一. (30 points) **半监督学习**

关于半监督学习，课堂上介绍了为什么需要半监督学习以及半监督学习的几种基本做法。在前沿研究中，无论是计算机视觉还是自然语言处理领域，半监督学习仍然是研究重点。下面三个问题分别针对传统半监督学习、深度半监督学习进行拓展介绍。

1. (10 points) 在 TSVM 中会对未标记的样本进行标记指派，这个本质上也是一种伪标签的应用。伪标签的含义就是给无标记数据赋予标签，最初这些标签可能绝大多数是错误的，但是会有一些是准确的，将这些准确的筛选出来加入有标记数据训练即可起到扩充训练集的效果。这也是 TSVM 的主要思想。因此，使用一个预训练模型给未标记数据打标签之后，该怎么做才能尽可能选出那些准确的样本呢？请提供至少三种启发式解决方案并说明理由。(提示：从置信度等角度出发)。试查阅主动学习 (Active Learning) 相关资料，介绍主动学习并分析其与基于伪标签半监督学习的区别。
2. (10 points) 标记传播 (Label Propagation) 是一种非常有效的图半监督学习算法，将样本看做图的顶点，然后将标记从有标记样本扩散到未标记样本。该过程类似于 PageRank 算法，试描述 PageRank 算法并比较其与 Label Propagation 的区别。其余相关的算法还包括：Query Diffusion (QD) 等。QD 算法先计算样本点之间相似度 $a_{ij} = s(x_i, x_j) \geq 0, \forall i, j \in [n]^2$(假设该相似度对称，即 $a_{ij} = a_{ji}$)，构成亲和矩阵 (Affinity Matrix) $A \in \mathcal{R}^{(n+m)\times(n+m)}$，然后对每个样本点计算近邻集合 $NN_k(i)$，k 代表近邻样本数目，记 $D = \text{Diag}(\text{sum}(A, 0))$ 表示亲和矩阵的列和组成的对角阵，那么标准化的亲和矩阵为 $S = D^{-1/2}AD^{-1/2}$。记初始的扩散标记为 $f_0 \in \mathcal{R}^{n+m}$，例如是前 n 个有标记数据的标记 (回归问题的预测值) 和以 0 填充的 m 个未标记数据的标记。那么试着给出其扩散的迭代公式，并分析其意义。
3. (10 points) 课堂上介绍的传统半监督学习有几种经典假设：聚类假设和流形假设。前者假设数据存在聚簇结构，同一簇的样本属于同一类别；后者假设数据分布在一个流形结构上，邻近的样本有相同的输出值。在深度半监督学习中，有某些其它常用的假设：一致性假设 (Consistency Assumption) 和低熵 (Low Entropy Assumption) 假设。一致性假设指的是相似的样本具有相似的输出，低熵假设则是指预测输出结果尽可能具有较低的信息熵。在分类问题中，深度学习常用损失为交叉熵损失。假设有标记样本为 $\{(x_i^l, y_i^l)\}_{i=1}^n$，其

中 $y_i^l \in \{1, 2, \cdots, C\}$ 代表类别，C 表示类别数目。无标记样本记为 $\{x_j^u\}_{j=1}^m$。记神经网络拟合的函数为 $f_\theta(x)$，θ 代表待优化参数。假设该神经网络最后一层没有任何激活函数，即输出是每个类别的打分，打分可正可负，即 $f_\theta(x) \in \mathcal{R}^C$。请写出交叉熵函数的具体形式 (可参考：Softmax + Cross Entropy)，以及带有一致性假设和低熵假设优化目标的损失函数 (一致性假设中相似的样本可以根据数据增强来获得，即:$\hat{x}_j^u = \text{DataAugmentation}(x_j^u)$)。

