姓名：毛九弢　学号：221900175　2024 年 12 月 22 日

一. (20 points) 降维

1. 基于 numpy 和 fetch_lfw_people 数据集实现主成分分析（PCA）算法，不可以调用 sklearn 库，完成下面代码并且可视化前 5 个主成分所对应的特征脸 (10 points)

解:

代码见 Prob1/task1/task1.py 和 Prob1/task1/task1Check.py, 结果见 Prob1/task1/out

可视化前 5 个主成分所对应的特征脸如下：

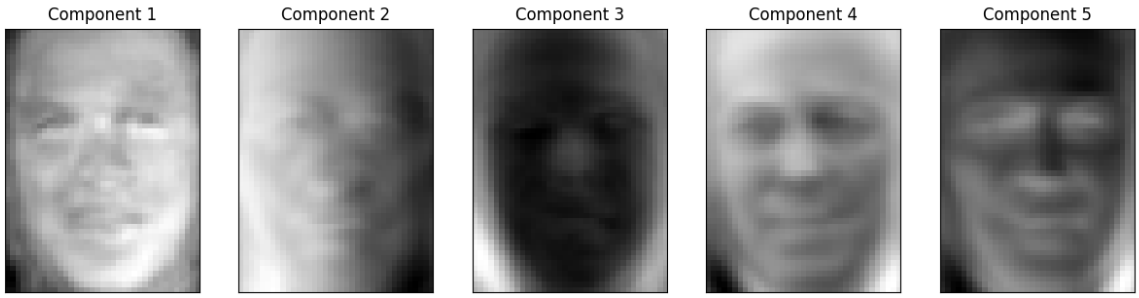


图 1: first 5 eigenfaces

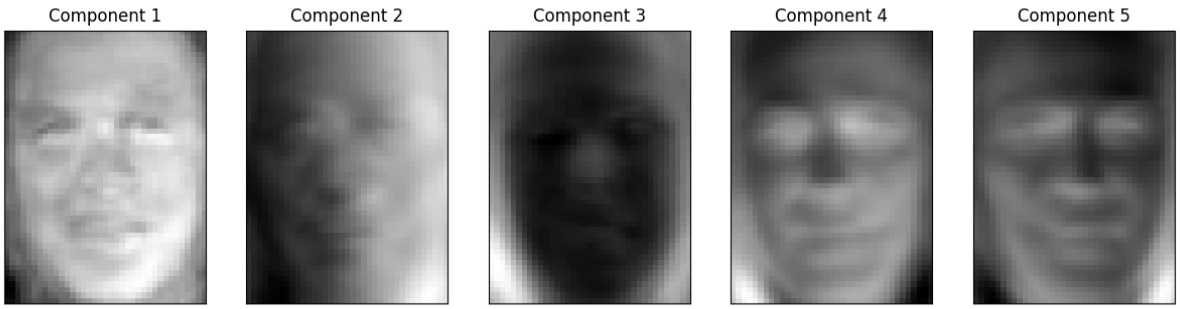


图 2: first 5 eigenfaces(by sklearn as a reference)

2. 根据局部线性嵌入 (Locally Linear Embedding, LLE) 的算法流程，尝试编写 LLE 代码，可以基于 sklearn 实现，并在瑞士卷数据集上进行实验降到 2 维空间。提交代码和展示多个在不同参数下的可视化的实验结果。请分析使用 LLE 时可能遇到哪些挑战 (10 points) [提示：瑞士卷数据集可以用 sklearn 的 make_swiss_roll(n_samples=3000, random_state=0) 生成 3000 个样本]

解:

