作业 1

1. 泊松回归

在这个问题中，我们将构建一种常用的广义线性模型（GLM）——泊松回归。在GLM中，指数族分布的选择通常基于手头的问题类型。如果我们要解决分类问题，那么我们使用支撑集离散的指数族分布（例如伯努利或类别分布（Categorical Distribution））。类似地，如果输出是实值，我们可以使用高斯或拉普拉斯分布（它们都属于指数族）。有时，期望的输出是关于某个事件的预期发生次数，例如，根据输入特征（也称为协变量）预测一天内预期的电子邮件数量，或预测下一小时进入商店的预期顾客数量等。你可能记得，泊松分布是一个支撑集在整数（即发生次数）上的概率分布，它也属于指数族。

在接下来的子问题中，我们将首先说明泊松分布属于指数族，推导相应的函数形式和训练模型的更新规则。

(a)

考虑参数为λ的泊松分布：

$$p(y;\lambda) = \frac{e^{-\lambda}\lambda^y}{y!}.$$

（这里y取正整数数值，$y!$是y的阶乘。）证明泊松分布属于指数族，并明确给出$b(y), \eta, T(y),$ 和$a(\eta)$的值。（请看第二题指数族的定义）

(b)

考虑使用GLM模型进行回归，并使用泊松分布的响应变量。此时对应的规范响应函数（Canonical response function）是什么？（你可以使用参数为λ的泊松随机变量的均值为λ这一事实。）

(c)

对于训练集$\{(x^{(i)}, y^{(i)}); i = 1, \ldots, n\}$，一个样本的对数似然为$\log p(y^{(i)}|x^{(i)};\theta)$。将对数似然函数对$\theta_j$求导，推导出当使用GLM模型、响应变量$y$为泊松分布且使用规范响应函数时，随机梯度下降的更新规则。

2. 广义线性模型的凸性

在这个问题中，我们将探讨并证明广义线性模型的一些关键性质。更确切的说，是探讨当GLM的响应变量满足指数族分布的情形。

通常，GLM 的训练使用负对数似然（NLL）作为损失函数。这与最大似然估计大致等价（即，最大化似然等价于最小化负对数似然）。在这个问题中，我们的目标是说明 GLM 的 NLL 损失关于模型参数是一个凸函数。需要提醒的一点是，凸函数具有一个很好的性质，即任何局部最小值也是全局最小值，并且有大量关于如何使用各种算法（如梯度下降或随机梯度下降）高效优化凸损失函数的研究。

回顾一下，指数族分布的概率密度可以表示为：

$$p(y;\eta) = b(y)\exp\left(\eta^T T(y) - a(\eta)\right),$$

其中η是分布的自然参数。此外，在广义线性模型中，η被表示为$\theta^T x$，其中$x \in \mathbb{R}^d$是样本的输入特征，$\theta \in \mathbb{R}^d$是可学习的参数。为了说明GLM的NLL损失是凸函数，我们将这个过程分解为几个子部分，并逐一解决。我们的策略是说明损失函数关于模型参数的二阶导数（即 Hessian 矩阵）在所有模型参数值处都是半正定的（PSD）。我们还会证明指数族分布的一些良好性质作为中间步骤。

为了方便起见，我们只考虑η是标量的情况。假设

$$p(Y|X;\theta) \sim \text{ExponentialFamily}(\eta),$$

其中$\eta \in \mathbb{R}$是标量，且$T(y) = y$。此时指数族的表达式简化为：

$$p(y;\eta) = b(y)\exp\left(\eta y - a(\eta)\right).$$

(a)

推导分布的均值表达式，证明$\mathbb{E}[Y;\eta] = \frac{\partial}{\partial\eta}a(\eta)$（注意因为$\eta = \theta^T x$，$\mathbb{E}[Y;\eta] = \mathbb{E}[Y|X;\theta]$）。换句话说，证明指数族分布的均值是关于自然参数的对数配分函数(log-partition function)的一阶导数。

提示：首先观察到

$$\frac{\partial}{\partial\eta}\int p(y;\eta)dy = \int \frac{\partial}{\partial\eta}p(y;\eta)dy.$$

(b)

接下来，推导分布的方差表达式。具体来说，证明$\text{Var}(Y) = \frac{\partial^2 a(\eta)}{\partial\eta^2}$（注意$\text{Var}(Y;\eta) = \text{Var}(Y|X;\theta)$）。换句话说，证明指数族分布的方差是关于自然参数的对数配分函数（log-partition function）的二阶导数。

提示：利用上一子问题的结果可以简化推导过程。

(c)

最后，写出损失函数$\ell(\theta)$，即分布的NLL损失作为θ的函数。然后计算损失关于θ的Hessian矩阵，并证明它总是半正定的（PSD）。最终导出结论：GLM的NLL损失是凸。

提示1：使用链式法则和前面部分的结果来简化推导。

提示2：任何概率分布的方差都是非负的。

3. 多元最小二乘法

到目前为止，我们在课程中只考虑了目标变量y是标量的情况。假设我们现在不是试图预测单个输出，而是有一个对于每个样本有多个输出的训练集：

$$\{(x^{(i)}, y^{(i)}), i = 1, \ldots, m\},\ x^{(i)} \in \mathbb{R}^n,\ y^{(i)} \in \mathbb{R}^p.$$

此时，对于每个训练样本，$y^{(i)}$是一个有p个元素的向量。我们希望像最小二乘法一样使用线性模型来预测输出，通过以下参数矩阵Θ来指定：

$$y = \Theta^T x,$$

其中$\Theta \in \mathbb{R}^{n\times p}$。

(a)

这种情况下，损失函数是：

$$J(\Theta) = \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{p}\left((\Theta^T x^{(i)})_j - y_j^{(i)}\right)^2.$$

