Homework 01 Math

北京大学 2024 春季人工智能基础第一次课程作业

Arthals 2110306206

zhuozhiyongde@126.com

2024.03

1 第一问

> 请简述什么是贝叶斯定理，什么是最大似然估计（MLE），什么是最大后验估计（MAP）。

贝叶斯定理：

$$P(A|B) = \frac{P(A,B)}{P(B)} = \frac{P(B|A)P(A)}{P(B)} \tag{1}$$

也即，**后验概率** 等于 **似然概率** 乘以 **先验概率除以边缘概率** 。

最大似然估计（MLE）：找到一组最大的模型参数，使得在这组参数下，观测数据出现的概率最大。在上式中，如果认为 A 是模型参数，B 是观测数据，那么 MLE 的目标就是最大化似然概率 $P(B|A)$ ：

$$\theta_{MLE} = \arg\max_{\theta} P(B|\theta) \tag{2}$$

最大后验估计（MAP）：在 MLE 的基础上，由于我们并不可能遍历所有的参数空间，所以我们需要引入先验概率 $P(A)$ ，也即对参数做出假设。MAP 的目标是最大化后验概率 $P(A|B)$ ，由于我们已知数据，所以 $P(B)$ 是一个常数，因此最大化后验概率等价于最大化似然概率乘以先验概率 $P(B|A)P(A)$ 。

$$\theta_{MAP} = \arg\max_{\theta} P(B|\theta)P(\theta) \tag{3}$$

2 第二问

> 设 $X \sim N(\mu, \sigma^2), \mu, \sigma^2$ 为未知参数，$x_1, x_2, \ldots, x_n$ 是来自 X 的样本值，求 μ, σ^2 的最大似然估计量。

设 $x_1, x_2, \ldots, x_n$ 是从正态分布 $N(\mu, \sigma^2)$ 中抽取的样本，似然函数为：

$$L(\mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right) \tag{4}$$

对数似然函数为：

$$\ln L(\mu, \sigma^2) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2 \tag{5}$$

对 μ 和 σ^2 分别求偏导数，并令偏导数等于 0，可得：

$$\frac{\partial \ln L}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i) - \mu = 0 \tag{6}$$

$$\frac{\partial \ln L}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}(x_i - \mu)^2 = 0 \tag{7}$$

解得：

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{8}$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2 \tag{9}$$

也即， $\hat{\mu}$ 就是样本均值，$\hat{\sigma}^2$ 是样本方差。

3 第三问

> 请简述分类问题与回归问题的主要区别。

分类问题：对于一个输入 x ，预测它属于哪一个类别。分类问题的输出是离散的，通常是一个类别标签。最常见的做法是在隐藏层后接一个 Softmax 层，输出每个类别的概率，然后以概率最大者作为我们的预测分类。

回归问题：对于一个输入 x ，预测一个与之相关的连续值 y。回归问题的输出是连续的，通常是一个实数。最常见的做法是在隐藏层后接一个 $n \times 1$ 的全连接层，输出一个实数。

4 第四问

> 请简述有监督学习与无监督学习的主要区别

有监督学习：提供数据集 x 的同时提供对应的标签数据 y，构成 **输入输出对** ，通过学习输入输出对之间的关系，来预测新的输入数据的输出。有监督学习的目标是找到一个函数 f ，使得 $f(x) \approx y$ 。

无监督学习：只提供数据集 x，没有对应的标签数据。主要用于发现数据中隐含的结构或者模式，常见的问题包括聚类。

5 第五问

> 给定数据 $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, 用一个线性模型估计最接近真实 γ_i (ground truth) 的连续标量 Y, $f(x_i) = w^T x_i + b$, 使得 $f(x_i) \approx y_i$.
>
> 求最优 (w^*, b^*) 使得 $f(x_i)$ 与 y_i 之间的均方误差最小：
>
> $$(w^*, b^*) = \arg\min_{(w,b)} \sum_{i=1}^{n}(f(x_i) - y_i)^2 \tag{10}$$
>
> 并解释 (w^*, b^*) 何时有 closed form 解，何时没有 closed form 解。

（下述内容直接摘抄 / 修改自课堂笔记）

我们可以通过最小二乘法来获取最优的 w 和 b。

以线性代数来表示这个问题，也就是 Least Squares Estimator（最小二乘估计）：

$$\begin{aligned}(w^*, b^*) &= \arg\min \textstyle\sum_{i=1}^{n} (f(x_i) - y_i)^2 \\ &= \arg\min \textstyle\sum_{i=1}^{n} (y_i - wx_i - b)^2\end{aligned} \tag{11}$$

我们将之转为矩阵形式：

$$\mathbf{A} = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix} = \begin{bmatrix} X_1^{(1)} & \dots & X_1^{(p)} \\ \vdots & \ddots & \vdots \\ X_n^{(1)} & \dots & X_n^{(p)} \end{bmatrix}, \quad Y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \quad \beta = \begin{bmatrix} w_1 \\ \vdots \\ w_p \end{bmatrix} \tag{12}$$