代码见 Prob1/task2/task2.py, 结果见 Prob1/task2/out

(a) 原 3d 的瑞士卷数据集可视化如下：

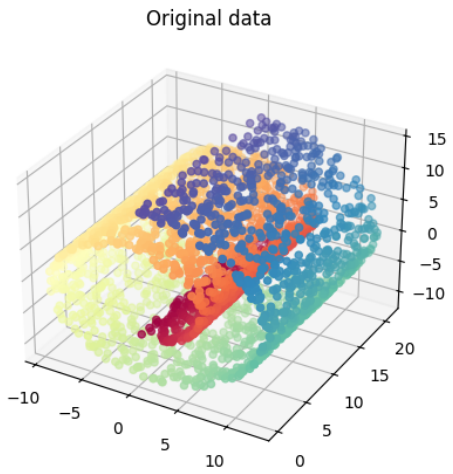


图 3: Swiss Roll

(b) 我们选取不同的近邻数目，进行 LLE 降维。

可以发现，近邻数目过多或过少都会导致降维后的数据不具有代表性。如果近邻数目过小，每个数据点的邻居仅限于非常局部的范围。如果近邻数目过大，每个数据点的邻居范围过于宽泛，可能包含了不相关的点。因此，我们需要选择合适的近邻数目。在这里，我们选择近邻数目为 5, 7, 12, 15, 20, 30, 50, 75, 100 进行降维来观察结果。

