
- 结论：清楚地展示你对问题的识别知识
- 书目，其中包括Bhola等人（2020），以及你在项目中使用的任何其他相关工作的参考文献。我们鼓励你使用APA 7的引用风格，但也可以使用不同的风格，只要你在报告中保持一致。

\*\* 意思是说，我们更想看到的是你对任务的思考，以及确定不同方法的相对表现的*原因*，而不是你选择的不同方法的原始分数。这并不是说你应该忽视不同的数据运行的相对性能，而是说你应该超越简单的数字，思考支撑这些数字的原因。

我们将提供LATEX和RTF风格的文件，我们希望你撰写报告时使用这些文件。**报告应以单个PDF文件的形式提交。**如果报告以PDF以外的任何格式提交，我们保留以0分退回报告的权利。

你的名字和学生证**不应该**出现在报告的任何地方，包括任何元数据（文件名等）。如果我们发现任何这样的信息，我们保留将报告以0分退回的权利。

## 5 项目第二阶段

在审查过程中，你将阅读两份由你的同学提交的匿名材料。这样做是为了帮助你思考处理项目的一些其他方法，并确保每个学生都能收到一些额外的反馈。你的目标是每篇评论共写150-300字，对三个“问题”做出回应：

- 用一段话（50-100字）简要概括作者的工作内容
- 用一段话（50-100字）指出你认为作者做得好的地方，以及为什么？
- 用一段话（50-100字）指出你认为可以改进的地方，以及为什么？

## 6 评估标准

该项目满分为30分，占你该科目总分的30%。分数将是：

### 报告质量：（26/30分）

你可以参考Canvas/Assignment 3页面上的评分标准，其中详细说明了我们将在报告中寻找的内容。

### Kaggle: (2/30 分)

用于向Kaggle竞赛提交（至少）一组模型预测。

### 评论：(2/30分)

你将为其他学生写的两份报告各写一篇评论；你将遵循上述准则。

## 7 使用Kaggle

**任务** Kaggle比赛的内容是预测**工资水平**（不是平均工资）。

**说明** Kaggle课内竞赛的网址将很快在LMS上公布。要参加比赛，请按以下步骤进行：

- 每个学生都应该用你的学生身份创建一个Kaggle账户（除非他们已经有一个）。
- 你每天最多可以提交8次。在Kaggle网站上可以找到一个提交文件的例子。
- 提交的数据将由Kaggle评估其准确性，只针对30%的测试数据，形成公共排行榜。
- 在比赛结束前，你可以从之前提交的作品中选择一个最终作品--在默认情况下，Kaggle会选择公共排行榜上得分最高的作品。
- 比赛结束后，公开的30%测试分数将被私人排行榜上的100%测试分数所取代。

## 8 分配政策

### 8.1 数据使用条款

该数据集来自[Bhola等人（2020）](#)发表的资源：

Akshay Bhola, Kishaloy Halder, Animesh Prasad, and Min-Yen Kan. 2020.从工作描述中检索技能：一个基于语言模型的极端多标签分类框架。载于第28届国际计算语言学会议论文集，第5832-5842页，西班牙巴塞罗那（在线）。International Committee on Computational Linguistics.

该参考文献**必须**在书目中引用。由于违反了使用条款，我们保留对任何缺乏此参考文献的提交文件标记为0的权利。

人口统计学标签是根据[新加坡统计局](#)的数据添加的。

请注意，该数据集是发布在万维网上的实际数据的样本。因此，它可能包含一些品味低下的信息，或可能被认为是攻击性的信息。我们会要求你尽可能地超越这一点，把注意力放在手头的工作上。如果你对这些条款有异议，请尽快与我们联系（[lea.frermann@unimelb.edu.au](mailto:lea.frermann@unimelb.edu.au)）。



## 项目规格的变更/更新

我们将使用Canvas公告任何大规模的变化（希望没有！）和Ed公告小的澄清。通过Canvas对项目规范进行的任何补充将取代本版本规范中的信息。

## 逾期提交政策

为确保同行评审工作顺利进行，**不允许逾期提交**。提交将于**5月19日下午5点**结束。对于那些明显无法及时提交完整解决方案的学生，我们可能会提供延期，但请注意，在这种情况下，你可能无法参与同行评审过程并从中受益。我们将根据具体情况寻求解决方案。请给 Hasti Samadi（[hasti.samadi@unimelb.edu.au](mailto:hasti.samadi@unimelb.edu.au)）发电子邮件，并附上延迟原因的文件。

## 学术不端行为

对于大多数学生来说，与同伴讨论想法将成为这个项目的自然组成部分。然而，这仍然是一项个人任务，因此重复使用想法或过度影响算法的选择和开发将被视为作弊。我们强烈建议你在本科目的Canvas上（重新）学习学术诚信培训模块。我们将检查提交的材料是否具有原创性，并将援引大学的学术不端行为政策<sup>4</sup>如果认为发生了不适当的串通或剽窃行为，我们将援引大学的学术不端行为政策。由生成性人工智能（包括但不限于ChatGPT）产生的内容不是你自己的作品，根据**大学的政策**，提交这样的内容将被视为学术不端行为的案例。

## 参考文献

Bhola, A., Halder, K., Prasad, A., and Kan, M.-Y. (2020). 从工作描述中检索技能：一个基于语言模型的极端多标签分类框架。 In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5832-5842, Barcelona, Spain（在线）。 International Committee on Computational Linguistics.

Elazar, Y. and Goldberg, Y. (2018). 从文本数据中去除人口属性的对抗性。 In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 11-21, Brussels, Belgium. 计算语言学协会。

Han, X., Shen, A., Cohn, T., Baldwin, T., and Frermann, L. (2022). 预测性公平性的系统评估。 见《*计算语言学协会亚太分会第二届会议暨第十二届自然语言处理国际联合会议论文集*》（第一卷：长论文），第68-81页，仅在线。 计算语言学协会。

Joshi, M., Das, D., Gimpel, K., and Smith, N. A. (2010). 电影评论和收入：文本回归的实验。 在 *人类语言技术中：The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 293-296, Los Angeles, California. 计算语言学协会。

---

<sup>4</sup><http://academichonesty.unimelb.edu.au/policy.html>

Reimers, N. and Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 在2019年自然语言处理经验方法会议和第九届自然语言处理国际联合会议 (EMNLP-IJCNLP) 论文集中, 第3982-3992页, 中国香港。计算语言学协会。

Schütze, H., Manning, C. D., and Raghavan, P. (2008). 信息检索简介, 第39卷。剑桥大学出版社剑桥。