解：

1. 使用启发式解决方案的重点在于设计启发式函数，故可以基于置信度等角度设计启发式函数，使得准确的样本的启发式函数值更高，从而将他们选取出来。
   我们可以将已知的一个类别和另一类别的中间位置作为分隔平面，可以将落在分隔平面两侧的样本分别赋予其对应区域类别的标签，距离分隔平面越近的数据点其标签的不确定性越高，故我们可以每次选择不确定性最高的数据点作为 “query instance” 进行判断。故启发式函数可用来度量不确定性，以选出 “query instance”。
   (1) 可简单考虑距离分隔平面最近的样本点，即不确定性最高的样本点，即后验概率最低的样本点，故可以使用置信度进行度量，选择置信度最小的样本点作为 “query instance”。

$$\begin{aligned} \mathrm{x}_{LC}^* &= \arg\min_x P_\theta(\hat{y}|x) \\ &= \arg\max_x 1 - P_\theta(\hat{y}|x) \end{aligned} \tag{1}$$

   (2) 也可考虑最难分离的两个样本点，可以使用样本属于某个类别的后验概率，计算最大的两个类别后验概率的差进行度量，选择差最小（即最难区分）的数据点作为 “query instance”。

$$\begin{aligned} \mathrm{x}_{M}^* &= \arg\min_x \left(P_\theta(\hat{y_1}|x) - P_\theta(\hat{y_2}|x)\right) \\ &= \arg\max_x \left(P_\theta(\hat{y_2}|x) - P_\theta(\hat{y_1}|x)\right) \end{aligned} \tag{2}$$

   其中 x 属于类别 $\hat{y_1}$ 的概率高于属于类别 $\hat{y_2}$ 的概率。
   (3) 也可考虑样本点的熵，熵越大混乱程度也越大，即不确定性

越高，选择熵最大的数据点作为“query instance”。

$$\begin{aligned} \mathrm{x}_H^* &= \arg\max_x H_\theta(Y|x) \\ &= \arg\max_x -\sum_y P_\theta(y|x)\log P_\theta(y|x) \end{aligned} \tag{3}$$

主动学习是一种半监督学习方法，可以在大量未标记数据中寻找最有价值的数据，进而提升模型的性能。其基本思想是每一轮主动地选择少部分样本进行查询，并进行标记，再训练模型，直到满足要求。

主动学习与基于伪标签半监督学习的区别在于主动学习进行多轮学习以给未标记的样本加上标记，且每一轮仅选择少量已标记样本进行学习；而基于伪标签的半监督学习则只进行一轮学习以给未标记的样本加上标记，且直接使用所有已标记样本进行学习。

2. PageRank 算法是一个计算网站重要程度的算法，它首先将各个网页的重要性得分初始化为相同的值，然后通过迭代或递归计算来更新每个页面的重要性得分，直到得分趋于稳定，即得到各个网站的重要程度，从而可以得到网站重要性的排名。基本想法是在有向图上定义一个一阶马尔科夫链，描述在有向图上随机游走访问各个节点的行为，最后访问每个节点的概率会收敛到状态平稳分布，此时各个节点的概率值就是节点的重要性得分。

   PageRank 算法与 Label Propagation 算法的区别主要在于 PageRank 算法可以视为对样本点进行回归，而且其基于马尔科夫链对每个样本点的值进行更新，而 Label Propagation 则是对样本点进行分类，而且其仅简单根据相邻节点的标签对样本点标签进行更新，即离得近的样本标记要尽量一致。

   若以 α 的概率进行随机游走，以 $1-\alpha$ 的概率以相等概率随机跳转到任意一个节点，则 t 时刻下 QD 算法扩散的迭代公式可表示为

$$\mathrm{f}_{t+1} = \alpha \mathrm{S}\mathrm{f}_t + (1-\alpha)\mathrm{y} \tag{4}$$

   其中 y 为随机跳转到各个节点的概率，可设置为 n 个 1 和 m 个 0，即跳转到每个有标记数据的概率相等，且对排序得分的贡献

均为 1。
其意义为以 α 的概率进行随机游走，以 $1-\alpha$ 的概率以相等概率随机跳转到任意一个节点后得到新的排序得分。

3. 定义指示函数

$$I(i,c)=\begin{cases}1, & y_i^l=c\\ 0, & y_i^l\neq c\end{cases} \tag{5}$$

则交叉熵函数的具体形式为

$$\mathrm{H}(I,f_\theta)=-\sum_{i=1}^{n}\sum_{c=1}^{C}I(i,c)\log\left(\frac{e^{\left(f_\theta\left(x_i^l\right)\right)_c}}{\sum_{j=1}^{C}e^{\left(f_\theta\left(x_i^l\right)\right)_j}}\right) \tag{6}$$