可视化结果如下：

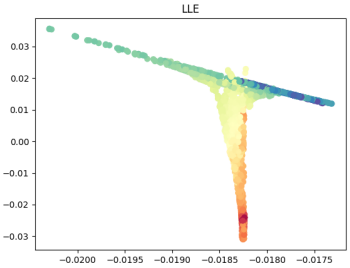


图 4: with 5 neighbors

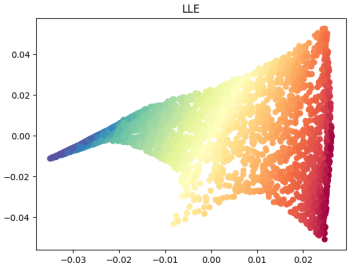


图 5: with 7 neighbors

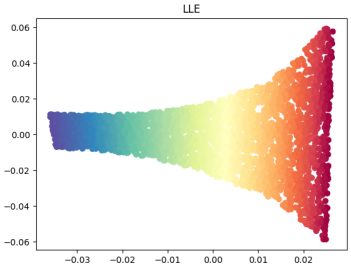


图 6: with 12 neighbors

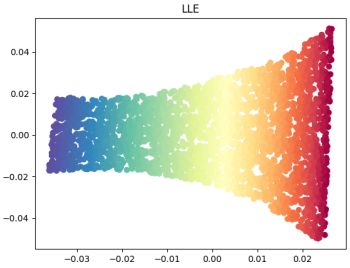


图 7: with 15 neighbors

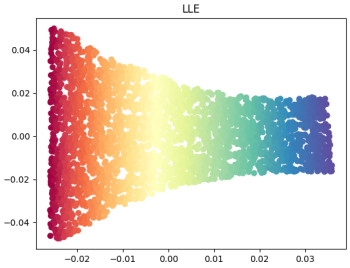


图 8: with 20 neighbors

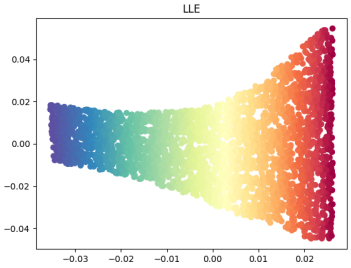


图 9: with 30 neighbors

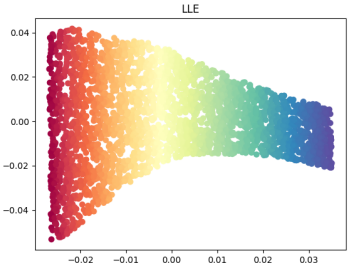


图 10: with 50 neighbors

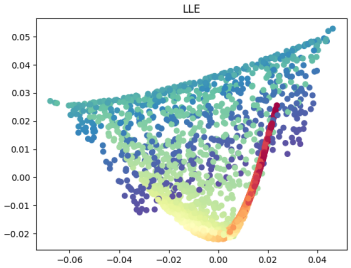


图 11: with 75 neighbors

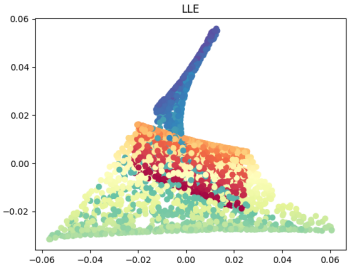


图 12: with 100 neighbors

(c) 使用 LLE 时可能遇到哪些挑战:（参考资料：LLE 算法详解）

- 算法本身的局限：算法所学习的流形只能是不闭合的，且样本集是稠密均匀的。
- 选择合适的近邻数目：近邻数目的选择对 LLE 的效果有很大影响，过小的近邻数目可能导致降维后的数据不具有代表性，过大的近邻数目可能导致降维后的数据过于稀疏。
- 数据分布不均匀：导致局部线性假设失效，无法很好地保留数据的局部结构。
- 数据噪声敏感：LLE 算法对数据的噪声敏感，噪声会对降维结果产生较大影响。

二. (20 points) 特征选择

1. （12 points）使用 Wine 数据集，比较过滤式（基于互信息）和包裹式（RFE）特征选择方法的性能。

(i) 实现两种特征选择方法，选择特征数量从 1 到全部特征

(ii) 使用交叉验证评估不同特征数量下的模型准确率

(iii) 绘制特征数量与准确率的关系图

(iv) 分析并比较：两种方法的最佳特征数量和对应准确率, 计算并解释每个特征被选择的频率

解:

(i)&(ii) 代码见 Prob2/task1/task1.py 和 Prob2/task2/task2.py, 结果见 Prob2/out

(iii) 特征数量与准确率的关系图如下：

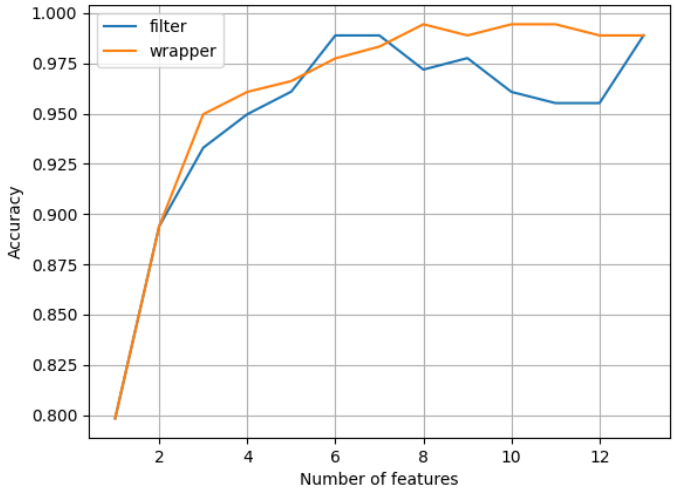


图 13: Feature Number vs Accuracy

(iv) 分析：两种方法的最佳特征数量和对应准确率, 计算并解释每个特征被选择的频率（特征重要性）

```
Filter method:
    Best number of features: 6
    their idxs are: [ 0  6  9 10 11 12]
    Best accuracy: 0.9888888888888889
Wrapper method:
    Best number of features: 8
    their idxs are: [ 0  1  2  3  6  9 11 12]
    Best accuracy: 0.9944444444444445
Features idx: range(0, 13)
Filter method feature frequency: [0.0989011  0.06593407 0.01098901 0.03296703 0.05494505 0.07692308 0.14285714 0.02197802 0.04395604 0.10989011 0.08791209 0.12087912 0.13186813]
Wrapper method feature frequency: [0.0989011  0.07692308 0.10989011 0.08791209 0.02197802 0.05494505 0.14285714 0.04395604 0.01098901 0.12087912 0.03296703 0.06593407 0.13186813]
```