注意这一步的 $\mathbf{A}$ 的每一行代表一个数据 X_i，每一行除了最后一列，都是 X_i 的特征，最后一列是 1，这是一个小的 trick，因为这样做的话，我们就可以把 b 合并到 β 中，也就是：

$$\mathbf{A} = \begin{bmatrix} X_1^{(1)} & \dots & X_1^{(p-1)} & 1 \\ \vdots & \ddots & \vdots & \vdots \\ X_n^{(1)} & \dots & X_n^{(p-1)} & 1 \end{bmatrix}, \quad \beta = \begin{bmatrix} w_1 \\ \vdots \\ w_{p-1} \\ b \end{bmatrix} \tag{13}$$

我们可以得到：

$$\begin{aligned}\hat{\beta} &= \arg\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} (X_i\beta - Y_i)^2 \\ &= \arg\min_{\beta} \frac{1}{n} (\mathbf{A}\beta - \mathbf{Y})^{\mathrm{T}} (\mathbf{A}\beta - \mathbf{Y})\end{aligned} \tag{14}$$

其中，$\hat{\beta}$ 是最优的 w，也就是我们要求的结果。

简化这个式子，去掉和优化无关的 $\frac{1}{n}$ ，我们可以得到：

$$\begin{aligned}J(\beta) &= (\mathbf{A}\beta - \mathbf{Y})^{\mathrm{T}} (\mathbf{A}\beta - \mathbf{Y}) \\ &= \beta^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{A}\beta - 2\beta^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{Y} + \mathbf{Y}^{\mathrm{T}}\mathbf{Y} \\ \frac{\partial J(\beta)}{\partial \beta} &= 2\mathbf{A}^{\mathrm{T}}\mathbf{A}\beta - 2\mathbf{A}^{\mathrm{T}}\mathbf{Y} = 0\end{aligned} \tag{15}$$

如果 $\mathbf{A}$ 可逆，那自然可以根据上式求出 β 的解：

$$\hat{\beta} = (\mathbf{A}^{\mathrm{T}}\mathbf{A})^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{Y} \tag{16}$$

但是，很多情况下，$\mathbf{A}$ 并不可逆（存在线性依赖），比如 $n < p$ 时，我们可以证明它一定不可逆（此时，数据点的数量小于特征的数量）。而且即使 $\mathbf{A}$ 可逆，当它的维度很大时，计算也是很昂贵的。所以我们经常使用梯度下降法来求解。

6 第六问

> Ridge regression 问题的解具有什么特点，为什么？
>
> Lasso regression 问题的解具有什么特点？为什么？

（下述内容直接摘抄 / 修改自课堂笔记）

这两个问题（惩罚项）解，具有如下特征：

- **L1 套索回归的正则化项**：$\lambda \sum_{j=1}^{p} |\beta_j|$，这是一个 L1 范数，它对所有系数施加相同的惩罚，这会导致一些系数直接为零，从而产生一个 **稀疏解**。**非零 w 更少**。
- **L2 岭回归的正则化项**：$\lambda \sum_{j=1}^{p} \beta_j^2$，这是一个 L2 范数，它对大的系数施加更大的惩罚，导致系数平滑地趋近于零。**有些 w 更小**。

7 第七问

> 请从 model function、loss function、optimization solution 三个方面比较 Linear regression 与 Logistic regression 的异同。

7.1 Model function

- 线性回归：$f(x) = w^T x + b$
- 逻辑回归：$f(x) = \frac{1}{1+e^{-w^T x - b}}$

7.2 Loss function

- 线性回归：$\frac{1}{n} \sum_{i=1}^{n} (f(x_i) - y_i)^2$，也即使用均方误差 MSE 作为损失函数。主要是衡量预测值与真实值之间的差异。
- 逻辑回归：$-\frac{1}{n} \sum_{i=1}^{n} y_i \log(f(x_i)) + (1 - y_i) \log(1 - f(x_i))$，也即使用交叉熵 Cross Entropy 作为损失函数。主要是衡量预测分布与真实分布之间的差异。

7.3 Optimization solution

- 线性回归：可以通过最小二乘法求解，也可以通过梯度下降法求解。
- 逻辑回归：一般只能通过梯度下降法求解。

8 第八问

> K - 近邻分类器的超参数是什么？怎么选择 K - 近邻分类器的超参数？

（下述内容直接摘抄 / 修改自课堂笔记）

K - 近邻分类器主要的超参数是 k ，即我们选择的邻居的个数。还有一个超参数是距离度量方式，包括欧氏距离、曼哈顿距离。

在训练过程中，我们将整个数据集划分为训练集（Training Set）、测试集（Test Set）和验证集（Validation Set）：

- 训练集：用于训练模型，让模型从数据中学习到特征。训练集通常是整个数据集的大部分，比如 70%~80%。
- 验证集：用于在训练过程中评估模型的性能，并调整超参数。验证集通常占整个数据集的一小部分，比如 10%~15%。
- 测试集：模型训练完成后，在测试集上评估模型的最终性能。测试集通常占整个数据集的 10%~20%，必须是模型从未见过的数据。（在选择参数时不考虑）

选择 k 的方法通常是通过交叉验证：将数据集分成几个部分，一部分用于训练，另一部分用于验证，并对不同的 k 值进行测试，选择使得验证集分类精度最高的 k 值。