姓名：张三
学号：1234567

## 一. (20 points) **神经网络基础**

给定训练集 $D = \{(\boldsymbol{x}_1, \boldsymbol{y}_1), (\boldsymbol{x}_2, \boldsymbol{y}_2), ..., (\boldsymbol{x}_m, \boldsymbol{y}_m)\}$. 其中 $\boldsymbol{x}_i \in \mathbb{R}^d$, $\boldsymbol{y}_i \in \mathbb{R}^l$ 表示输入示例由 $d$ 个属性描述, 输出 $l$ 维实值向量. 图 **??**给出了一个有 $d$ 个输入神经元、$l$ 个输出神经元、$q$ 个隐层神经元的多层神经网络, 其中输出层第 $j$ 个神经元的阈值用 $\theta_j$ 表示, 隐层第 $h$ 个神经元的阈值用 $\gamma_h$ 表示. 输入层第 $i$ 个神经元与隐层第 $h$ 个神经元之间的连接权为 $v_{ih}$, 隐层第 $h$ 个神经元与输出层第 $j$ 个神经元之间的连接权为 $w_{hj}$. 记隐层第 $h$ 个神经元接收到的输入为 $\alpha_h = \sum_{i=1}^{d} v_{ih} x_i$, 输出层第 $j$ 个神经元接收到的输入为 $\beta_j = \sum_{h=1}^{q} w_{hj} b_h$, 其中 $b_h$ 为隐层第 $h$ 个神经元的输出.

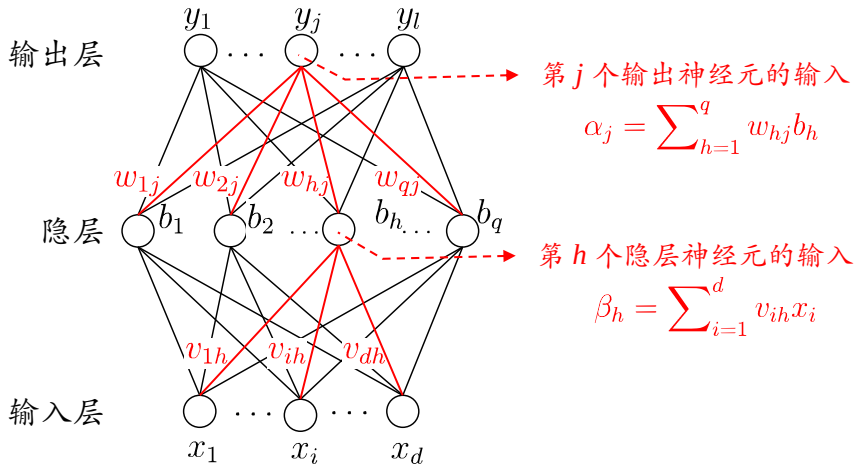


Figure 1: 多层神经网络（教材图 5.7）

不同任务中神经网络的输出层往往使用不同的激活函数和损失函数, 本题介绍几种常见的激活和损失函数, 并对其梯度进行推导.

1. 在二分类问题中 $(l = 1)$ , 标记 $y \in \{0, 1\}$, 一般使用 Sigmoid 函数作为激活函数, 使输出值在 $[0, 1]$ 范围内, 使模型预测结果可直接作为概率输出. Sigmoid 函数的输出一般配合二元交叉熵 (Binary Cross-Entropy) 损失函数使用, 对于一个训练样本 $(\boldsymbol{x}, y)$ 有

$$\ell(y, \hat{y}_1) = -\left[y \log(\hat{y}_1) + (1 - y) \log(1 - \hat{y}_1)\right] \tag{1}$$

记 $\hat{y}_1$ 为模型对样本属于正类的预测结果, 请计算 $\frac{\partial \ell(y, \hat{y}_1)}{\partial \beta_1}$,

2. 当 $l > 1$, 网络的预测结果为 $\hat{\boldsymbol{y}} \in \mathbb{R}^l$, 其中 $\hat{y}_i$ 表示输入被预测为第 $i$ 类的概率. 对于第 $i$ 类的样本, 其标记 $\boldsymbol{y} \in \{0,1\}^l$, 有 $y_i = 1$, $y_j = 0, j \neq i$. 对于一个训练样本 $(\boldsymbol{x}, \boldsymbol{y})$, 交叉熵损失函数 $\ell(\boldsymbol{y}, \hat{\boldsymbol{y}})$ 的定义如下

$$\ell(\boldsymbol{y}, \hat{\boldsymbol{y}}) = -\sum_{j=1}^{l} y_j \log \hat{y}_j \tag{2}$$

在多分类问题中, 一般使用 Softmax 层作为输出, Softmax 层的计算公式如下

$$\hat{y}_j = \frac{e^{\beta_j}}{\sum_{k=1}^{l} e^{\boldsymbol{\beta}_k}} \tag{3}$$

