报告写作说明

在每个作业中，您都被要求编写一份报告，其中展示您的数据/结果并讨论它们。这些报告的主要目的是(i)明确地表明您了解您正在使用的工具，(ii)可以批判性地评估数据和结果。仅仅运行一个程序并引用它输出的数字是不够的。此外，虽然你将使用软件作为一个工具来执行计算，可视化数据等，通常你将主要结合代码片段在讲座和软件是一个相对较小的工作的一部分你所做的工作（这是一种手段，而不是结束本身）。考虑到这一点，这里有一些关于写好报告的建议。

- *数值结果*。在提供数值数据时，重要的是：
  - 描述如何计算这些值，以及它们代表了什么。如果你只是提供了一组数字，例如[0.1,0.5，-0.6]，没有进一步的评论，那么我只能猜猜它们是什么意思。相反，你可能会说你“训练了一个线性模型$y=\theta_1 x_1+\theta_2 x_2+\theta_3 x_3$有三个特征，训练后的参数值为$\theta_1 = 0.1,\ \theta_2 = 0.5,\ \theta_3 = -0.6$”.
  - 评论/讨论数据。E.g.在上面的例子中，你可能会注意到，“自从$\theta_2$以来”大于$\theta_1$然后由模型在特征x上放置更大的权重比特性x上要多，自$\theta_3$以来是负的，然后是特征x的增加会导致产量y下降。“。根据上下文，您可能会简要地评论这种行为是否合理/预期。
- 绘图数据。良好地可视化数据对于帮助理解数据很重要。在绘制数据时，您应该：
  - 花点时间思考如何以一种清晰的方式呈现数据——有时这需要一些尝试和错误。
  - 确保线足够厚，散点图中使用的标记足够大，很容易看到。选择颜色和标记的形状，使情节易于阅读。
  - 始终标记轴。
  - 确保所有文本（包括轴标签/刻度号、任何图例等）都足够大，以便清晰可见
  - 在支持文本中(i)描述图中的值是如何计算的以及它们代表什么，(ii)评论/讨论绘制的数据（类似于显示数值数据时，参见上文）。

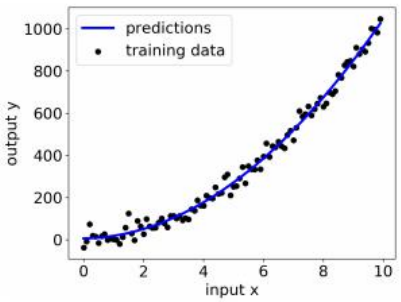



图1：示例图

  - 图1显示了一个良好实践的例子。注意两个轴都有标记，文本清晰。黑点表示训练数据和来自机器学习模型的训练数据和蓝线预测。通过使用不同的颜色，数据更容易阅读，线和点足够大到很容易阅读。由于训练数据是有噪声的，并且由单个数据点组成，因此最好将其绘制为一个单独的点（即作为一个散点图），而不是将这些点与一条线连接起来。没有必要把情节本身变大。
- *执行计算的代码*。把代码视为一个能产生神秘数字的黑盒子是一个严重的错误。像滑雪学习这样的包使这个v很容易做，所以你需要做一个积极的努力来避免这种情况。

- 当您使用技能学习函数时，请确保您了解输出值表示什么以及如何计算它们。例如，一些技能学习函数生成“分数”值，如果你使用其中一个值，那么你理解如何计算分数值是非常重要的，即它是否是均方误差、分类精度等等。你必须在报告中说明这一点，否则我就不知道你是否明白它真正的价值是什么。
- 有时你会被要求从头开始实现机器学习计算，而不是使用技能学习。在这种情况下，您需要在报告中讨论/解释代码是做什么，例如。通过在你的报告中包含相关的代码片段，以及解释它做什么的文本。不要说“查看代码”，并保留它，即使代码包含注释。
- 不需要解释用于生成图或输入数据的代码，需要解释的仅仅是进行数值计算的代码。
- 请使用有意义的变量名等保持代码的简短和干净。

- 讨论。通常在作业中，你会被要求讨论/批判性地评估数据和结果。然后有必要表达一个意见，而且通常相同的数据可以支持不同的意见，这样就没有“正确”的答案。

  - 因此，在发表意见时，需要记住的关键问题是提出由数据/结果支持的支持性论点，以证明该意见是合理的。仅仅陈述一个观点本身，例如。“我认为模型a比模型B更好”，没有支持的论点，几乎没有价值。
  - 将讨论看作是一个演示您对如何解释您正在使用的工具/技术的输出的理解的机会，以及一个实践使用数据来支持您的推理的机会。
  - 在讨论方面，时间越长通常也好不到哪里去。“大脑倾倒”是缺乏理解的明显症状，长而散漫的文本也是如此。相反，目标是一个简短的、集中的讨论，清楚地说明你想要提出的要点，如果你喜欢它，要点格式是可以的。

- 报告长度。作为一个粗略的指南，一个作业报告应该大约有5页长。如果你发现自己超过10页就太长了——也许你可以绘制多个曲线来减少数字的数量和节省空间，也许你的讨论/描述太长和冗长，需要提炼，也许文本格式过度利用空间。

- 报告格式。请使用pdf格式的报告，并使用输入的文本，而不是手写的。在pdf文档的附录中包含代码，以便我们很容易检查。此外，请还在一个单独的zip文件中提交代码和数据，我们可以运行该文件来检查报告中的结果。