将$J(\Theta)$写成矩阵-向量的表示（即，不使用任何求和符号）。[**提示**：从$m \times n$设计矩阵开始：

$$X = \begin{bmatrix} (x^{(1)})^T \\ (x^{(2)})^T \\ \vdots \\ (x^{(m)})^T \end{bmatrix}$$

和$m \times p$目标矩阵：

$$Y = \begin{bmatrix} (y^{(1)})^T \\ (y^{(2)})^T \\ \vdots \\ (y^{(m)})^T \end{bmatrix}$$

然后推导如何用这些矩阵表示$J(\Theta)$。]

(b)

找到最小化$J(\Theta)$的闭式解Θ。

(c)

假设我们不是同时考虑多元向量$y^{(i)}$，而是分别计算每个$j = 1, \ldots, p$的变量$y_j^{(i)}$。在这种情况下，我们有p个单独的线性模型，形式为：

$$y_j^{(i)} = \theta_j^T x^{(i)},\ j = 1, \ldots, p.$$

（这里每个$\theta_j \in \mathbb{R}^n$。）这$p$个独立的最小二乘问题的参数与（2）中直接求的解有何区别？

4. 不完整标签的训练

在这个问题中，我们将考虑在标签不完全可见的情况下训练二元分类器。特别是，我们考虑一个在现实生活中并不罕见的场景，即我们只拥有部分正例的标签。所有负例和其余的正例都没有标签。

令$\{(x^{(i)}, t^{(i)})\}_{i=1}^n$是一个标准的独立同分布样本集。这里$x^{(i)}$是输入/特征，而$t^{(i)}$是标签。现在考虑如下情形，我们无法观察到$t^{(i)}$的值。相反，我们只能观察到部分正例的标签。具体来说，假设我们观察到的$y^{(i)}$是由以下方式生成的：

$$\forall x, \quad p(y^{(i)} = 1 \mid t^{(i)} = 1, x^{(i)} = x) = \alpha,$$

$$\forall x, \quad p(y^{(i)} = 0 \mid t^{(i)} = 1, x^{(i)} = x) = 1 - \alpha,$$

$$\forall x, \quad p(y^{(i)} = 1 \mid t^{(i)} = 0, x^{(i)} = x) = 0,$$

$$\forall x, \quad p(y^{(i)} = 0 \mid t^{(i)} = 0, x^{(i)} = x) = 1,$$

其中$\alpha \in (0,1)$是一个未知的标量。换句话说，如果未观察到的“真实”标签$t^{(i)}$是1，那么我们有α的概率观察到标签$y^{(i)} = 1$。另一方面，如果未观察到的“真实”标签$t^{(i)}$是0，那么我们总是观察到标签$y^{(i)} = 0$。

在问题中的最终目标是构建一个分类器h来预测真实标签t，只使用部分标签y。换句话说，我们希望构建h使得$h(x^{(i)}) \approx p(t^{(i)} = 1 \mid x^{(i)})$，只使用$x$和$y$。

在接下来的子问题中，我们将尝试在只有部分观测的情况下解决这个问题。也就是说，我们只能访问$\{(x^{(i)}, y^{(i)})\}_{i=1}^n$，并尝试预测$p(t^{(i)} = 1 \mid x^{(i)})$。

(a)

证明在我们的假设下，对于任何i，

$$p(t^{(i)} = 1 \mid y^{(i)} = 1, x^{(i)}) = 1\Theta$$

也就是说，观察到正标签$y^{(i)} = 1$时，我们可以确定隐藏的真实标签是1。注意使用贝叶斯定理进行推导（非数学解释不会得分）。

(b)

证明对于任何样本，真实标签$t^{(i)}$为正的概率是部分标签为正的概率的$1/\alpha$倍。也就是说，证明：

$$p(t^{(i)} = 1 \mid x^{(i)}) = \frac{1}{\alpha} \cdot p(y^{(i)} = 1 \mid x^{(i)}).$$

注意，上面的方程表明，如果我们知道α的值，那么我们可以通过乘以因子$1/\alpha$，将一个近似预测$p(y^{(i)} = 1 \mid x^{(i)})$的函数$h(\cdot)$转换为一个近似预测$p(t^{(i)} = 1 \mid x^{(i)})$的函数。

(c) 估计α

前一子问题中的解决方案需要知道α的值，但我们不知道它。现在我们将设计一种方法来基于函数$h(\cdot)$估计α，该函数可近似预测$p(y^{(i)} = 1 \mid x^{(i)})$。

为了简化分析，假设我们已经获得了一个完美预测$p(y^{(i)} = 1 \mid x^{(i)})$的函数$h(x)$，即$h(x^{(i)}) = p(y^{(i)} = 1 \mid x^{(i)})$。

我们做出一个关键假设，即$p(t^{(i)} = 1 \mid x^{(i)}) \in \{0, 1\}$。这个假设意味着生成“真实”标签$t^{(i)}$的过程是无噪声的。这个假设并非很不合理。注意，我们并未假设观测到的标签$y^{(i)}$是无噪声的，这将是一个不合理的假设!

现在我们将证明：

$$\alpha = \mathbb{E}[h(x^{(i)}) \mid y^{(i)} = 1]$$

为了证明这一点，说明当$y^{(i)} = 1$时，$h(x^{(i)}) = \alpha$。

（上述结果为以下算法提供了动机，通过估计上式右边来估计α：令V_+是验证集中标记为正的样本集合，即$V_+ = \{x^{(i)} \in V \mid y^{(i)} = 1\}$。然后我们使用如下估计： $\alpha \approx \frac{1}{|V_+|} \sum_{x^{(i)} \in V_+} h(x^{(i)})$.