Linear Regression

narutoacm

2015 年 1 月 4 日

Contents

1 回归问题 (Regression Problem)

回归是研究自变量和因变量关系的问题。设 X 是自变量，Y 是因变量，其对应的关系用函数 f 表示，即有 $Y = f(X)$。因此，理想情况下，每个 X 的值都只可能出现一个 Y 值。例如 $f(X) = X^2$，那么如果 $X = 2$，那么 Y 一定为 4。但是，由于噪声等干扰的影响，导致产生了不确定性，即每个 X 的值，可能出现多个不同的 Y 值。如上例的情况，在 $X = 2$ 时，$Y = 4.2$ 也是可能的，$Y = 8$ 也是可能的，但是，可以想象，出现后一种情况的概率应该很低。这样，由于噪声影响，X 与 Y 的关系不在是固定的 f，而是 $Y = f(X) + \epsilon$，其中 ϵ 是由于噪声引起的，它具有不确定性。这种不确定性我们可以用概率来描述，即 X，Y 满足联合分布 $P(X, Y)$。

1.1 预测函数

现在假如我们知道这个联合分布 $P(X, Y)$，对 X 的一个值 x，我们应该如何去猜测对应的 y 呢？如果没有噪声干扰的话，知道 $P(X, Y)$ 也就相当于知道了它们的关系 $f(X)$，我们也可以确定对应的 y 一定等于 $f(x)$。但是由于已经有不确定性，你就不能这么简单的猜了。首先是 $f(X)$ 没法知道，其次，即使知道 $f(X)$，你也不能简单的猜结果就是 $f(x)$，因为实际上结果是 $f(x) + \epsilon$，这是个不确定的值。也就是，我们需要找一个“确定的”函数 $g(x)$，让它来**代表**“不确定的”函数 $f(x) + \epsilon$。对每个 x，我们就猜它对应的 Y 值是 $g(x)$。

1.2 损失函数与期望损失

怎么找呢？一种简单但是很自然的想法是，我们会想取得使 $P(Y|X = x)$ 最大的那个 y，因为这个 y 值是最可能出现的，也就是让 $g(x) = \arg\max_y P(y|X = x)$。但是这样做就完全排除了其他值的可能，比如虽然某个 Y 值的概率不是最大的，但它也可能的。一种合理的办法是，我们先定义一个损失函数 (Loss function)$L(y, g(x))$，它代表我们的猜测 $g(x)$ 与实际出现的值之间的差异，当然差异越大，这个损失也越大。如一种很常见的损失函数叫做平方损失 (square loss)

$$L_{square}(y, g(x)) = (y - g(x))^2 \tag{1}$$

然后，我们需要这个损失函数的期望值最小，即

$$E(L) = \int\int L(y, g(x))p(x, y)\,\mathrm{d}x\mathrm{d}y \tag{2}$$

要求的即是

$$g^*(x) = \arg\min_{g(x)} E(L) \tag{3}$$

以上的 $E(L)$ 叫做期望损失或期望风险 (expected loss)，这个思想叫做最小化期望风险准则。

为什么以这个 $g^*(x)$ 做代表要更有说服力？我们看到，如果某个 $p(x,y)$ 很大，同时 $L(y,g(x))$ 很小的话，那整个期望也会比较小。这符合我们之前在上文的一个简单想法作出的猜测，即概率越大它越可能是。但同时，它也考虑了其他的可能，并不像上面那个简单想法一样，把其他可能全都排除掉。

1.3 平方损失与回归函数

那从现在开始，我们只考虑 L_{square} 的情况，先来看看在这个情况下，最佳的代表 g^* 应该是什么样的。

$$\begin{aligned} E(L_{square}) &= \int\int (y-g(x))^2 p(x,y)\,\mathrm{d}x\mathrm{d}y \\ &= \int\int (y-E(Y|X=x)+E(Y|X=x)-g(x))^2 p(x,y)\,\mathrm{d}x\mathrm{d}y \\ &= \int\int (y-E(Y|X=x))^2\,\mathrm{d}x\mathrm{d}y + \\ &\quad \int\int (E(Y|X=x)-g(x))^2\,\mathrm{d}x\mathrm{d}y + \\ &\quad \int\int 2(y-E(Y|X=x))(E(Y|X=x)-g(x))]p(x,y)\,\mathrm{d}x\mathrm{d}y \end{aligned} \tag{4}$$

先看第 1 部分，$\int\int (y-E(Y|X=x))^2 p(x,y)\,\mathrm{d}x\mathrm{d}y$，它和 $g(x)$ 无关，已知分布 $P(X,Y)$ 时，这部分的值就是固定的，它是“内在的损失”，是由噪声引起的。

再看第 3 部分，常数项 2 不管，

$$\begin{aligned} & \int\int (y-E(Y|X=x))(E(Y|X=x)-g(x))]p(x,y)\,\mathrm{d}x\mathrm{d}y \\ = & \int \left(\int (y-E(Y|X=x))p(y|x)\,\mathrm{d}y\right)(E(Y|X=x)-g(x))p(x)\,\mathrm{d}x \\ = & \ 0 \end{aligned} \tag{5}$$

最后是第 2 部分，由于被积函数是个平方项，因此这部分的值至少是非负的。而如果要使整个 $E(L)$ 最低，那令这部分的值等于 0 的时候就是我们能做到的极限了，也就是令

$$g^*(x) = E(Y|X=x) \tag{6}$$

这个 g 才能使得我们的 $E(L)$ 的值最低。上述这个函数非常重要，统计学里专门把这个函数叫做“回归函数”(Regression Function)。可以看到，在没有噪声干扰，或噪声的期望是 0 时，$g^*(x) = f(x)$。即在这种情况下，我们以最小化期望平方损失为标准所找到的最佳代表就是它原本的“真正的关系”$f(x)$。而一

般噪声的期望就可以认为是 0，所以我们就可以认为上述函数 $g^*(x)$ 就是真正的 $f(x)$。

1.4 统计推断与决策理论

如果知道了 $P(X,Y)$，那 $E(Y|X)$ 也当然可以求出来。但是实际上我们不可能知道这个分布，于是我们就需要做“推断”(inference) 了。我们需要从这个分布里抽取的某些样本 $(X_1,Y_1),\ldots,(X_n,Y_n)$ 中，推断出 $P(X,Y)$。但实际上，我们只需要推断出 $P(Y|X)$ 就行了，因为知道了这个条件概率就足够我们求出 $E(Y|X)$ 了。这样就有了 2 种不同的求回归函数的手段，一种是先推断出 $P(X,Y)$，有了联合分布就可以得到 $P(Y|X)$，继而求出回归函数，这种方式叫做“生成方法”(generative method)。另一种是直接推断出 $P(Y|X)$，然后求出回归函数，这种方式叫做“判别式方法”(discriminant method)。

所以，回归实际上做的就是，从样本中推断一个 $g^*(x)$ 的估计 $h(x)$，这个 $h(x)$ 就是最终我们所得到的目标函数。实际上有一个这样的关系：

$$f(x)+\epsilon \approx_{\text{probably generaly equally}} g^*(x) \approx_{\text{depend on the estimator}} h(x)$$

前一个约等式是决策理论的工作，选定某种损失函数，我们选一个最佳代表。后一个约等式是统计推断的工作，给定样本，去推断这个最佳代表。

2 线性回归 (Linear Regression)

线性回归做了以下 2 个假设。1) 假设 $f(x)$ 是 x 的线性函数，即 $f(x)=w^Tx+w_0$，其中 w 与 w_0 是参数。2) 假设噪声 ϵ 和因变量 X 是独立的，即不受 X 的值的不同而改变分布，并且它分布的期望为 0，方差是 σ^2。为了方便，以下假设给自变量扩充第 0 维，值为 1，同时把 w_0 一起写到整个向量 w 中，这样在不引起误解的情况下，线性关系直接写作 $f(x)=w^Tx$。注意这个式子中的 w 和 x 都扩充了一维。