记 $\hat{x}_i^u=\mathrm{DataAugmentation}(x_i^u)$, 其对应的类别标签为 $\hat{y}_i^u$, 定义数据增强样本的指示函数

$$\hat{I}(i,c)=\begin{cases}1, & \hat{y}_i^u=c\\ 0, & \hat{y}_i^u\neq c\end{cases} \tag{7}$$

则带有一致性假设和低熵假设优化目标的损失函数为

$$\begin{aligned}\mathrm{L}=&-\sum_{i=1}^{n}\sum_{c=1}^{C}I(i,c)\log\left(\frac{e^{\left(f_\theta\left(x_i^l\right)\right)_c}}{\sum_{j=1}^{C}e^{\left(f_\theta\left(x_i^l\right)\right)_j}}\right)\\ &-\sum_{i=1}^{m}\sum_{c=1}^{C}\hat{I}(i,c)\log\left(\frac{e^{(f_\theta(x_i^u))_c}}{\sum_{j=1}^{C}e^{(f_\theta(x_i^u))_j}}\right)\end{aligned} \tag{8}$$

二. (30 points) **概率图模型基础**

本题考查概率图模型中的隐马尔科夫模型和条件随机场相关内容。

1. (10 points) 概率图模型是一类用图来表达变量相关关系的概率模型，其中，隐马尔可夫模型 (Hidden Markov Model, HMM) 和条件随机场 (Conditional Random Field, CRF) 皆用于解决序列标注任务。请对比两种模型的区别。
2. (10 points) 条件随机场与对数几率回归 (Logistic Regression) 都可看作是对数线性模型 (Log-linear model) 的特例，假设 x 是样本点，

y 是标签，对数线性模型可写成如下的形式：

$$p(y \mid x; w) = \frac{\exp \sum_j w_j F_j(x, y)}{\sum_{y'} \exp \sum_j w_j F_j(x, y')} \tag{9}$$

其中，F_j 表示第 j 个特征函数，w_j 为对应的权重，分母部分为归一化项。请根据以上公式说明条件随机场和对数几率回归的差异。

3. (10 points) 在 HMM 的使用过程中会发现以下两个缺点：一、对于每个状态，HMM 模型只能捕捉到该状态对应观测的依赖关系，拿文本翻译任务来说，是需要考虑周围单词的信息，甚至是整段文本的信息；二、并且目标函数和预测目标函数并非匹配，具体指的是 HMM 学习了状态和观测的联合分布 $P(\mathrm{Y}, \mathrm{X})$，但在预测时使用的是条件概率 $P(\mathrm{Y} \mid \mathrm{X})$. (a) 具体说明最大熵马尔可夫模型 (Maximum Entropy Markov Model, MEMM) 如何解决以上两点，是否还有其他方法呢? (b) MEMM 中出现了新的问题：标签偏移问题 (label bias problem)，请说明为什么会出现该问题，并思考如何解决。参考文献: Maximum entropy Markov models for information extraction and segmentation.

解：

1. 隐马尔可夫模型与条件随机场的区别在于：
   (1) 隐马尔可夫模型是有向图模型；而条件随机场是无向图模型。
   (2) 隐马尔可夫模型是生成式模型，直接对联合分布建模；而条件随机场是判别式模型，对条件分布建模。
   (3) 隐马尔可夫模型中含有隐变量；而条件随机场则不含隐变量。
   (4) 隐马尔可夫模型使用矩阵定义参数；而条件随机场则使用函数定义参数。
   (5) 隐马尔可夫模型中观测变量的取值只依赖于当前时刻的状态变量，而当前时刻的状态变量只依赖于前一时刻的状态变量；而条件随机场中观测序列会影响标记变量，即标记变量与前面几个时刻的观测变量都有关。

2. 对数线性模型可写成如下的形式：

$$p(y \mid x; w) = \frac{\exp \sum_j w_j F_j(x, y)}{\sum_{y'} \exp \sum_j w_j F_j(x, y')} \tag{10}$$

具体而言，对于条件随机场，其模型可以表示为

$$p(\mathrm{y} \mid \mathrm{x}) = \frac{\exp\left(\sum_j \sum_{i=1}^{n-1} \lambda_j t_j\left(y_{i+1}, y_i, \mathrm{x}, i\right) + \sum_k \sum_{i=1}^{n} \mu_k s_k\left(y_i, \mathrm{x}, i\right)\right)}{\sum_{y'} \exp\left(\sum_j \sum_{i=1}^{n-1} \lambda_j t_j\left(y'_{i+1}, y'_i, \mathrm{x}, i\right) + \sum_k \sum_{i=1}^{n} \mu_k s_k\left(y'_i, \mathrm{x}, i\right)\right)} \tag{11}$$

对于对数几率回归，其模型可以表示为

$$\mathrm{p}(y=i|x) = \begin{cases} \dfrac{\mathrm{e}^{\mathrm{w_i^T x + b_i}}}{1+\sum_{j=1}^{\mathrm{K-1}} \mathrm{e}^{\mathrm{w_j^T x + b_j}}}, & i = 1, 2, 3, ..., K-1 \\ \dfrac{1}{1+\sum_{j=1}^{\mathrm{K-1}} \mathrm{e}^{\mathrm{w_j^T x + b_j}}}, & i = K \end{cases} \tag{12}$$

分析两种模型的表达式，发现分母都是归一化项，差别主要在于分子部分。分析分子的表达式可以看出，条件随机场与对数几率回归的差异主要在于：
条件随机场处理的是序列化数据（如时序数据），除样本数据外，其分子部分与观测序列和相邻标记变量都有关；而对数几率回归处理的是非序列化数据，除样本数据外，其分子部分仅与当前类别对应的参数有关。

3. (a) 对于第一点，MEMM 中当前时刻的状态变量不仅与当前时刻的观测变量有关，也与前一时刻的状态变量有关，即 $P(\mathrm{Y} \mid \mathrm{X}) = \prod_{i=1}^{n} P(y_i \mid y_{i-1}, x_i)$，建立了某个状态与其相邻状态的关系；对于第二点，MEMM 直接通过最大熵模型学习条件概率，故目标函数与预测目标函数匹配了。
也可以采用 CRF 的思路来改进 HMM，即抛弃标记仅与当前状态有关的假设，某时刻的标记变量会受到先前观测序列的影响，使模型能建模上下文信息；同样改为直接对条件概率建模；改用无向图模型；在所有状态上建立一个统一的概率模型。
(b) 因为 MEMM 做的是局部归一化，导致有更少转移的状态拥有的转移概率普遍偏高，MEMM 倾向于选择有更少转移的状态，故概率最大路径更容易出现有更少转移的状态。可以改用全局归一化来解决。

三. (40 points) **概率图模型进阶**

本题研究概率图模型里面的变分推断技术。现定义联合分布如下：

$$p(\mathrm{t}, \boldsymbol{w}, \alpha) = p(\mathrm{t} \mid \boldsymbol{w}) p(\boldsymbol{w} \mid \alpha) p(\alpha)$$

其中各具体分布为：

$$p(\mathrm{t} \mid \boldsymbol{w}) = \prod_{n=1}^{N} \mathcal{N}\left(t_n \mid \boldsymbol{w}^T \boldsymbol{\phi}_n, \beta^{-1}\right)$$

$$p(\boldsymbol{w} \mid \alpha) = \mathcal{N}\left(\boldsymbol{w} \mid 0, \alpha^{-1} \boldsymbol{I}\right)$$

$$p(\alpha) = \mathrm{Gam}\left(\alpha \mid a_0, b_0\right) = \frac{{b_0}^{a_0} \alpha^{a_0 - 1} e^{-b_0 \alpha}}{\Gamma(a_0)}$$

这里，$\mathcal{N}$ 代表高斯分布，$\mathrm{Gam}\left(\alpha \mid a_0, b_0\right)$ 表示变量为 α，参数为 a_0, b_0 的 Gamma 分布。

1. (10 points) 请使用盘式记法表示联合分布 $p(\mathrm{t}, \boldsymbol{w}, \alpha)$。
2. (20 points) 现在需要寻找对后验概率分布 $p(\boldsymbol{w}, \alpha \mid \mathrm{t})$ 的一个近似。使用变分框架进行分解，得到变分后验概率分布的分解表达式为 $q(\boldsymbol{w}, \alpha) = q(\boldsymbol{w}) q(\alpha)$。首先计算 $q^*(\alpha)$，考虑 α 上的概率分布，利用教材公式（14.39），只保留与 α 有函数依赖关系的项，试证明

   $$\ln q^*(\alpha) = (a_0 - 1) \ln \alpha - b_0 \alpha + \frac{M}{2} \ln \alpha - \frac{\alpha}{2} \mathbb{E}\left[\boldsymbol{w}^T \boldsymbol{w}\right] + \text{常数}$$

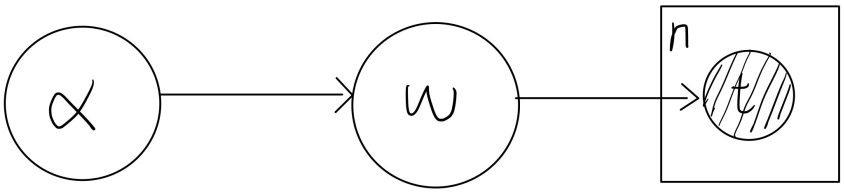
   这里 M 表示与 $\boldsymbol{w}$ 和 α 无关的常数。
3. (10 points) 由上一题得到的结果，观察发现这是 Gamma 分布的对数，因此 $q^*(\alpha)$ 仍然服从如下 Gamma 分布

   $$q^*(\alpha) = \mathrm{Gam}\left(\alpha \mid a_N, b_N\right)$$

   问：(a) 计算 a_N, b_N 的具体值（提示：由上一题计算 Gamma 分布对数 $\ln p(\alpha)$ 的结果观察 α 和 $\ln \alpha$ 的系数便可简单解决）(b) 比较 $q^*(\alpha)$ 与 $p(\alpha)$ 的相似度，总结变分推断的效果。（简单作答即可）

解：

1. 联合分布 $p(\mathrm{t}, \boldsymbol{w}, \alpha)$ 使用盘式记法可表示为



2. 证明如下：

$$\begin{aligned}\ln p(\alpha) &= \ln\left(\frac{{b_0}^{a_0}\alpha^{a_0-1}e^{-b_0\alpha}}{\Gamma(a_0)}\right)\\ &= \ln\left({b_0}^{a_0}\alpha^{a_0-1}e^{-b_0\alpha}\right) - \ln\Gamma(a_0)\\ &= a_0\ln b_0 + (a_0-1)\ln\alpha - b_0\alpha - \ln\Gamma(a_0) \qquad (13)\end{aligned}$$

$$\begin{aligned}\ln p(\boldsymbol{w}\mid\alpha) &= \ln\left(\frac{1}{\sqrt{2\pi\alpha^{-1}\boldsymbol{I}}}e^{-\frac{\boldsymbol{w}^T\boldsymbol{w}}{2\alpha^{-1}\boldsymbol{I}}}\right)\\ &= -\frac{\boldsymbol{w}^T\boldsymbol{w}}{2\alpha^{-1}\boldsymbol{I}} - \frac{1}{2}\ln\left(2\pi\alpha^{-1}\boldsymbol{I}\right)\\ &= -\frac{\boldsymbol{w}^T\boldsymbol{w}}{2\alpha^{-1}\boldsymbol{I}} - \frac{1}{2}\ln 2\pi + \frac{1}{2}\ln\alpha - \frac{1}{2}\ln\boldsymbol{I} \qquad (14)\end{aligned}$$

