北京理工大学 2021-2022 学年春季学期

《数据科学与大数据技术的数学基础》期末考试

考试时间：2022-06-20 15:30-17:30

总共 11 题，满分 70 分。请注意写清解答步骤，确保字迹清晰。

1. 在利用最小哈希估计数据流$x_1, x_2, \ldots, x_n$中不同元素个数时，将由独立的哈希函数$h_1(\cdot), h_2(\cdot), \ldots, h_k(\cdot)$得到的数据流元素的最小哈希值分别记为$s_1, s_2, \ldots, s_k \in [0,1]$（假设对任意元素$x$，$h_i(x)$是在$[0,1]$均匀分布的随机变量）。请写出如何用$s_1, s_2, \ldots, s_k$完成对不同元素个数的估计。[4 分, 本题写出表达式即可，不要求做理论推导]

2. 在 JL 转换中，给定数据$\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathbb{R}^k$，利用$d \times k$矩阵$\mathbf{A}$对数据作线性变换，通过变换所得的数据之间依然能（近似）保持原数据$\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$之间的欧氏距离。请完整写出前一句话中的划线部分对应的不等式（即写出欧几里得低失真嵌入问题中需满足的不等式，若引入新的数学符号，需要解释其含义）。[4 分, 本题不要求做理论推导]

3. 在用布隆过滤器解决近似从属判断问题时，记集合$\mathcal{S}$的不同元素个数为n，为每个元素都根据独立的哈希函数$h_1(\cdot), h_2(\cdot), \ldots, h_k(\cdot)$分别计算$k$个哈希值，并将长为$m$的布隆过滤器的相应位置置为 1（假设对任意元素$x$，$h_i(x)$是在$\{1,2,\ldots,m\}$均匀分布的随机变量）。此后，

（1）若元素x属于$\mathcal{S}$，那么$h_1(x)$在布隆过滤器中的位置被置 1 的概率是多少？

（2）若元素x不属于$\mathcal{S}$，那么$h_1(x)$在布隆过滤器中的位置被置 1 的概率是多少？

（3）若元素x不属于$\mathcal{S}$，而且额外已知布隆过滤器的m个位置中共有t个已被置 1，那么$h_1(x)$在布隆过滤器中的位置已被置 1 的概率是多少？[7 分, 本题三问写出表达式即可，不要求利用$m \to \infty$条件作近似]

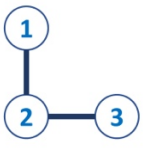
4. 在一致性哈希中，记机器的个数为n，若只用一个随机哈希函数为机器做映射，为证明以较大概率至少一个机器在哈希值空间（即“圆圈”）上占得的比例较小，可采取如下步骤：

第一步，将哈希值空间均匀分成m段，推导每个段都最多包含一个机器的概率；

第二步，证明在一定条件下该概率低于一定阈值。

请：（1）写出第一步中划线的概率的表达式；（2）假设m值较大，请利用$\left(1-\frac{1}{m}\right) \approx e^{-\frac{1}{m}}$及类似约等式对（1）中的概率近似、整理成指数的形式（即写成$e^{g(n,m)}$的形式，$g(n,m)$为待求函数）。[6 分]

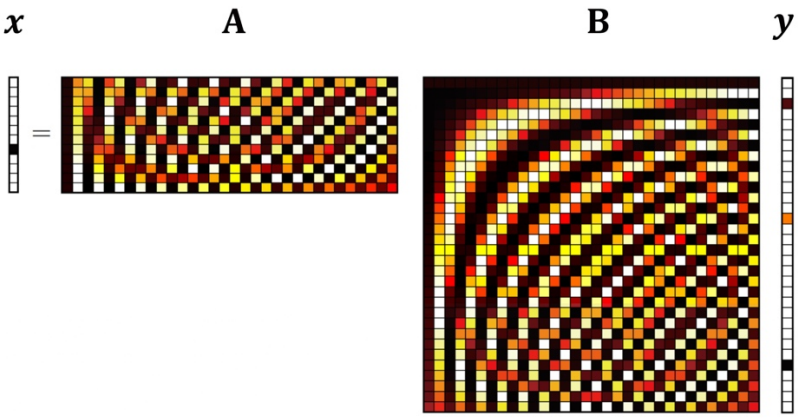
5. 如下图所示，现有一个无向简单图。（1）请写出该图的拉普拉斯矩阵；（2）请简单描述如何利用拉普拉斯矩阵为图中三个顶点分别计算一个一维表示，使相邻顶点有相似但不相同的表示（即解决谱嵌入问题，每个顶点的嵌入属于$\mathbb{R}$）。[4 分，本题不要求计算顶点的一维表示，仅要求用一至两句话简述方法]