都柏林三一学院
计算机科学与统计学院

**模拟每周分配** CS7CS4/CSU44061机器学习

**游戏规则：**

- **可以与他人讨论，但不要显示您写给其他人的任何代码。你必须用你自己的语言写答案，并完全由你自己编写代码。所有提交的作品都将被检查是否有剽窃。**
- **报告必须输入（请不要手写答案），并作为黑板上单独的pdf提交（请不要作为zip文件的一部分）。**
- **重要提示：对于每个问题，你的主要目的是清楚地说明你理解你在做什么——而不仅仅是运行一个程序并引用它输出的数字。冗长而散漫的答案和“大脑转储”并不是实现这一目标的方法。如果您编写代码来执行计算，您需要讨论/解释该代码会做什么，如果您提供数值结果，您需要讨论它们的解释。一般来说，大部分的功劳是为了解释/分析，而不是代码/数字答案。说“看到代码”还不够好，即使代码包含注释。同样，没有进一步评论的独立数字或图也不够好。**
- **当您的答案包含一个图时，请确保(i)标记轴，(ii)确保所有文本（包括轴标签/蜱虫）都足够大以清晰可见，并(iii)在文本中说明图所显示的内容。**
- **在您提交的pdf报告中包含作为附录为作业编写的代码来源。还包括一个单独的邮政文件，包含可执行代码和所需的任何数据文件。程序应该运行用Python编写的代码，并且应该在运行时加载数据等，这样我们就可以解压缩您提交的文件，并直接运行它来检查它是否工作。使用有意义的变量名等保持代码简短和干净。**
- **报告通常应为5页左右，上限为10页（不包括带代码的附录）。如果你浏览了超过10页，那么额外的页面将不会被标记。**

下载数据集

- 从<url>下载分配数据集重要信息：您必须获取自己的数据集副本，不要使用其他人下载的数据集。.
- 请剪切并粘贴数据文件的第一行（以#开头），并包含在标识数据集的提交中。
- 数据文件由两列数据（加上第一条头行）组成。第一列是输入特征，第二列是目标值。

.

分配

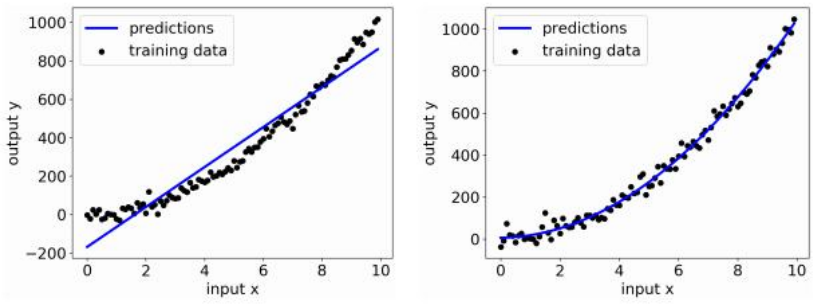
(a)(i)使用sklearn的线性回归函数来训练数据上的线性模型。将训练数据和预测绘制在同一图上，以便于进行比较。并给出了训练模型的参数。讨论。

(ii)现在通过包含第二个特征等于第一个特征的平方来增加你的模型。训练这个新模型并讨论它的预测

模型解决方案

(a)(i)图2(a)显示了线性模型y=θ的训练数据和预测$_0$+ θ$_1$x使用sk学习的线性回归函数进行训练，其中x是输入特征，y是输出预测。训练模型的参数值为θ$_0$ = 158.12, θ$_1$ = 102.请参见下文(ii)中的讨论。可以看出，训练数据遵循一条曲线。线性模型的预测是通过这些曲线数据的最佳拟合直线，但由于预测不能捕捉到曲线的形状，它们显示出一致的误差，特别是在数据的边缘和中心。

图2：训练数据加上(a)线性模型和(b)二次模型的预测图




(a) linear model (b) quadratic model

(ii)使用numpy：将nunpy导入numpy作为np

Xtrain_poly=np.column_stack((Xtrain, Xtrain* * 2 ))

模型=Line=和()。f（Xtrain_poly和Xtrain_poly）

这里Xtrain是一个包含训练数据输入的1000 1 向量，它只是一个特征，所以Xtrain的每一行对应一个训练数据点)。然后使用column_stack构造一个1000 2 矩阵Xtrain_poly，其中每一行再次对应一个训练数据点，但现在有两个输入特征。图2(b)显示了该扩展模型y=θ生成的预测$_0$+ θ$_1x$+ $\theta_2x^2$可以看出，由于添加了第二个二次特征，该模型现在能够捕获数据中的曲线，因此预测更加准确。训练模型的参数值为θ$_0$ = 7.00, θ$_1$ = 1.88, θ$_3$ = 10.21.添加二次特征可以得到截距参数θ$_0$和重量θ$_1$分配给原始输入特性x的值变得要低得多（以前它们分别是-158.12和102.97）。

附录

从sklearn开始。inear_model导入一个模型=一个()。f（Xtrain，
ytrain）
模型(模型。intercept_，模型。
coef_)ypred=模型。it(Xtrain)
导入matplotlib。作为pltplt。中华民
国(“字体”，size=18)
p l t .如果我喜欢你的话。constrained_layout.使用
‘]=Trueplt。(Xtrain，你的火车，　　或=的黑色‘)
p l t .(Xtrain，ypred，=的蓝色，还有=3)。xlabel（“输
入x”）；plt。y输出(“输出y“)
p l t .legend（“数据”，“数据”）plt。显示()

将数字导入为np
Xtrain_poly=np.column_stack((Xtrain，Xtrain* * 2 ))
模型=Line=和()。Xtrain_poly(Xtrain_poly，Xtrain_poly)，模
型。intercept_，模型。coef_)

ypred=模型。p r e d i c t(Xtrain_poly)
p l t .(Xtrain，你的火车，　　或=的黑色‘)
p l t .(Xtrain，ypred，=的蓝色，还有=3)。xlabel（“输入x”）；plt。y输出(“输出y“)
p l t .legend（“数据”，“数据”）plt。显示()