人工智能基础

作业一

1. 贝叶斯定理是描述如何通过观察到的现象修正先验概率的定理。其数学表达式为

$$p(Y|X) = \frac{p(X|Y)p(Y)}{\sum_Y p(X|Y)p(Y)}.$$

其中 $p(Y|X)$ 是后验概率（posterior），$p(X|Y)$ 是似然（likelihood），$p(Y)$ 是先验概率（prior）。

最大似然估计是一种估计参数数值的方法。在最大似然估计中，我们通过调整参数的值，使得数据的似然 $p(\text{data}|\theta)$ 最大。

而最大后验估计是另一种估计参数数值的方法。最大后验估计中，我们主要依据贝叶斯公式，通过调整参数的值，使得数据的后验概率最大。在这一方法中，我们需要先假设一个先验分布 $p(\theta)$，而我们优化后验分布正比于先验分布乘似然 $p(\theta|\text{data}) \propto p(\text{data}|\theta)p(\theta)$。

2. 采样数据的似然为

$$p(x_1, \dots, x_n|\mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}.$$

取对数后求极值

$$\log(p) = \sum_{i=1}^{n} \left(-\frac{(x_i-\mu)^2}{2\sigma^2} - \log(\sigma) \right) + \text{const.}$$

极值处对 μ 和 σ 的偏导数为零，即

$$\frac{\partial \log(p)}{\partial \mu} = \sum_{i=1}^{n} \frac{x_i-\mu}{\sigma^2} = 0,$$

$$\frac{\partial \log(p)}{\partial \sigma} = \sum_{i=1}^{n} \left(\frac{(x_i - \mu)^2}{\sigma^3} - \frac{1}{\sigma} \right) = 0.$$

解得 μ 和 σ^2 的最大似然估计量为

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

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2.$$

其中 $\bar{x} = \frac{1}{n} \sum_i x_i$ 是样本的平均值。

3. 分类问题关注的是离散值的预测，模型输出通常是一个概率分布，我们会选择一个最高概率的类别作为分类结果；而回归问题关注的是连续值的预测，模型的输出量通常就是我们的预测结果。
4. 有监督学习训练集上的每一个样本都是有标签的，其目标是学习一个函数或者模型，该模型能够通过输入特征预测标签值，常见的有监督学习有分类问题、回归问题等；而无监督学习使用的是没有标签的数据集，其目标是发现数据的内在规律，常见的无监督学习有聚类、降维等。
5. 为方便运算，将 w 和 b 合并成一个新的参数向量 $\beta = (b, w^T)^T$，将所有数据排成一个矩阵

$$X = \begin{pmatrix} 1 & x_1^T \\ \vdots & \vdots \\ 1 & x_n^T \end{pmatrix},$$

那么我们需要最小化的函数就可以写为

$$J(\beta) = (X\beta - Y)^T (X\beta - Y).$$

均方误差取极值的时候，对 β 的偏导数为零

$$\frac{\partial J(\beta)}{\partial \beta} = 2X^T(X\beta - Y) = 0,$$

或者

$$(X^TX)\beta = X^TY,$$

假设 x_i 是一个 k 维的列向量，那么 X^TX 就是一个 $(k+1)\times(k+1)$ 的矩阵。当且仅当 X^TX 是一个满秩的矩阵，即 $\text{rank}(X^TX) = k+1$ 时，$\beta = (b, w^T)^T$ 才会有 closed form 的解:

$$\hat{\beta} = (X^TX)^{-1}X^TY.$$

否则，就不会有 closed form 解，需要借助正则化（regularization）才能得到参数的估计值。

6. Ridge regression 是在 loss function 后加上一个 L2 正则化项（即参数的平方和），其解通常都具有较小的系数，但通常不会缩减为 0，这是因为 L2 正则化会惩罚较大的系数。通过 Ridge regression 能够减少过拟合，提高模型稳定性。Lasso regression 是在 loss function 后加上一个 L1 正则化项（即参数的绝对值之和），其解通常都会比较稀疏，这是因为 L1 正则化项在零处的梯度是不连续的，这使得系数更容易缩减为 0。Lasso regression 能够在处理高维数据的时候更好的进行特征选择，提高模型的解释性。
7. 从 model function 来看，linear regression 的模型函数可以表示为 $y = w^Tx + b$；Logistic regression 的模型函数基于线性关系，但是通过 Sigmoid 函数映射到[0,1]区间，表示某个类别发生的概率。

   从 loss function 来看，虽然不同的方法用不同的损失函数在理论上都可行，但是 linear regression 通常会选择均方误差（MSE）作为 loss function，而 logistic regression 通常会选择交叉熵损失

（cross-entropy loss）作为 loss function。

从 optimization solution 来看，linear regression 通常会使用最小二乘法或者梯度下降法来求解参数，logistic regression 通常也会采用梯度下降法求解参数。

8. K-近邻分类器的超参数主要有两个：K 值（模型预测时，所选取的与待分类样本距离最近的样本数量），和 distance metric（如果度量距离，如 L1 距离、L2 距离等）。

选取超参数的一个常用方法是交叉验证（cross-validation）。通过划分数据集为训练集（training data）和验证集(validation data)或更多份的折叠数据集（folds），观察不同超参数对模型在验证集上性能的影响，最后选出使得模型在验证集上性能最佳的超参数。