图 14: 两种方法的最佳特征数量、对应准确率和每个特征被选择的频率

（文件见Prob2/out/task1/task1_4.log）

2.（8 points）使用 L1 正则化的 Logistic 回归（LASSO）进行特征选择。

要求：

(i) 实现基于 LASSO 的特征选择，给出代码

(ii) 分析：

(a) 被选择的特征（系数非零）

(b) 特征的重要性排序（基于系数绝对值大小）

(c) 基于 Lasso 选择出特征（对应 Logistic 回归系数非 0），计算对应的模型准确率

(d) 对比相同特征数量下，三种特征选择方法的模型准确率

解:

(i) 代码见 Prob2/task2/task2.py, 结果见 Prob2/out

(ii) 分析：

(a) 一共有 13 个特征，为 0——12.

(b) 被选择的特征为：6, 11, 12, 10, 3（按照重要性排序）

(c) 重要性排序如上，系数的绝对值为：0.2653, 0.1959, 0.1517, 0.0679, 0.0674.

(d) 基于 Lasso 选择出特征（对应 Logistic 回归系数非 0），计算对应的模型准确率为 0.9497.

(e) 对比相同特征数量 (feature=5) 下，三种特征选择方法的模型准确率如下：

特征选择方法	准确率
过滤式（基于互信息）	0.96095
包裹式（RFE）	0.96619
LASSO	0.94968

表 1: 三种特征选择方法的模型准确率

三. (20 points) 半监督

1. 在本题中使用朴素贝叶斯模型和 SST2 数据集进行半监督 EM 算法的实践，代码前面部分如下，请补充完后续代码，只保留 10% 的标注数据，置信度设为 0.7，训练 5 轮，给出训练后模型在验证集上的分类结果 (10 points)

解:

代码见 Prob3/task1.py, 结果见 Prob3/out

由于验证集数据也比较多，我们以验证集上的准确率作为模型在验证集上的分类结果。

- 未使用半监督 EM，模型在未标注数据上的准确率为 0.8134，在验证集上的准确率为 0.7775
- 使用半监督 EM，置信度为 0.7，训练 5 轮，模型在未标注数据上的准确率为 0.7924，在验证集上的准确率为 0.7695

2. 伪标签的置信度大小对模型的训练结果会有一定的影响，通常会有固定置信度和动态设置置信度两种方式，请你完成这两种方式，并统计不同方式下每次迭代中伪标签的错误率，并分析这两种方式的优劣 (5 points)

解:

代码见 Prob3/task1.py, 结果见 Prob3/out

- 由于错误率等于 1-准确率，以下展示的是每次迭代中伪标签的准确率:（epoch 0 表示未经 EM 迭代时的模型准确率）

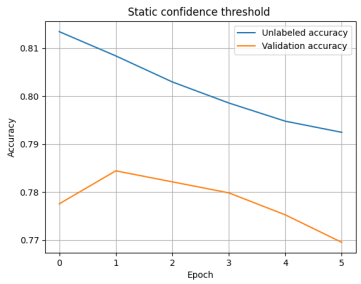


图 15: 固定置信度

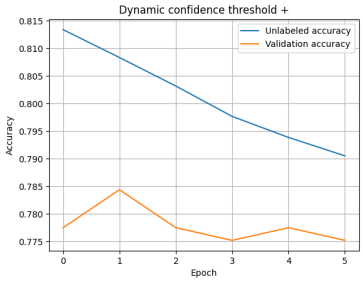


图 16: 动态置信度 (每轮变大)

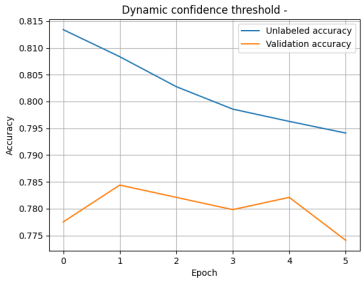


图 17: 动态置信度 (每轮变小)