易见 Softmax 函数输出的 $\hat{\boldsymbol{y}}$ 符合 $\sum_{j=1}^{l} \hat{y}_j = 1$, 所以可以直接作为每个类别的概率. Softmax 输出一般配合交叉熵 (Cross Entropy) 损失函数使用, 请计算 $\frac{\partial \ell(\boldsymbol{y}, \hat{\boldsymbol{y}})}{\partial \beta_j}$,

3. 分析在二分类中使用 Softmax 和 Sigmoid 的联系与区别.
4. KL 散度 (Kullback-Leibler divergence) 定义了两个分布之间的距离, 对于两个离散分布 $Q(x)$ 和 $P(x)$, 其定义为

$$D_{\mathrm{KL}}(P \parallel Q) = \sum_{x \in \mathcal{X}} P(x) \log \left( \frac{P(x)}{Q(x)} \right) \tag{4}$$

其中 $\mathcal{X}$ 为 $x$ 的取值空间. 试分析交叉熵损失函数和 KL 散度的关系.

**解:**

二. (20 points) **运算的向量化**

在编程实践中, 一般需要将运算写成向量或者矩阵运算的形式, 这叫做运算的向量化 (vectorization). 向量化可以充分利用计算机体系结构对矩阵运算的支持加速计算, 大部分数学运算库例如numpy也对矩阵计算有专门的优化. 另一方面, 如果一个运算可以写成向量计算的形式, 会更容易写出其导数形式并进行优化. 本题中举两个简单的例子

1. 给定示例矩阵 $\boldsymbol{X} \in \mathbb{R}^{m \times d}$, 表示 $m$ 个示例 (向量), 每个示例有 $d$ 维, 计算 $m$ 个示例两两之间的距离矩阵 $\boldsymbol{D} \in \mathbb{R}^{m \times m}$, 两个向量之间的欧

式距离定义为 $\|\boldsymbol{x}-\boldsymbol{y}\|_2=\sqrt{\sum_{i=1}^{d}(x_i-y_i)^2}$. 求距离矩阵可以通过循环的方式, 即plain_distance_function中实现的方法;

```python
import numpy as np

def plain_distance_function(X):
    # 直观的距离计算实现方法
    # 首先初始化一个空的距离矩阵D
    D = np.zeros((X.shape[0], X.shape[0]))
    # 循环遍历每一个样本对
    for i in range(X.shape[0]):
        for j in range(X.shape[0]):
            # 计算样本i和样本j的距离
            D[i, j] = np.sqrt(np.sum((X[i] - X[j])**2))
    return D
```

2. 输入一个矩阵 $\boldsymbol{X}\in\mathbb{R}^{m\times d}$, 表示 $m$ 个向量, 每个向量有 $d$ 维, 要求对输入矩阵的行按照一个给定的排列 $\boldsymbol{p}=\{p_1,p_2,...,p_m\}$ 进行重新排列. 即输出一个新的矩阵 $\boldsymbol{X}'$, 其中第 $i$ 行的内容为输入矩阵的第 $p_i$ 行. 假设重排列为一个函数 perm 即 $\boldsymbol{X}'=\mathrm{perm}(\boldsymbol{X})$, 已知梯度 $\frac{\partial\ell}{\partial\boldsymbol{X}'}$, 需要计算 $\frac{\partial\ell}{\partial\boldsymbol{X}}$. 对矩阵的行进行排列可以采用简单的循环实现, 例如plain_permutation_function中的实现方法.

```python
import numpy as np

def plain_permutation_function(X, p):
    # 初始化结果矩阵，其中每一行对应一个样本
    permuted_X = np.zeros_like(X)
    for i in range(X.shape[0]):
        # 采用循环的方式对每一个样本进行重排列
        permuted_X[i] = X[p[i]]
    return permuted_X
```

请给出上述两种任务的向量化实现方案，并分析上述实现方法和向量化实现方法之间运行时间的差异。（提示：比如可以针对不同规模的矩阵大小来尝试分析主要操作的运行时间）

**解：**

三. (20 points) **支持向量机**

考虑标准的 SVM 优化问题如下 (即教材公式 (6.35)),

$$\begin{aligned}\min_{\boldsymbol{w},b,\xi_i}\quad & \frac{1}{2}\|\boldsymbol{w}\|^2+C\sum_{i=1}^{m}\xi_i\\ \text{s.t.}\quad & y_i\left(\boldsymbol{w}^\top\boldsymbol{x}_i+b\right)\geq 1-\xi_i\\ & \xi_i\geq 0, i\in[m].\end{aligned}\tag{5}$$

注意到, 在 (2.1) 中, 对于正例和负例, 其在目标函数中分类错误的“惩罚”是相同的. 在实际场景中, 很多时候正例和负例错分的“惩罚”代价是不同的（参考教材 2.3.4 节）. 比如考虑癌症诊断问题, 将一个确实患有癌症的人误分类为健康人, 以及将健康人误分类为患有癌症, 产生的错误影响以及代价不应该认为是等同的. 所以对负例分类错误的样本 (即 false positive) 施加 $k > 0$ 倍于正例中被分错的样本的“惩罚”. 对于此类场景下

1. 请给出相应的 SVM 优化问题.
2. 请给出相应的对偶问题, 要求详细的推导步骤, 如 KKT 条件等.