2.1 最大似然估计与最小二乘估计

为了推断 $P(Y|X)$，我们进一步假设噪声的分布为正态分布 $N(0,\sigma^2)$。因此，由于 $Y=f(X)+\epsilon$，我们有

$$P(Y|X=x)=N(Y|w^Tx,\sigma^2) \tag{7}$$

用最大似然法去估计 w，给定样本 $(X_1, Y_1), \ldots, (X_n, Y_n)$，似然函数为

$$L(w) = \prod_{i=1}^{n} N(Y_i | w^T X_i, \sigma^2) \tag{8}$$

注意这个 L 不是上节所讲的损失函数。对数似然函数为

$$\begin{aligned} \ln L(w) &= \sum_{i=1}^{n} \ln(N(Y_i | w^T X_i, \sigma^2)) \\ &= -\frac{1}{\sigma^2} E_D(w) + n \ln \sigma + const \end{aligned} \tag{9}$$

其中，我们令

$$E_D(w) = \frac{1}{2} \sum_{i=1}^{n} (Y_i - w^T X_i)^2 \tag{10}$$

可以看到，对数似然函数对参数 w 的依赖只有 $E_D(w)$ 一项，由于该项前的系数是负的，因此要使的对数似然函数最大，就应当使得 $E_D(w)$ 最小。因此我们对该项求对 w 的偏导

$$\nabla_w E_D(w) = -\sum_{i=1}^{n} (Y_i - w^T X_i) X_i \tag{11}$$

令上述偏导等于 0，就求得使似然函数最大的 w 的解析解为

$$w_{ML} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y} \tag{12}$$

其中，$\mathbf{X} = (X_1 \ldots X_n)^T$，$\mathbf{Y} = (Y_1, \ldots, Y_n)^T$。

有了这个估计量，我们就可以对 x 做预测了。由正态分布的期望容易得到 $h(x) = g^*(x) = E(Y|X = x) = w_{ML}^T x$，这个就是我们用最大似然估计得到的最好的预测函数了。

我们看到，在假定噪声服从正态分布时，最大似然估计等价于最小化 $E_D(w)$。定义错误函数 (error function)$err(h(X_i), Y_i)$，代表我们的预测函数 h 对第 i 个样本所产生的预测错误。平方错误函数为

$$err_{square}(h(X_i), Y_i) = (Y_i - h(X_i))^2 \tag{13}$$

需要注意这个错误函数和上节介绍的损失函数是 2 个不同的东西！因此，最大似然估计等价于最小化样本的平方错误和。而直接由这个思想出发，即通过最小化样本的平方错误和得到的回归函数的估计叫做最小二乘估计。因此，最小二乘估计本质上是在假设噪声为正态分布下的最大似然估计。

2.2 最大后验估计与正则化

最大后验估计通常也叫做“半贝叶斯”方法。之所以叫半，是因为它不是完全贝叶斯的，它引入了参数的先验分布，但仍然只求了一个参数的点估计。

我们假设参数 w 的先验分布也服从正态分布，为 $N(w|\alpha,\Sigma)$，注意这是个多元正态分布。我们求 w 在已知样本情况下的后验分布

$$p(w|\mathbf{Y},\mathbf{X}) \propto p(\mathbf{Y}|w,\mathbf{X})p(w) \tag{14}$$

后验分布正比于似然函数和先验分布的乘积，分母是个和 w 无关的边缘分布，因此我们只看分子的形式。注意对数似然函数关于 w 是一个 w 的 2 次多项式，而先验分布本身也是正态分布，它的对数关于 w 也是 w 的 2 次多项式，所以后验分布的对数关于 w 也是一个 2 次多项式，这说明后验分布也是一个正态分布！因此我们设

$$p(w|\mathbf{Y},\mathbf{X}) = N(w|m,S) \tag{15}$$

因此我们只要看关于 w 的项，就可以找出后验分布的参数 m 和 S，而其他部分只是一些归一化的常数 (对于 w 来说是常数)。

首先看似然函数 $p(\mathbf{Y}|w,\mathbf{X})$，它就是我们上边的 $L(w)$，我们把它的对数形式写成矩阵形式

$$\begin{aligned} \ln L(w) &= -\tfrac{1}{\sigma^2}E_D(w) + other \\ &= -\tfrac{1}{2}w^T(\tfrac{\mathbf{X}^T\mathbf{X}}{\sigma^2})w + w^T(\tfrac{\mathbf{X}^T\mathbf{Y}}{\sigma^2}) + other \end{aligned} \tag{16}$$

其他的项之所以写成 $other$ 是因为这些项与 w 无关。下面是先验分布 $p(w)$，

$$\ln p(w) = -\frac{1}{2}w^T(\Sigma^{-1})w + w^T(\Sigma^{-1}\alpha) + other \tag{17}$$

因此，我们可以写出 w 的后验分布的对数关于 w 的项的形式：

$$\ln p(w|\mathbf{Y},\mathbf{X}) = -\frac{1}{2}w^T(\frac{\mathbf{X}^T\mathbf{X}}{\sigma^2} + \Sigma^{-1})w + w^T(\frac{\mathbf{X}^T\mathbf{Y}}{\sigma^2} + \Sigma^{-1}\alpha) + other \tag{18}$$

这是个关于 w 的 2 次多项式，也证实了后验分布是正态分布的观点。这样，我们就可以从这个式子里直接写出后验分布的参数：

$$\begin{aligned} S &= (\tfrac{\mathbf{X}^T\mathbf{X}}{\sigma^2} + \Sigma^{-1})^{-1} \\ m &= S(\tfrac{\mathbf{X}^T\mathbf{Y}}{\sigma^2} + \Sigma^{-1}\alpha) \end{aligned} \tag{19}$$

最大后验估计是选后验分布的最大值点作为参数的估计，而后验分布是正态分布，因此最大后验估计就是后验分布的期望

$$w_{MAP} = m = (\frac{\mathbf{X}^T\mathbf{X}}{\sigma^2} + \Sigma^{-1})^{-1}(\frac{\mathbf{X}^T\mathbf{Y}}{\sigma^2} + \Sigma^{-1}\alpha) \tag{20}$$

利用这个估计，我们求得的预测函数为 $h(x) = w_{MAP}^T x$。

考虑先验分布的一种特殊情况，我们假设它的期望是 0 向量，即 $\alpha = \mathbf{0}$，同时协方差矩阵是对角矩阵，即 w 的各分量是独立的。这时令

$$\Sigma = \beta \mathbf{I} \tag{21}$$

这时，后验分布为

$$\ln p(w|\mathbf{Y}, \mathbf{X}) = -\frac{1}{\sigma^2} E_D(w) - \frac{1}{2\beta} w^T w + other \tag{22}$$

这样，最大后验估计等价于求最小化样本平方错误与一个正则化项的和。

$$w_{MAP} = \arg\max_w \ln p(w|\mathbf{Y}, \mathbf{X}) = \arg\min_w E_D(w) + \lambda w^T w \tag{23}$$

其中，$\lambda = \frac{\sigma^2}{2\beta}$，叫做正则化系数。

在非概率的视角下，机器学习通常可以看作 3 步，首先找一个样本错误函数，其次找一个优化目标函数，如最小化样本的错误和，最后是找到一个有效的求解该优化的算法。而机器学习理论确保了，在足够的条件下，如样本量够大，样本的错误情况会和普遍的错误情况差不多，这就是确保了模型有泛化能力。而在第 2 步中，如果只考虑让模型拟合样本，显然越复杂的模型越能更好的拟合样本，这样就会导致过拟合，这是由于此时模型的复杂度过高。所以优化目标还需要考虑模型的复杂度，这就是正则化项。

而以概率的视角看，正则化项实际上等价于对参数加上一个先验分布。直觉上，参数的先验分布代表我事先在不知道样本情况的时候，认为参数大致是个怎么样的分布，如最可能是什么值。等我看到样本之后，我在根据样本去调整参数。而添加了正则化项的目标函数，在优化该目标函数时，参数也是带有“偏向性”，就是偏向使得该正则化项值小的方向。如上述的 $\lambda w^T w$，参数偏向于 0。

2.3 贝叶斯线性回归

上小节所讲的是“半贝叶斯”的，而在全贝叶斯的观点下，在知道参数的后验分布后，对新的 x 的预测 y 的分布为

$$p(y|x, \mathbf{Y}, \mathbf{X}) = \int p(y|x, w, \mathbf{Y}, \mathbf{X}) p(w|\mathbf{Y}, \mathbf{X}) \, \mathrm{d}x \tag{24}$$

下面为了简便，把条件项 x，$\mathbf{Y}$，$\mathbf{X}$ 都省略。首先看联合分布 $p(y, w)$。

$$p(y, w) = p(y|w)p(w) = N(y|x^T w, \sigma^2) N(w|m, S) \tag{25}$$

我们看该联合概率的对数

$$\begin{aligned}\ln p(y,w) &= -\tfrac{1}{2\sigma^2}(y-x^Tw)^T(y-x^Tw)-\tfrac{1}{2}(w-m)^TS^{-1}(w-m)+other \\ &= -\tfrac{1}{2}\begin{pmatrix}y\\w\end{pmatrix}^T\begin{bmatrix}\frac{1}{\sigma^2} & \frac{x^T}{\sigma^2}\\ \frac{x}{\sigma^2} & \frac{xx^T}{\sigma^2}+S^{-1}\end{bmatrix}\begin{pmatrix}y\\w\end{pmatrix} \\ &\quad +\begin{pmatrix}y\\w\end{pmatrix}^T\begin{pmatrix}0\\S^{-1}m\end{pmatrix} \\ &\quad +other\end{aligned} \tag{26}$$

其中 $other$ 是不依赖 y 和 w 的项。这是一个关于 (y,w) 的 2 次式，因此该联合概率也是正态分布。从上式中我们可以直接写出该联合分布的参数，即期望 m_{union} 与协方差矩阵 S_{union} 为：

$$\begin{aligned}S_{union} &= \begin{bmatrix}\frac{1}{\sigma^2} & -\frac{x^T}{\sigma^2}\\ -\frac{x}{\sigma^2} & \frac{xx^T}{\sigma^2}+S^{-1}\end{bmatrix}^{-1}=\begin{bmatrix}\sigma^2+x^TSx & x^TS\\ Sx & S\end{bmatrix} \\ m_{union} &= S_{union}\begin{pmatrix}0\\S^{-1}m\end{pmatrix}=\begin{pmatrix}x^Tm\\m\end{pmatrix}\end{aligned} \tag{27}$$

有了联合分布，就可以求出 y 的边缘分布，利用附录中的结果得到，y 也服从正态分布，其分布为

$$p(y|x,\mathbf{Y},\mathbf{X})=N(y|m^Tx,\sigma^2+x^TSx) \tag{28}$$

因此，预测函数为

$$h(x)=g^*(x)=E(Y|X=x,\mathbf{Y},\mathbf{X})=m^Tx \tag{29}$$

预测函数的形式和最大后验估计得到的一样，但是意义不同。从该分布的方差看，σ^2 是噪声的方差，这是数据内在的。而 x^TSx 代表了对参数的不确定性，参数的不确定性越小，该项越小，整个方差也越小，分布的不确定性也就越小。当样本数越多，这一项会趋近于 0。

3 附录

3.1 正态分布的边缘分布

已知联合分布 $(x^T, y^T)^T$ 服从多元正态分布 $N(m, \Sigma)$，令联合分布的参数期望和协方差矩阵为

$$\begin{aligned} m &= \begin{pmatrix} m_x \\ m_y \end{pmatrix} \\ \Sigma &= \begin{bmatrix} \Sigma_1 & \Sigma_2 \\ \Sigma_2^T & \Sigma_3 \end{bmatrix} \end{aligned} \tag{30}$$

为了方便，我们定义协方差的逆矩阵为

$$\Sigma^{-1} = \begin{bmatrix} \Lambda_1 & \Lambda_2 \\ \Lambda_2^T & \Lambda_3 \end{bmatrix} \tag{31}$$

可以求出两矩阵有如下的对应关系

$$\begin{aligned} \Sigma_1 &= (\Lambda_1 - \Lambda_2 \Lambda_3^{-1} \Lambda_2^T)^- 1 \\ \Sigma_2 &= -\Sigma_1 \Lambda_2 \Lambda_3^{-1} \\ \Sigma_3 &= \Lambda_3^{-1} + \Lambda_3^{-1} \Lambda_2^T \Sigma_1 \Lambda_2 \Lambda_3^{-1} \end{aligned} \tag{32}$$

该联合分布的边缘分布 $p(x)$ 为

$$p(x) = \int p(x, y) \, \mathrm{d}y \tag{33}$$

由于是对 y 积分，我们可以把 $p(x, y)$ 中的有关 x 的项看成是常数项，这样就可以移到积分外，先看看 $p(x, y)$ 的对数形式，我们把有关 x 与 y 的项单独提出来。

$$\begin{aligned} \ln p(x, y) &= -\tfrac{1}{2} \begin{pmatrix} x - m_x \\ y - m_y \end{pmatrix}^T \begin{bmatrix} \Lambda_1 & \Lambda_2 \\ \Lambda_2^T & \Lambda_3 \end{bmatrix} \begin{pmatrix} x - m_x \\ y - m_y \end{pmatrix} + other \\ &= -\tfrac{1}{2} x^T \Lambda_1 x + x^T (\Lambda_1 m_x + \Lambda_2 m_y) \\ &\quad + -\tfrac{1}{2} \left(y - \Lambda_3^{-1} (\Lambda_2^T m_x - \Lambda_2^T x + \Lambda_3 m_y) \right)^T \Lambda_3 \left(y - \Lambda_3^{-1} (\Lambda_2^T m_x - \Lambda_2^T x + \Lambda_3 m_y) \right) \\ &\quad + \tfrac{1}{2} \left(\Lambda_3^{-1} (\Lambda_2^T m_x - \Lambda_2^T x + \Lambda_3 m_y) \right)^T \Lambda_3 \left(\Lambda_3^{-1} (\Lambda_2^T m_x - \Lambda_2^T x + \Lambda_3 m_y) \right) \\ &\quad + other \end{aligned} \tag{34}$$

可以看到，中间关于 y 的项本身是正态分布的形式，因此对 y 积分后的值只是正态分布的 normalize 的项，而从正态分布的形式知道该项只和其协方差矩阵

有关，即 Λ_3^-1，这是和 x 无关的项，因此，

$$\begin{aligned}\ln p(x) &= \ln \int p(x,y)\,\mathrm{d}y \\ &= -\tfrac{1}{2}x^T\Lambda_1 x + x^T(\Lambda_1 m_x + \Lambda_2 m_y) \\ &\quad +\tfrac{1}{2}\left(\Lambda_3^{-1}(\Lambda_2^T m_x - \Lambda_2^T x + \Lambda_3 m_y)\right)^T \Lambda_3 \left(\Lambda_3^{-1}(\Lambda_2^T m_x - \Lambda_2^T x + \Lambda_3 m_y)\right) \\ &\quad +other\end{aligned} \tag{35}$$

这是一个 x 的 2 次式，因此 x 的边缘分布也是一个正态分布！设分布的期望和协方差矩阵分别是 u_x 和 v_x。对上式继续提出有关 x 的项，然后用配比的方法就可以求出这些值。

$$\begin{aligned}\ln p(x) &= -\tfrac{1}{2}x^T(\Lambda_1 - \Lambda_2\Lambda_3^{-1}\Lambda_2^T)x \\ &\quad +x^T(\Lambda_1 m_x - \Lambda_2\Lambda_3^{-1}\Lambda_2^T m_x) \\ &\quad +other\end{aligned} \tag{36}$$

所以

$$\begin{aligned}v_x &= (\Lambda_1 - \Lambda_2\Lambda_3^{-1}\Lambda_2^T)^{-1} = \Sigma_1 \\ u_x &= v_x(\Lambda_1 - \Lambda_2\Lambda_3^{-1}\Lambda_2^T)m_x = m_x\end{aligned} \tag{37}$$