- 两种方式的优劣:
  - 固定置信度：优点是简单易行。缺点是过小的固定置信度可能会导致模型在训练过程中过度自信从而过度依赖伪标签，从而影响模型的预测能力；过大的固定置信度可能会导致模型在训练中放弃过多伪标签，从而影响泛化能力。
  - 动态置信度：优点是可以根据模型的训练情况动态调整伪标签的置信度，使得模型在训练过程中考虑伪标签的质量和数量。缺点是需要更多的调参工作，并且可能会导致模型的稳定性下降。

3. 修改代码，设置不同的迭代次数（如 3 次、5 次、15 次）。在验证集上分析：不同迭代次数下，模型性能如何变化？分析为什么在过多迭代的情况下，模型性能可能下降？(5 points)

解:

代码见 Prob3/task1.py, 结果见 Prob3/out

- 不同迭代次数下，模型性能如下：

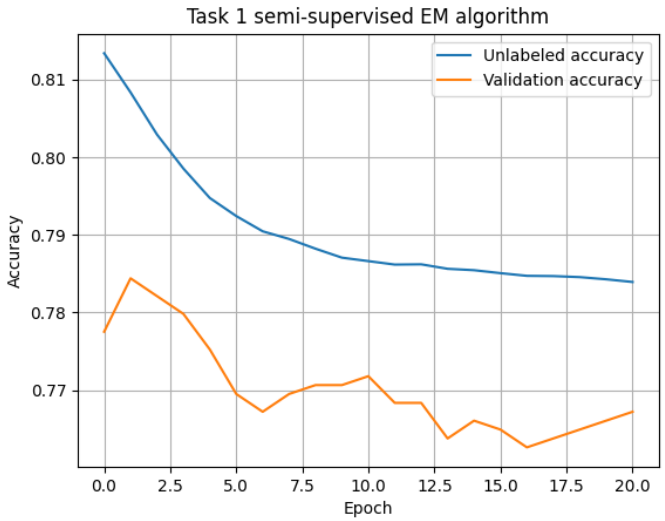


图 18: 不同迭代次数下，模型性能

- 分析：在过多迭代的情况下，模型性能下降了。可能是由于模型在多轮迭代过程中逐步过度拟合伪标签，导致过拟合，从而导致模型的泛化能力下降，从而导致模型在验证集上的性能下降。
- 参考资料

四. (20 points) 概率图模型

1. 证明 $a \perp\!\!\!\perp (b,c) \mid d$ 蕴含 $a \perp\!\!\!\perp b \mid d$。（5 points）

解:

由于 $a \perp\!\!\!\perp (b,c) \mid d$ ，我们有：$P(a \mid b,c,d) = P(a \mid d)$

由于条件概率的性质，得到：$P(a \mid b,d) = \sum_c P(a \mid b,c,d)P(c \mid b,d)$

根据 $a \perp\!\!\!\perp (b,c) \mid d$，我们有 $P(a \mid b,c,d) = P(a \mid d)$，代入上式得到：

$$P(a \mid b,d) = \sum_c P(a \mid d)P(c \mid b,d)$$

由于 $P(a \mid d)$ 与 c 无关，可以从求和符号中提取出来：

$$P(a \mid b,d) = P(a \mid d)\sum_c P(c \mid b,d)$$

根据概率的全概率公式，$\sum_c P(c \mid b,d) = 1$，因此：

$$P(a \mid b,d) = P(a \mid d)$$

即 $a \perp\!\!\!\perp b \mid d$。因此，$a \perp\!\!\!\perp (b,c) \mid d$ 蕴含 $a \perp\!\!\!\perp b \mid d$。

