人工智能基础第一次作业

刘卓洋 2300017729

1、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)。

解答:

1) 贝叶斯公式:

设有事件 A,B, 其发生的概率分别为 $P(A)$,$P(B)$, 且设 A 发生的条件下 B 发生的概率为 $P(A|B)$, 则有公式:

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

同样的，若有一系列事件 X，一系列事件 Y，则有:

$$P(Y|X) = \frac{P(X|Y)P(Y)}{\sum P(X|Y)P(Y)}(2)$$

贝叶斯公式在形式上刻画了两事件 X,Y 之间的概率与条件概率的关系，贝叶斯公式体现出其“预测性”，也即通过贝叶斯公式，可以在两独立事件之间建立关联性，借此预测其中某一时间发生的概率，在 AI 中，贝叶斯公式被用于最大似然估计 (MLE) 与最大后验估计 (MAP) 中。

2) 最大似然估计 (MLE):

最大似然估计是一种调整参数的逻辑，与频率学派相对应，其核心思想是:

给定一个概率分布 D, 并且已知其概率密度函数或概率质量函数 f_D, 以及一个分布参数 θ, 依托此分布, 当给定一组数据时, 即可得到对应的似然函数, 并对对应的参数进行调整, 以达到和真实值越发接近的目的，即:

i) 各个变量之间符合某种分布，且为独立同分布:

$$x_i \sim i.i.d(p(x|\theta)), p(X|) = \prod_{i=1}^{n} p(x_i|\theta)$$

ii) 此时我们希望得到的最优的参数 θ 即为:

$$\theta_{MLE} = \arg\max_{\theta} logp(x|\theta) = \arg\max_{\theta} \sum_{i=1}^{n} p(x_i|\theta)$$

(其中使用 log 的对数形式是为了避免多个属于 (0,1) 的概率相乘后的结果太小而无法保证

浮点数精度，因此对数化化为相加）
随后，可以通过优化方法来改进 θ，以使达到离结果最近的估计。
3) 最大后验估计 (MAP):
最大后延估计同样是一种调整参数的逻辑，与贝叶斯学派相对应，在 MLE 中，我们提到需要给定一个确定的概率分布 D，以及其概率密度函数或概率质量函数 f_D，以及一个分布参数 θ，但是，在 MAP 中，我们认为不存在这样一个已知的分布参数 θ，因此，其核心思想是：
认为 θ 是一个随机变量，用一个设好参数的 β 分布作为 θ 的先验分布 $p(\theta)$，即对参数有一个可能的信念，并基于此，由贝叶斯公式：

$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)}$$

我们刚才通过假设得到了 $P(\theta)$ 和 $P(X|\theta)$，而下方的 $P(X)$ 为归一化常量，以此可以求得 $P(\theta|X)$，称为后验分布，其可用于衡量我们估计的 θ 与真实的 θ 之间的差异，并对其调整以达到得到结果的目的。

2、设 $X \sim N(\mu, \sigma^2), \mu, \sigma$ 为未知参数,x_1, x_2, $\cdots$,x_n 是来自 X 的样本值，求 μ, σ 的最大似然估计量。
解：
LaTeX 代码有些难写，附图：

计算。

解：求 μ, σ^2 的最大似然估计量的基本思路是对 μ, σ 求偏导

∵ $x_1, x_2 \cdots x_n$ 为来自 X 的样本值，设 $x = (x_1, x_2 \cdots x_n)^T$

而 $X \sim N(\mu, \sigma^2)$ ∴ $x_i \sim \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

且 $X_1, X_2 \cdots X_n \sim i.i.d\, N(\mu, \sigma^2)$

$$\therefore p(x) = p(x_1)\, p(x_2) \cdots p(x_n) = \prod_{i=1}^{n} p(x_i) = \prod_{i=1}^{n} N(x_i)$$

$$\therefore \ln p(x) = \sum_{i=1}^{n} \ln\left[(2\pi)^{-\frac{1}{2}} \cdot \sigma^{-1} \cdot e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}\right]$$

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

$$= -\frac{1}{2\sigma^2}\cdot\sum_{i=1}^{n}(x_i-\mu)^2 - \frac{n}{2}\ln\sigma^2 - \frac{n}{2}\ln(2\pi)$$ 此为最大似然函数

对 μ 求偏导得

$$\frac{\partial[\ln p(x)]}{\partial\mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu) = 0 \qquad \therefore \mu_{ML} = \frac{\sum_{i=1}^{n} x_i}{n}$$

对 σ^2 求偏导得

$$\frac{\partial[\ln p(x)]}{\partial(\sigma^2)} = \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i-\mu)^2 - \frac{n}{2\sigma^2} = 0 \qquad \therefore (\sigma^2)_{ML} = \frac{\sum_{i=1}^{n}(x_i-\mu_{ML})^2}{n}$$

3、请简述分类问题与回归问题的主要区别。

解答：

1) 求解对象：回归问题是连续数值的预测，分类问题是离散数值的预测。

2) 求解目的：回归问题是用一个线性模型得出最接近真实值的一个结果，其目的是为了得到最接近真实值的预测值；而分类问题是用一个线性模型与分类函数，先提取特征，再进行分类，其目的是为了对目标进行区分。

3) 求解方法：回归问题通过最小二乘估计，对结果进行线性拟合；而分类问题在此基础上，利用一个逻辑函数进行分类化。

4) 目标函数的估计：回归问题利用最小二乘；而分类问题通过交叉熵衡量目标函数。

4、请简述有监督学习与无监督学习的主要区别。
解答:
主要体现在训练数据上，有监督学习的训练数据是被标注的，因此可以理解为是受人类监督的;而无监督学习的训练数据是没有任何标注的,因此可以理解为是不受人类监督的。(当然，其测试数据都是有标注的)

5、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 y_i (ground truth) 的连续标量 Y，$f(x_i) = w^T x_i + b$，such that $f(x_i) \approx y_i$。求最优 (w^*, b^*) 使得 $f(x_i)$ 与 y_i 之间的均方误差最小:

$$(w^*, b^*) = \arg\min_{(w,b)} \frac{1}{n} \sum_{i=1}^{n} (f(x_i) - y_i)^2$$

并解释 (w^*, b^*) 何时有 closed form 解，何时没有 closed form 解。
解答:
记 $\beta^* = \arg\min_{(w,b)} \frac{1}{n} \sum_{i=1}^{n} (f(x_i) - y_i)^2$,并且记矩阵 $A = [(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)]^T$,$\beta = [w_1, w_2, b]^T$，$Y = [y_1, \ldots, y_n]^T$，则有:

$$\beta^* = \arg\min_{(w,b)} \frac{1}{n} (A\beta - Y)^T (A\beta - Y)$$

令 $J(\beta) = (A\beta - Y)^T (A\beta - Y)$，则当

$$\frac{\partial J(\beta)}{\partial \beta}|_{\beta^*} = 2A^T A\beta - 2A^T Y = 0$$

时，可取得最优的 (w^*, b^*)
由

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

可得

$$(A^T A)\beta^* = A^T Y$$

因此，若 $A^T A$ 可逆，则

$$\beta^* = (A^T A)^{-1} A^T Y$$

这也即是对应的 closed form 解，有 closed form 解的条件为 $A^T A$ 可逆;
否则，我们应用正则化，应用 MAP 为 β^* 提供先验分布，再利用 Ridge regression 和 Lasso regression 求解。

6、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？

解答：

Ridge regression 的解最终求得 β 较小

原因：Ridge regression 采用对 β 采取的先验分布为二项分布的方法，即假设

$$\beta_{MAP} = \arg\min_{\beta} \sum_{i=1}^{n} (Y_i - X_i\beta)^2 + \lambda||\beta||_2^2$$

其惩罚 ζ 为 $\zeta = ||w||_2^2$，给定一个 ζ 值，则在坐标上体现为圆形，当与先验分布（二项分布）取最小二乘解时，其体现出均匀性，因此往往可以得到最优的解。（即 β 值最小的解）

Lasso 的解最终求得的 β 较分散

原因：Lasso 采用对 β 采取的先验分布为拉普拉斯分布的方法，即假设

$$\beta_{MAP} = \arg\min_{\beta} \sum_{i=1}^{n} (Y_i - X_i\beta)^2 + \lambda||\beta||_1$$

其惩罚 ζ 为 $\zeta = ||w||_1$，给定一个 ζ 值，则在坐标上体现为方形，当与先验分布（拉普拉斯分布）取最小二乘解时，其体现出非均匀性，解往往落在坐标轴上，导致其解向量中有很多维度变为 0，其他得到的最优解是分散的。

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

解答：

1)model function: 不同

Linear regression 采用最小二乘估计法，输出值可能为任意值

Logistic regression 采用最小二乘估计法后，采用一个 Logistic function 将结果转化为 0 到 1 的值

2)loss function: 不同

Linear regression 的 loss function 为 $f_n = \arg\min_f \frac{1}{n}\sum_{i=1}^{n}(f(x_i) - y_i)^2$(square error) 即最小二乘估计的逻辑。

Logistic regression 采用交叉熵为 loss function，即 $C(f(x^n), y^n) = -[y^n ln f(x^n) + (1 - y^n) ln(1 - f(x^n))]$(cross entropy)，其目的是为了避免优化过程中的梯度为 0 的问题。（梯度

即最小二乘法对 w_i 的导数，当 f(x) 被转化为 0,1 分布后，其梯度为 0）
3)optimization solution: 相同
均采用梯度下降法和反向传播，通过沿梯度自行下降，使参数趋向于使得 loss function 最小的值。

8、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?
解答：
K-近邻分类器为一种分类方法，输入没有标签（标注数据的类别），即没有经过分类的新数据，首先提取新数据的特征并与测试集中的每一个数据特征进行比较；然后从测试集中提取 K 个最邻近（最相似）的数据特征标签，统计这 K 个最邻近数据中出现次数最多的分类，将其作为新的数据类别。
其中的参数 Distance metric(测距方法) 和 K 即为超参数，因为其完全由分类器的制作者决定，且 Distance metric 和 K 的选取没有一个统一的合适的方法，在实际情况中，对 Distance metric 和 K 的选取采用以下方法：
1) 主要思想：尝试所有的数据，以确定 Distance metric 和 K 应该选取什么值，此时，我们需要训练数据与测试数据，但是我们不能使用真实的测试数据，因此需要将一部分训练数据划分为测试数据;
2) 将训练数据 n 登分为 $T_1, T_2, \ldots\ldots, T_n$ 取 T_i 作为测试数据，记为 $validation_i$，利用剩余的训练数据，逐一尝试 Distance metric 和 K 值，选出最优的 Distance metric 和 K_i，i 从 1 到 n，选取平均结果最好的 Distance metric 和 K，作为最合理的超参数 Distance metric 和 K。(一般的选用的 Distance metric 为 Euclidean distance 和 Manhattan distance 两种)