概率统计在机器学习一元回归问题中的应用

付容天
北京邮电大学计算机学院（国家示范性软件学院）
学号 2020211616　班级 2020211314

2021 年 11 月 17 日

1　总体叙述

1.1　概率统计在机器学习

机器学习问题从本质上就是利用概率论与数理统计、微积分、线性代数等数学知识，通过现代计算机强大的计算能力，让计算机能够通过“计算”来分析和理解事件，并作出相应的决策。一个合理的机器学习模型必须要在大部分情况下都能产生令人满意的预测结果，并且能够具有自我学习、更新原模型的能力。其中，微积分和线性代数的知识主要对数据进行处理，导出一个拟合模型；而概率论与数理统计的知识主要用于判断假设的正确性和模型拟合的优劣程度，并根据所得的统计量对已有模型进行修改，使之更加贴合实际情况，并且在预测的时候利用概率论与数理统计的知识对预测量的可靠性和预测区间进行分析，从而更好地服务现实应用。

1.2　概率统计在一元回归问题

在机器学习领域，一类重要的基本问题就是一元情况的回归分析问题。从一元情况的回归分析问题中发展出来的方法可以轻松地迁移到多元回归分析、逻辑回归等更加复杂的问题当中。一元情况的回归分析的基本流程是：1）观察实际问题是否属于一元情况的回归分析问题；2）给定输入数据集；3）根据数据集得到参数的估计值；4）分析参数估计值的各项统计学特征并以此判断设定的模型是否满意；5）根据 4 的结果对模型的设定进行修改，重复上述过程直到得到满意的参数估计值。

对于一个回归问题，完全反映其真实情况的回归方程称为**总体回归方程**；但是，受限于数据集规模、系统误差等多方面的原因，我们一般不能得到总体回归方程。在有限的数据集下得到的回归方程称为**样本回归方程**，它反映了对于给定的数据集，因变量与自变量之间的关系。在多数情况下，我们只能得到样本回归方程。以样本回归方程作为总体回归方程的

一个估计是有道理的，如果我们选定的解释变量足够解释该回归问题，那么其余的影响因素都可以认为是非主要因素，当数据集的规模足够大的时候，这些非主要因素往往呈现出确定的概率分布，这样，我们就在样本回归方程和总体回归方程之间建立了关系。需要注意的是，不仅各项误差属于非主要因素，其他对该回归问题影响程度远小于解释变量的因素也属于非主要因素。

我们现在已经选定了一个解释变量，并认为其他所有因素要么是误差、要么是影响程度远小于被选定的解释变量的因素，下面的论述将基于这个前提展开。除解释变量以外，其他的非主要因素都可以被看作是对真实情况的“扰动”，我们将这些扰动看作是一组相互独立的随机变量，其各自的期望均是 0，这是因为在我们的假设中，当数据集足够大时，这些随机变量对真实情况的影响几乎可以忽略。同理，我们也可以清楚地知道：所有随机变量的方差一定是一个有限数，否则就违背了我们论述的前提。那么，根据中心极限定理，无论我们的非主要因素有多少个具体的随机变量、也无论他们具体的分布如何，只要相互独立、各自的期望和方差存在并且一致有界，则其随机变量之和（也就是非主要因素全体）就符合正态分布：

$$\lim_{n\to+\infty} P\left\{\frac{\sum_{i=1}^{n}(X_i-\mu_i)}{\sqrt{\sum_{i=1}^{n}{\sigma_i}^2}}\le a\right\}\to\Phi(a)$$

根据数学期望和方差的性质便得到：

$$\frac{1}{n}\sum_{i=1}^{n}X_i\to N\left\{\mu,\frac{\sigma^2}{n}\right\},\qquad \frac{\frac{1}{n}\sum_{i=1}^{n}X_i-\mu}{\sigma/\sqrt{n}}\to N(0,1)$$

总结上面的论述，我们实际上就是为了解决一元回归问题的方便而给出了一组合理的假设，列出如下：

- **一元回归问题的唯一解释变量假设**：这是容易理解的，对于一元线性回归问题，我们可以选定唯一的解释变量
- **非主要因素的零均值假设**：扰动因素会使得真实情况上下浮动，由于我们已经选定了唯一的解释变量，那么扰动因素的影响就应该随着数据集规模的增加而逐渐减小，且这些非主要因素最终呈现出正态分布的特征
- **线性关系假设**：一元回归问题可以认为是求解选定的解释变量和被解释变量之间的线性方程的问题，即使对于给定的输入集，严格的线性关系可能会受到非主要因素的破坏，但我们依然认为一元回归问题是线性关系问题