由教材公式 (14.39) 可得

$$
\begin{aligned}
\ln q^*(\alpha) &= \mathbb{E}_{i\neq j}\left[\ln p\left(\boldsymbol{w}, \alpha\right)\right] + const \\
&= \left(a_0 \ln b_0 + \left(a_0 - 1\right)\ln\alpha - b_0\alpha - \ln\Gamma(a_0)\right) \\
&\quad + \left(-\frac{\alpha}{2}\mathbb{E}\left[\boldsymbol{w}^T\boldsymbol{w}\right] - \frac{1}{2}\ln 2\pi + \frac{1}{2}\ln\alpha - \frac{1}{2}\ln\boldsymbol{I}\right) + const \\
&= \left(a_0 - 1\right)\ln\alpha - b_0\alpha + \frac{1}{2}\ln\alpha - \frac{\alpha}{2}\mathbb{E}\left[\boldsymbol{w}^T\boldsymbol{w}\right] \\
&\quad + \left(a_0 \ln b_0 - \ln\Gamma(a_0) - \frac{1}{2}\ln 2\pi - \frac{1}{2}\ln\boldsymbol{I}\right) + const \\
&= \left(a_0 - 1\right)\ln\alpha - b_0\alpha + \frac{M}{2}\ln\alpha - \frac{\alpha}{2}\mathbb{E}\left[\boldsymbol{w}^T\boldsymbol{w}\right] + \text{常数}
\end{aligned}
\tag{15}
$$

原式得证。

3. (a) 由第 2 问中计算 Gamma 分布对数 $\ln p(\alpha)$ 的结果可得 $a_N = a_0 + \frac{M}{2},\ b_N = b_0 + \frac{\mathbb{E}\left[\boldsymbol{w}^T\boldsymbol{w}\right]}{2}$。
   (b)

(1) $q^*(\alpha)$ 与 $p(\alpha)$ 的相似度为可用 KL 散度来刻画。

$$
\begin{aligned}
D_{KL}(p,q^*) &= \int_\alpha p(\alpha)\ln\frac{p(\alpha)}{q^*(\alpha)}d\alpha \\
&= \int_\alpha p(\alpha)\left(\ln p(\alpha)-\ln q^*(\alpha)\right)d\alpha \\
&= \int_\alpha \left(p(\alpha)\ln p(\alpha)-p(\alpha)\ln q^*(\alpha)\right)d\alpha \\
&= \int_\alpha (\frac{{b_0}^{a_0}\alpha^{a_0-1}e^{-b_0\alpha}}{\Gamma(a_0)}\left(a_0\ln b_0+(a_0-1)\ln\alpha-b_0\alpha-\ln\Gamma(a_0)\right) \\
&\quad -\frac{{b_0}^{a_0}\alpha^{a_0-1}e^{-b_0\alpha}}{\Gamma(a_0)}((a_0-1)\ln\alpha-b_0\alpha+\frac{M}{2}\ln\alpha-\frac{\alpha}{2}\mathbb{E}\left[\boldsymbol{w}^T\boldsymbol{w}\right] \\
&\quad +\ \text{常数}))d\alpha \\
&= \int_\alpha (\frac{{b_0}^{a_0}\alpha^{a_0-1}e^{-b_0\alpha}}{\Gamma(a_0)}(a_0\ln b_0+(a_0-1)\ln\alpha-b_0\alpha-\ln\Gamma(a_0) \\
&\quad -(a_0-1)\ln\alpha+b_0\alpha-\frac{M}{2}\ln\alpha+\frac{\alpha}{2}\mathbb{E}\left[\boldsymbol{w}^T\boldsymbol{w}\right]-\ \text{常数}))d\alpha \\
&= \int_\alpha (\frac{{b_0}^{a_0}\alpha^{a_0-1}e^{-b_0\alpha}}{\Gamma(a_0)}(a_0\ln b_0+-\ln\Gamma(a_0) \\
&\quad -\frac{M}{2}\ln\alpha+\frac{\alpha}{2}\mathbb{E}\left[\boldsymbol{w}^T\boldsymbol{w}\right]-\ \text{常数}))d\alpha \qquad (16)
\end{aligned}
$$

(2) 也可直接分析 Gamma 分布 $q^*(\alpha)$ 与 $p(\alpha)$ 的参数，即 a_N, b_N 与 a_0, b_0，又 $a_N = a_0 + \frac{M}{2}$，$b_N = b_0 + \frac{\mathbb{E}\left[\boldsymbol{w}^T\boldsymbol{w}\right]}{2}$，由之前的计算可知 $M=1$，且 $\boldsymbol{w}$ 服从均值为 0，方差为 $\alpha^{-1}\boldsymbol{I}$ 的高斯分布，故 $\mathbb{E}\left[\boldsymbol{w}^T\boldsymbol{w}\right]$ 也不大。$q^*(\alpha)$ 与 $p(\alpha)$ 的参数相差不大，变分推断的效果较好。