6. 下图对应的是压缩感知中的感知过程，根据图中符号该过程相应的等式为$\boldsymbol{x}=\mathbf{A}\mathbf{B}\boldsymbol{y}$。

（1）请分别指出$\boldsymbol{x},\mathbf{A},\mathbf{B},\boldsymbol{y}$的名称（如原始信号、恢复信号、测量矩阵、正交矩阵、测量向量、稀疏向量等）；

（2）在图中，$p\times n$矩阵$\mathbf{A}$的各行向量是$\mathbf{B}$的后p个列向量的转置（$p<n$）。请分析此类$\mathbf{A}$对压缩感知的影响，如需修改$\mathbf{A}$，请指出一种关于$\mathbf{A}$的替代方案。[8 分]



7. 假设$\boldsymbol{u}_0=\begin{bmatrix}0\\1\end{bmatrix}$而且$2\times 2$对称矩阵$\mathbf{A}$的正交对角化为$\mathbf{A}=\mathbf{Q}\mathbf{D}\mathbf{Q}^T=\begin{bmatrix}\frac{\sqrt{3}}{2} & -\frac{1}{2}\\ \frac{1}{2} & \frac{\sqrt{3}}{2}\end{bmatrix}\begin{bmatrix}3 & 0\\0 & 1\end{bmatrix}\begin{bmatrix}\frac{\sqrt{3}}{2} & \frac{1}{2}\\ -\frac{1}{2} & \frac{\sqrt{3}}{2}\end{bmatrix}$。在理解$\mathbf{A}$左乘$\boldsymbol{u}_0$的几何意义时，可以在平面直角坐标系中画出$\boldsymbol{u}_0$对应的$(0,1)$点，并将$\mathbf{A}$左乘$\boldsymbol{u}_0$理解为$\mathbf{Q}^T,\mathbf{D},\mathbf{Q}$依次左乘向量$\boldsymbol{u}_0$。请：

（1）分别指出三次左乘的几何意义（如将$\boldsymbol{u}_0$对应的点顺时针/逆时针旋转多少度等）；

（2）画一个直角坐标系，标注出向量$\mathbf{A}\boldsymbol{u}_0$（即$\mathbf{A}$左乘$\boldsymbol{u}_0$所得的$2\times 1$向量），以及标注出矩阵$\mathbf{A}$最大特征值对应的特征向量所在的直线。[7 分，画图起到示意效果即可，对比例不做严格要求]

8. 给定矩阵$\mathbf{A}=\begin{bmatrix}-1 & -1\\ 1 & 0\\ 0 & 1\end{bmatrix}$，定义 F-范数为$\|\mathbf{A}\|_F=\sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}\mathrm{A}_{ij}^2}$。

（1）计算$\mathbf{A}$的奇异值分解（第一种定义或第二种定义都可以，按第二种定义写更简便）；

（2）利用$\mathbf{A}$的奇异值分解解决（目标秩为 1 的）低秩矩阵近似问题：$\min\limits_{\mathbf{B}\in\mathbb{R}^{3\times 2}}\|\mathbf{A}-\mathbf{B}\|_F\quad \text{s.t. } \mathrm{rank}(\mathbf{B})\le 1$。要求写出最优解$\mathbf{B}^*$（即最优近似矩阵），以及$\|\mathbf{A}-\mathbf{B}^*\|_F$的值（可以用公式）；

（3）矩阵$\mathbf{A}$可视作三个二维数据（即$\boldsymbol{a}_1=[-1,-1]^T,\boldsymbol{a}_2=[1,0]^T,\boldsymbol{a}_3=[0,1]^T$）构成的数据矩阵。利用主成分分析，可将$\boldsymbol{a}_1,\boldsymbol{a}_2,\boldsymbol{a}_3$分别近似为$\tilde{\boldsymbol{a}}_1,\tilde{\boldsymbol{a}}_2,\tilde{\boldsymbol{a}}_3$。其中，$\tilde{\boldsymbol{a}}_1,\tilde{\boldsymbol{a}}_2,\tilde{\boldsymbol{a}}_3$可分别写成同一个向量$\boldsymbol{v}=[v_1,v_2]^T$乘以一个系数的形式。定义$3\times 2$矩阵$\tilde{\mathbf{A}}$为$\begin{bmatrix}\tilde{\boldsymbol{a}}_1^T\\ \tilde{\boldsymbol{a}}_2^T\\ \tilde{\boldsymbol{a}}_3^T\end{bmatrix}$，求向量$\boldsymbol{v}$和矩阵$\tilde{\mathbf{A}}$（不用在主成分分析前对数据做平移、比例缩放等预处理）。[17 分]

9. 若矩阵$\mathbf{A}$是一个$n \times n$可逆矩阵，将它的奇异值分解记为$\mathbf{A} = \mathbf{USV}^T$，将它的奇异值分解的等价形式记为$\mathbf{A} = \sum_{i=1}^{n} s_i \boldsymbol{u}_i \boldsymbol{v}_i^T$。其中，$s_i$是$\mathbf{A}$的奇异值，$\boldsymbol{u}_i$是$\mathbf{A}$的左奇异向量，$\boldsymbol{v}_i$是$\mathbf{A}$的右奇异向量。由$\mathbf{A}$可逆可知$s_1, s_2, \dots, s_n \neq 0$。求矩阵$\mathbf{A}^{-1}$的奇异值分解的等价形式（$\mathbf{A}^{-1}$为$\mathbf{A}$的逆矩阵，满足$\mathbf{AA}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$）。[5 分]

10. 根据马尔可夫不等式，若X是非负随机变量且$\mathbb{E}[\mathrm{X}] \neq 0$，那么对任何常数$c > 0$，有$\Pr[\mathrm{X} \geq c\mathbb{E}[\mathrm{X}]] \leq \frac{1}{c}$。

请用马尔可夫不等式证明若Y是随机变量，对任意$a \in \mathbb{R}$及$t > 0$，有$\Pr[\mathrm{Y} \geq a] \leq \frac{\mathbb{E}[e^{t\mathrm{Y}}]}{e^{ta}}$。[4 分]

11. 在相似搜索中，给定文件A, B，可以定义函数$f(\cdot)$、将它们分别映射为一个实数。例如，若考虑Jaccard 相似度下的相似搜索，可以分别定义$f(\mathrm{A}), f(\mathrm{B})$为文件A, B中所有元素哈希值的最小值，继而能证明$\Pr[f(A) = f(B)] = J(A, B)$。

现考虑一种新的相似度下的相似搜索，即用向量$\boldsymbol{x}_\mathrm{A}, \boldsymbol{x}_\mathrm{B} \in \mathbb{R}^n$分别表示文件A, B，用向量$\boldsymbol{x}_\mathrm{A}, \boldsymbol{x}_\mathrm{B}$之间的夹角反映文件A, B的相似度：夹角$0^\mathrm{o}$表示相似度最高，夹角$180^\mathrm{o}$表示相似度最低。针对这种新的相似度定义，可以如此定义$f(\mathrm{A}), f(\mathrm{B})$：随机生成向量$\boldsymbol{r} \in \mathbb{R}^n$（$\boldsymbol{r}$的各元素分别独立根据标准正态分布取值），

$$f(\mathrm{A}) = \begin{cases} 1, & \text{若} \boldsymbol{r}^T \boldsymbol{x}_\mathrm{A} \geq 0, \\ -1, & \text{若} \boldsymbol{r}^T \boldsymbol{x}_\mathrm{A} < 0, \end{cases} \qquad f(\mathrm{B}) = \begin{cases} 1, & \text{若} \boldsymbol{r}^T \boldsymbol{x}_\mathrm{B} \geq 0, \\ -1, & \text{若} \boldsymbol{r}^T \boldsymbol{x}_\mathrm{B} < 0. \end{cases}$$

在上式中，$\boldsymbol{r}^T \boldsymbol{x}_\mathrm{A}$等于列向量$\boldsymbol{r}$与列向量$\boldsymbol{x}_\mathrm{A}$的内积$\langle \boldsymbol{r}, \boldsymbol{x}_\mathrm{A} \rangle$。计算$f(\mathrm{A}), f(\mathrm{B})$时采用的是相同的$\boldsymbol{r}$。

请考虑$\boldsymbol{x}_\mathrm{A}, \boldsymbol{x}_\mathrm{B}$都是二维向量的简化情况（即$\boldsymbol{x}_\mathrm{A}, \boldsymbol{x}_\mathrm{B} \in \mathbb{R}^2$），求$\Pr[f(A) = f(B)]$与向量$\boldsymbol{x}_\mathrm{A}, \boldsymbol{x}_\mathrm{B}$之间夹角的关系。[4 分, 可以用作图的方式进行分析说明]