上面的三个假设是合理的，在现实生活中，我们的经验常常可以验证这三个假设的正确性。

2 一元回归问题

根据上一部分的讨论，对于一元回归分析问题，我们采用线性模型来进行分析。选定一个解释变量之后，所有其他无关紧要的变量都被纳入了非主要因素中，并且随着输入集规模

的增加，各个非主要因素变量近似服从期望为 0 的正态分布。

总体回归方程使用系统的线性项和非系统的随机项构成，在理论上能够精确地描述总体回归问题，但是受限于多方面的因素，一般来说非系统的随机项服从的分布规律不能先验地知道，所以从理论上确定精确的总体回归方程是极为困难的。我们采用样本回归方程作为总体回归方程的一个近似，设总体回归方程为：

$$y_i = w_0 + w_1 x_i + u_i \tag{1}$$

其中 u_i 为非主要因素，并设样本回归方程为：

$$y_i = \hat{w_0} + \hat{w_1} x_i + e_i \tag{2}$$

其中 e_i 是线性模型 $\hat{y_i} = \hat{w_0} + \hat{w_1} x_i$ 和给定因变量 y_i 之间的差值，称为**残差**。

为了描述我们的线性模型的精确性，一个自然的想法就是计算给定的因变量和我们计算得到的估计因变量之间的关系，也就是计算偏差量 e_i 的总和，考虑到 e_i 的正负，我们计算 ${e_i}^2$ 的总和。最小二乘法的原理就是使得残差平方和最小，也就是使

$$\sum {e_i}^2 = \sum (y_i - \hat{y_i})^2 = \sum (y_i - \hat{w_0} - \hat{w_1} x_i)^2$$

最小，根据二元函数极值的知识，求此函数的极值需要对变量 $\hat{w_0}$ 和 $\hat{w_1}$ 求偏导并令其为零，最后解得：

$$\hat{w_0} = \frac{\sum {x_i}^2 \sum y_i - \sum x_i \sum x_i y_i}{n \sum {x_i}^2 - (\sum x_i)^2} \tag{3}$$

$$\hat{w_1} = \frac{n \sum x_i \sum y_i - \sum x_i \sum y_i}{n \sum {x_i}^2 - (\sum x_i)^2} \tag{4}$$

设 $\bar{x} = \frac{1}{n} \sum x_i$，$\bar{y} = \frac{1}{n} \sum y_i$，那么有：

$$\hat{w_0} = \bar{y} - \hat{w_1} \bar{x} \tag{5}$$

$$\hat{w_1} = \frac{\sum x_i y_i - n \bar{x} \bar{y}}{\sum {x_i}^2 - n \bar{x}^2} \tag{6}$$

方便起见，设 $x_i' = x_i - \bar{x}$ 与 $y_i' = y_i - \bar{y}$，并称其为**离差**，这样，上述表达式进一步表示成：

$$\hat{w_0} = \bar{y} - \hat{w_1} \bar{x} \tag{7}$$

$$\hat{w_1} = \frac{\sum x_i' y_i'}{\sum {x_i'}^2} \tag{8}$$

这样就得到了一元回归分析的参数估计。

我们这里得到的样本回归函数的参数估计，在统计学上可以证明是总体回归方程的参数的**无偏估计**，下面具体说明这一点。在统计学中，无偏估计指的是被估计量的数学期望等于被估计参数，该性质保证了样本回归函数对总体回归函数的近似关系，也进一步说明了用样本回归函数来近似总体回归函数的正确性和可行性。如前所述，各个非主要因素服从期望为 0 的正态分布。因为：

$$\bar{y} = \frac{1}{n}\sum y_i = w_0 + w_1 x + \bar{u}$$

所以变量 y 的离差可以表示为：

$$y_i' = y_i - \bar{y} = w_1 x_i + (u_i - \bar{u})$$

考虑到：

$$\hat{w_0} = \bar{y} - \hat{w_1}\bar{x}$$

则残差 e_i 可以表示为：

$$e_i = y_i - \hat{y_i} = y_i' + \bar{y} - w_0 - w_1(x_i + \bar{x}) = y_i' - \hat{w_1}x_i' = (u_i - \bar{u}) - (\hat{w_1} - w_1)x_i'$$

在 e_i^2 表达式的两侧同时取期望，不难得到：

$$E(\sum e_i^2) = E\big[(\hat{w_1} - w_1)\sum {x_i'}^2\big] + E\big[\sum (u_i - \bar{u})^2\big] - 2E\big[(\hat{w_1} - w_1)\sum x_i'(u_i - \bar{u})\big] \quad (9)$$

其中，$u_i \sim N(0, \sigma^2)$，那么可将上式的第二项进行化简如下：

$$E\big[\sum (u_i - \bar{u})^2\big] = (n-1)^2\sigma^2$$

因为 $\bar{y}$ 是一个常数，那么：

$$\sum x_i' y_i' = \sum x_i'(y_i - \bar{y}) = \sum x_i' y_i$$

将此式代入之前已经得到的 $\hat{w_1}$ 的表达式 (9)，可以得到：

$$\hat{w_1} = \frac{\sum x_i' y_i'}{\sum {x_i'}^2} = \frac{\sum x_i' y_i}{\sum {x_i'}^2} = \sum k_i y_i, \quad k_i = \frac{x_i'}{\sum {x_i'}^2}$$

从此式反映的意义来说，可以将 $\hat{w_1}$ 理解成给定数据集下的 y_i 的一种加权平均。那么，由 y_i 的表达式知道，y_i 由系统的线性项 $w_0 + w_1 x_i$ 和非系统的非主要因素 u_i 组成，那么 y_i 的概率分布就和 u_i 的概率分布同型，那么 $\hat{w_1}$ 作为 y_i 的一种加权平均，其概率分布也与 u_i 的概率分布同型，通过下面的推导：

$$\hat{w_1} = \sum k_i y_i = \sum k_i(w_0 + w_1 x_i + u_i) = w_0\sum k_i + w_1\sum k_i x_i + \sum k_i u_i = w_1 + \sum k_i u_i$$

便知道 $\hat{w_1}$ 的期望确实就是 w_1，因此属于无偏估计。进一步，将此结果代入 $E(\sum e_i^2)$ 的表达式，就能得残差平方和的期望：

$$E(\sum e_i^2) = (n-2)\sigma^2 \tag{10}$$

我们可以使用此式得到总体方差的一个无偏估计量：

$$\hat{\sigma}^2 = \frac{\sum(y_i - \hat{y}_i)^2}{n-2} \tag{11}$$

这个式子具有重要意义，它意味着可以使用有限数据集的输入来估计总体回归问题的方差。

由于非主要因素的影响，得到的参数估计可能并不精确，因此我们需要使用概率统计的知识对得到的参数估计进行分析，确定其可靠性。已经知道各项非主要因素变量近似服从正态分布，而且从参数估计 $\hat{w_0}$ 和 $\hat{w_1}$ 的表达式知道这两个参数估计的概率分布仅与非主要因素有关，且这两个参数估计为非主要因素的线性函数，因此这两个参数估计服从正态分布，根据概率统计的知识和上面的分析，不难得到：

$$\hat{w_0} \sim N\left(w_0, \frac{\sigma^2 \sum {x_i}^2}{n \sum {x_i'}^2}\right) \tag{12}$$

$$\hat{w_1} \sim N\left(w_1, \frac{\sigma^2}{\sum {x_i'}^2}\right) \tag{13}$$

这就是两个参数估计的概率分布，根据概率分布，我们可以得到两个参数估计的一些统计学特征，比如，可以知道估计量 $\hat{w}$ 和理论值 w 相差不超过 3σ 的概率为 99.73%。

3 参数分析与假设检验

在前面的叙述中，我们已经得到了一元线性回归问题中的参数估计，现在的问题就是如何衡量我们得到的参数估计的可信程度，这种对模型优劣的检测思路和方法是十分重要的。下面我们将进一步通过概率统计的知识来对我们得到的结果进行分析。

与之前引入残差平方和的思路类似，我们需要找到合适的衡量手段来进行分析。将接下来要使用的三个概念先列出如下：

$$\sum {y_i'}^2 = \sum (y_i - \bar{y})^2$$

称为**离差平方和**，记为 J

$$\sum e_i^2 = \sum (y_i - \hat{y}_i)^2$$

称为**残差平方和**，记为 K

$$\sum {\hat{y}_i}^2 = \sum (\hat{y}_i - \bar{y}_i)^2$$

称为**回归平方和**，记为 L。当数据集给定的时候，J 的值是确定的，而 $J = K + L$，并且 K 和 L 在 J 中的占比是由我们估计的参数所决定的，那么，为了使残差平方和 K 尽可能小，实际上就是要找到合适的参数估计，使回归平方和 L 在此参数估计下尽可能地大。引入判定系数 $R^2 = \frac{L}{J}$ 来判断模型的优劣程度，如果 R^2 越大，那么说明回归平方和在离差平方和中的占比越大，从而残差平方和在离差平方和中的占比越小，此时，我们认为模型拟合程度较好。如果 R^2 越小，那么说明模型拟合程度较差。

以上讨论了在已知线性关系的条件下，对模型进行拟合优劣成程度的检验。但有些时候，我们可能错误地判断出线性关系、或者选定的解释变量不足以解释该回归问题，下面要讨论的办法就是用定量分析的方法来确定假设是否可靠。一般情况下，离差平方和 J 中的因变量有 n 个被观测值，该 n 个被观测值受到式 $\sum y_i' = (y_i - \bar{y}) = 0$ 的制约，因此 J 的自由度为 $n-1$。回归平方和 L 的自由度取决于选择的解释变量的个数，在一元回归问题中，L 的自由度就为 1。因此残差平方和 K 的自由度就为 $n-2$。根据前面的多处分析，不难知道 $\hat{y_i}$ 和 e_i 都是服从正态分布的，那么由概率统计的知识就得到它们的平方和就服从 χ^2 分布，即为：

$$\sum {\hat{y_i}}^2 \sim \chi_1^2 \tag{14}$$

$$\sum e_i^2 \sim \chi_{n-2}^2 \tag{15}$$

故 $\sum {\hat{y_i}}^2$ 和 $\sum e_i^2$ 经过自由度处理之后的比值就符合 F 分布，即为：

$$F = \frac{(\sum {\hat{y_i}}^2)/1}{(\sum e_i^2)/(n-2)} \sim F(1, n-2) \tag{16}$$

接下来将要引入 F **检验法**来判断假设是否合理。F 检验法的原理是对方差的齐性进行检验，理想情况下，我们选定的假设变量足以解释此时的回归问题。现假设选中的解释变量不能解释此问题，即 $w_1 = 0$，在此假设下计算统计量 F。给定 $\alpha = 0.05$，并查表得到分位点 $F_{0.05}(1, n-2)$，此时的含义为：统计量 F 大于分位点 $F_{0.05}(1, n-2)$ 的概率仅为 0.05，因此，如果一次计算得到的 F 的值确实大于 $F_{0.05}(1, n-2)$，那么就有足够的把握认为假设错误，即选中的变量可以解释此时的一元线性回归问题。

值得一提的是，上面讨论的 F 检验法可以平滑地迁移到多元情况的回归分析问题。

4 预测区间

最后简要地讨论一元回归模型的预测问题。

对于给定的输入 x_0，我们可以得到一个输出 y_0，并以此作为 x_0 的预测，但是，有时候我们需要知道预测的可靠程度，这时就需要引入**预测区间**的概念。设误差 $e_0 = y_0 - \hat{y_0}$，由于 $\hat{y_0}$ 服从正态分布，那么 e_0 也服从正态分布，可以得到：

$$E(e_0) = E(y_0 - \hat{y_0}) = 0 Var(e_0) = Var(y_0 - \hat{y_0}) = Var(y_0) + Var(\hat{y_0})$$

注意到 y_0 和 $\hat{y_0}$ 相互独立又有：

$$Var(y_0) = Var(w_0 + w_1 x_0 + u_0) = Var(u_0) = \sigma^2 \tag{17}$$

$$Var(\hat{y_0}) = E[\hat{y_0} - E(y_0)]^2 = \sigma^2 \big[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum {x_i'}^2}\big] \tag{18}$$

根据抽样理论，标准化之后的 e_0 服从自由度为 $n-2$ 的 T 分布，也即：

$$\frac{e_0}{\sigma\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum {x_i'}^2}}} \sim N(0,1) \tag{19}$$

那么就能得到 y_0 在显著水平 α 下的预测区间为：

$$\hat{y_0} - \sigma \cdot t_{\frac{\alpha}{2}} \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum {x_i'}^2}} \le y_0 \le \hat{y_0} + \sigma \cdot t_{\frac{\alpha}{2}} \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum {x_i'}^2}} \tag{20}$$

5 总结

概率论与数理统计的知识在机器学习领域中的应用极为广泛，本文只是简单探讨了一元回归分析问题中的一小部分内容。机器学习问题的各个阶段都离不开概率统计的知识，从数据输入到数据处理、从模型检验到模型修正……概率统计既针对机器学习中的随机性给出了分析方法，也针对统计数据提供了分析方法。概率统计在机器学习领域的应用之广，让我们不禁感叹机器学习其实就可以认为是概率统计的一种应用。