**解:**

四. (20 points) **核函数**

教材 6.3 节介绍了 Mercer 定理, 说明对于一个二元函数 $k(\cdot,\cdot)$, 当且仅当对任意 $m$ 和 $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m\}$, 它对应的 Gram 矩阵（核矩阵）是半正定的时, 它是一个有效的核函数. 核矩阵 $\boldsymbol{K}$ 中的元素为 $K_{ij} = \kappa(\boldsymbol{x}_i, \boldsymbol{x}_j)$. 请根据 Mercer 定理证明以下核函数是有效的.

1. $\kappa_3 = a_1\kappa_1 + a_2\kappa_2$, 其中 $a_1, a_2 \geq 0$.
2. $f(\cdot)$ 是任意实值函数, 由 $\kappa_4(\boldsymbol{x}, \boldsymbol{x}') = f(\boldsymbol{x})f(\boldsymbol{x}')$ 定义的 $\kappa_4$.
3. 由 $\kappa_5(\boldsymbol{x}, \boldsymbol{x}') = \kappa_1(\boldsymbol{x}, \boldsymbol{x}')\kappa_2(\boldsymbol{x}, \boldsymbol{x}')$ 定义的 $\kappa_5$.
4. 由 $\kappa_6(\boldsymbol{x}, \boldsymbol{x}') = f(\boldsymbol{x})\kappa_1(\boldsymbol{x}, \boldsymbol{x}')f(\boldsymbol{x}')$ 定义的 $\kappa_6$

**解:**

五. (20 points) **主成成分分析**

$\boldsymbol{x} \in \mathbb{R}^d$ 是一个随机向量, 其均值和协方差分别是 $\boldsymbol{\mu} = \mathbb{E}(\boldsymbol{x}) \in \mathbb{R}^d$, $\boldsymbol{\Sigma} = \mathbb{E}(\boldsymbol{x} - \boldsymbol{\mu}_{\boldsymbol{x}})(\boldsymbol{x} - \boldsymbol{\mu}_{\boldsymbol{x}})^\top \in \mathbb{R}^{d\times d}$. 定义随机变量 $\{y_i = \boldsymbol{u}_i^\top \boldsymbol{x} + a_i \in \mathbb{R}, i = 1, \ldots, d' \leq d\}$ 为 $\boldsymbol{x}$ 的主成分, 其中 $\boldsymbol{u}_i \in \mathbb{R}^d$ 是单位向量 ($\boldsymbol{u}_i^\top \boldsymbol{u}_i = 1$), $a_i \in \mathbb{R}$, $\{y_i\}_{i=1}^{d'}$ 是互不相关的零均值随机变量, 它们的方差满足 $\mathrm{var}(y_1) \geq \mathrm{var}(y_2) \geq \cdots \geq \mathrm{var}(y_{d'})$. 假设 $\boldsymbol{\Sigma}$ 没有重复的特征值.

1. 请证明 $\{a_i = -\boldsymbol{u}_i^\top \boldsymbol{\mu}\}_{i=1}^{d'}$.

2. 请证明 $\boldsymbol{u}_1$ 是 $\boldsymbol{\Sigma}$ 最大的特征值对应的特征向量. (提示: 写出要最大化的目标函数, 写出约束条件, 使用拉格朗日乘子法)
3. 请证明 $\boldsymbol{u}_2^\top \boldsymbol{u}_1 = 0$, 且 $\boldsymbol{u}_2$ 是 $\boldsymbol{\Sigma}$ 第二大特征值对应的特征向量. (提示: 由 $\{y_i\}_{i=1}^d$ 是互不相关的零均值随机变量可推出 $\boldsymbol{u}_2^\top \boldsymbol{u}_1 = 0$, 可作为第二小问的约束条件之一)
4. 通过 PCA 进行降维, 得到的随机变量满足 $\mathrm{var}(y_1) \geq \mathrm{var}(y_2) \geq \cdots \geq \mathrm{var}(y_d)$, 也就是降维后的数据在不同维度上有不同的方差, 从而导致不同维度的数值范围差异很大, 如果想要降维后的样本在不同维度具有大致相同的数值范围, 应该怎么做?

**解:**