2. 假设你有一组 d 个二元随机变量 $\{X_1, ..., X_d\}$。（5 points）

(i) 在不做任何独立性假设的情况下，完全描述联合分布所需的最小参数个数是多少？
提示: 由于总概率之和必须为 1，所需的参数个数比结果的总数少一个。

解: $2^d - 1$

因为是二元随机变量，每个随机变量有两个取值，所以每个随机变量的联合分布有 2^d 种可能的取值。由于提示，所以所需的最小参数个数是 $2^d - 1$

(ii) 假设这些随机变量的结构为马尔可夫链，其中每个 X_i 仅依赖于 X_{i-1}。在这种情况下，所需的最小参数个数是多少？

解: $2d - 1$

在马尔可夫链结构中，X_1 的分布需要 1 个参数（因为 X_1 是二元的），每个 X_i 的条件分布 $P(X_i \mid X_{i-1})$ 需要 2 个参数（因为 X_{i-1} 有 2 种取值，每种取值下 X_i 的条件概率需要 1 个参数）。因此，总参数个数为：

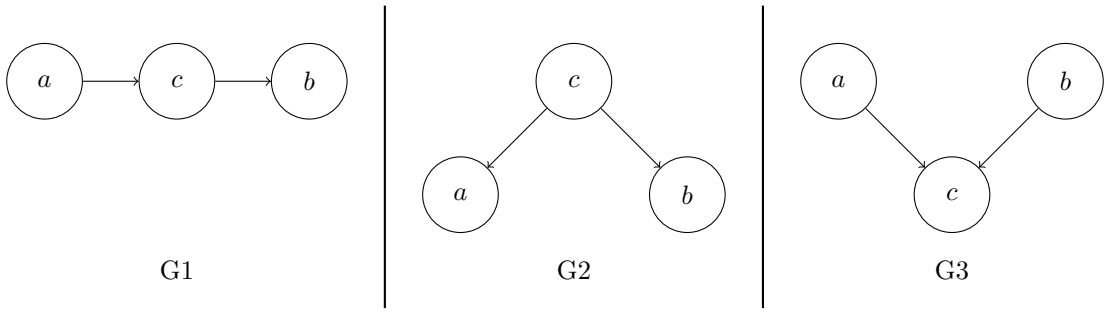
$$1 + 2(d-1) = 2d - 1$$

(iii) 从不做独立性假设到引入马尔可夫假设，参数复杂度是如何变化的？

解:

参数复杂度从指数级 $O(2^d)$（$2^d - 1$）降低为线性级 $O(d)$（$2d - 1$），显著减少了参数的数量。

3. 考虑以下三种结构各异的图模型。



请将每个情境与最合适的图模型进行匹配。（5 points）

(i) 一个家庭的旅行决定 (c) 会受到父母的工作安排 (a) 孩子的学校假期 (b) 的影响。

解: G3

(ii) 破纪录的大雪 (c) 会同时刺激滑雪度假村的预订量 (a) 和冬季服装的需求 (b)。

解: G2

(iii) 个人的锻炼习惯 (a) 会影响自身的能量水平 (c)，进而影响工作效率 (b)。

解: G1

(iv) 一个地区的气候 (a) 决定了生长的植被类型 (c)，而植被类型又会影响野生动物的数量 (b)。

解: G1

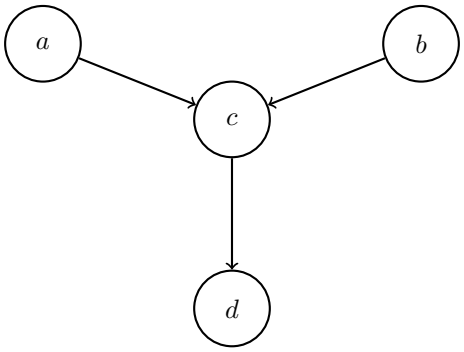
(v) 一个国家的经济稳定性 (c) 会影响就业率 (a) 和消费者的消费习惯 (b)。

解: G2

(vi) 餐厅的受欢迎程度 (c) 取决于食品质量 (a) 和社交媒体的曝光度 (b)。

解: G3

4. 考虑以下有向图，其中所有变量均为未观测变量。（5 points）



(i) 在给定的图中，如何将联合分布 $p(a, b, c, d)$ 表示为边际分布和条件分布的组合？

解:

(ii) 假设 a, b, c, d 是二元随机变量，求建模该联合分布所需的最小参数个数。

解:

(iii) 证明 a 和 b 是相互独立的，即 $a \perp\!\!\!\perp b$。

解:

(iv) 假设对于图中任意节点 x , 都有 $p(x \mid \mathrm{pa}(x)) \neq p(x)$, 其中 $\mathrm{pa}(\cdot)$ 表示节点 x 的父节点集合。证明当观测到 d 时, a 和 b 不再相互独立, 即 $a \not\perp\!\!\!\perp b \mid d$.

解:

五. (20 points) 强化学习

在本问题中，你将思考如何通过在马尔可夫决策过程（MDP）中连续做决策来最大化奖励，并深入了解贝尔曼方程——解决和理解 MDP 的核心方程。

考虑经典的网格世界 MDP，其中智能体从单元格 (1, 1) 开始，并在环境中导航。在这个世界中，智能体每个格子里可以采取四个动作：上、下、左、右。格子用 (水平，垂直) 来索引；也就是说，单元格 (4, 1) 位于右下角。世界的转移概率如下：如果智能体采取一个动作，它将以 0.8 的概率移动到动作的方向所在的格子，并以 0.1 的概率滑到动作的相对右或左的方向。如果动作（或滑动方向）指向一个没有可通过的格子（即边界或 (2, 2) 格子的墙壁），那么该动作将保持智能体处于当前格子。例如，如果智能体在 (3, 1) 位置，并采取向上的动作，它将以 0.8 的概率移动到 (3, 2)，以 0.1 的概率移动到 (2, 1)，以 0.1 的概率移动到 (4, 1)。如果智能体在 (1, 3) 位置并采取右移动作，它将以 0.8 的概率移动到 (2, 3)，以 0.1 的概率移动到 (1, 2)，以 0.1 的概率停留在 (1, 3)。当智能体到达定义的奖励状态时（在 (4, 2) 和 (4, 3) 单元格），智能体将获得相应的奖励，并且本次回合结束。

回顾计算 MDP 中每个状态的最优价值, $V^*(s)$ 的贝尔曼方程，其中我们有一组动作 A，一组状态 S，每个状态的奖励值 $R(s)$，我们的世界的转移动态 $P(s'|s,a)$，以及折扣因子 γ：

$$V^*(s) = R(s) + \gamma \max_{a \in A} \sum_{s' \in S} P(s'|s,a) V^*(s')$$

最后，我们将策略表示为 $\pi(s) = a$ 其中策略 π 指定了在给定状态下采取的行动。

(a) 考虑一个智能体从单元格 (1, 1) 开始，在第 1 步和第 2 步分别采取向上和向上的动作。计算在每个时间步内，根据这一动作序列，智能体可以到达哪些单元格，以及到达这些单元格的概率。（6 points）

解:

(b) 考虑当前没有奖励值的所有状态的奖励函数 $R(s)$（即除了 (4, 2) 和 (4, 3) 以外的每个单元格）。定义在以下奖励值下，智能体的最优策略：(i.) $R(s) = 0$, (ii.) $R(s) = -2.0$, and (iii.) $R(s) = 1.0$. 你可以假设折扣因子接近 1，例如 0.9999。画出网格世界并标出在每个状态下应采取的动作可能会对你有所帮助（记住，策略是在 MDP 中对所有状态进行定义的!）（7 points）

注意: 你不需要算法上计算最优策略。你必须列出每种情况的完整策略，但只需要提供直观的理由。

解:

(c) 有时，MDP 的奖励函数形式为 $R(s,a)$ 它依赖于所采取的动作，或者奖励函数形式为 $R(s,a,s')$ ，它还依赖于结果状态。写出这两种形式的最优价值函数的贝尔曼方程。（7 points）

解: