机器学习与深度学习面试系列四（线性回归）

什么是线性回归?

- 线性：两个变量之间的关系**是**一次函数关系的——图象**是直线**，叫做线性。
- 非线性：两个变量之间的关系**不是**一次函数关系的——图象**不是直线**，叫做非线性。
- 回归：人们在测量事物的时候因为客观条件所限，求得的都是测量值，而不是事物真实的值，为了能够得到真实值，无限次的进行测量，最后通过这些测量数据计算**回归到真实值**，这就是回归的由来。

线性回归的一般表达式?

线性回归模型的最简单的形式也是输入变量的线性函数：

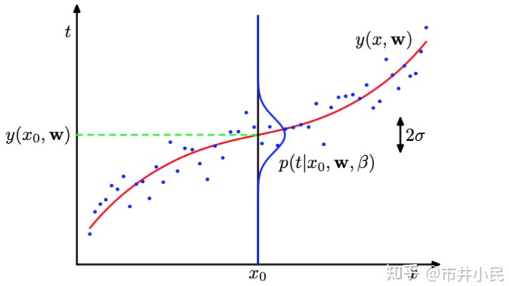
$$y=\sum_{i=1}^{M}w_ix_i+b$$

但是，通过将一组输入变量的非线性函数进行线性组合，我们可以获得一类更加有用的函数，被称为基函数(basis function)。这样的模型是参数的线性函数：

$$y=\sum_{i=1}^{M}w_i\phi_i(x)+b$$

常见的基函数包括：多项式基函数、“高斯”径向基函数、sigmoid基函数、傅里叶基函数等。

均方差损失函数与最大似然



均方差损失函数最大似然估计推导

用不确定性的观点看目标变量。对于任意一点x，都将其目标变量 t 看作是以 $t=wx+b$ 为均值的，β^{-1} *为方差的高斯分布。即：* $p(t_n|x_n,\bar{w},\beta)=N(t_n|y(x_n,\bar{w}),\beta^{-1})$

那么对于整个训练集 $\{x_n, t_n\}$，通过最大似然估计来确定w的值。似然函数为：

$$p(\bar{t}|\bar{x},\bar{w},\beta)=\prod_{n=1}^{N}p(t_n|x_n,\bar{w},\beta)=\prod_{n=1}^{N}N(t_n|y(x_n,\bar{w}),\beta^{-1})$$

要求使上述函数取到最大值的w，往往我们求使最大对数似然函数的w值（因为似然函数存在连乘符号，不好处理，取对数可以将连乘转化为连加）。对数似然函数为：

$$lnp(\bar{t}|\bar{x},\bar{w},\beta)=-\frac{\beta}{2}\sum_{n=1}^{N}\{y(x_n,\bar{w})-t_n\}^2+\frac{N}{2}ln\beta-\frac{N}{2}ln(2\pi)$$

观察式子等号右边，只有第一项含有参数w，要求使上式取到最大值的w，就是求使上式第一项取到最大值的w，也就是使 $\sum_{n=1}^{N}\{y(x_n,\bar{w})-t_n\}^2$ 取到最小值的w。这与均方差损失函数形式完全一样。

过拟合怎么解决?

L1正则化（Lasso回归）。 $L=\frac{1}{2}||Xw-y||^2+\lambda||w||_1$

L2正则化（岭回归）。 $L=\frac{1}{2}||Xw-y||^2+\lambda||w||_2^2$

ElasticNet正则化（ElasticNet回归）。 $L=\frac{1}{2}||Xw-y||^2+\lambda_1||w||_2^2+\lambda_2||w||_1$

正则化的使用场景?

只要数据线性相关，用线性回归拟合的不是很好，**需要正则化**，可以考虑使用岭回归(L2), 如何输入特征的维度很高,而且是稀疏线性关系的话， 岭回归就不太合适,考虑使用Lasso回归。L1正则化(Lasso回归)可以使得一些特征的系数变小,甚至还使一些绝对值较小的系数直接变为0。

在我们发现用Lasso回归太过(太多特征被稀疏为0),而岭回归也正则化的不够(回归系数衰减太慢)的时候，可以考虑使用ElasticNet回归来综合，得到比较好的结果。

正则化损失函数与最大后验估计的关系?

前文说到，不含正则项的损失函数，实际上是由最大似然法导出的，其实含正则项的损失函数是由最大后验估计导出的。

由贝叶斯定理：

$p(\bar{w}|\bar{x},\bar{t}) = \frac{p(\bar{t}|\bar{x},\bar{w})p(\bar{w})}{\sum p(\bar{t}|\bar{x},\bar{w})p(\bar{w})}$ 分母可以看作归一化参数，则：

$p(\bar{w}|\bar{x},\bar{t})$ 正比于 $p(\bar{t}|\bar{x},\bar{w})p(\bar{w})$ 。使用上面求的： $p(\bar{t}|\bar{x},\bar{w}) = \prod_{n=1}^{N} N(t_n|y(x_n,\bar{w}),\beta^{-1})$ ，对于 $p(\bar{w})$ 我们给予一个先验假设 $p(\bar{w}) = N(\bar{w}|0,\alpha I)$,即w满足均值为0的一个高斯分布。经过取对数，这样可以求得： $p(\bar{w}|\bar{x},\bar{t})$ 反比于 $\sum_{n=1}^{N}\{y(x_n,\bar{w})-t_n\}^2 + \frac{\alpha}{2}w^Tw$ ，问题转为求使这个式子最小化的w，该形式与添加了L2正则化项的损失函数形式一样，也就是说L2正则化损失函数相当于给原损失函数引入了一个关于w的高斯先验。

同样的，L1正则化损失函数相当于给原损失函数引入了一个关于w的拉普拉斯先验，不再赘述。

另外，最大似然法实际上也可以看作是一个关于w的均匀分布先验。

为什么L1正则化可以得到稀疏解?

可以从两个方面来理解：

从约束优化的角度：假定仅有两个权重 w_1 和 w_2 ，我们将其作为两个坐标轴，然后在图中绘制等值线，再分别绘制出 L1 范数与 L2 范数的等值线。最优解要在平方误差项 与正则化项之间折中，即出现在图中平方误差项等值线与正则化项等值线相交处.由下图可看出，采用 L1 范数时平方误差项等值线与正则化项等值线的 交点常出现在坐标轴上，即 w_1 或 w_2 为0， 而在采用L2范数时，两者的交点常出现在某个象限中，即 w_1 或 w_2 非0。换言之采用L1范数比L2范数更易于得到稀疏解.

从先验概率的角度：L1和L2正则化可以看作是满足关于w的拉普拉斯先验和高斯先验。可以看一下拉普拉斯概率分布函数的图像，明显拉普拉斯分布相较于正态分布，他的中心点更尖，也就是说他落在中心点（也即0值）的概率更大，因此它更容易得到稀疏解。

拉普拉斯概率分布

为什么正则化可以减轻过拟合?

从先验概率角度：无正则化的损失函数可以认为是满足关于w的均匀分布先验，而L1和L2正则化可以看作是满足关于w的拉普拉斯先验和高斯先验，所以L1和L2实际上是对w的取值范围做了约束，

以此减少模型的方差（偏置-方差困境），从而减轻了过拟合（过拟合可以看作是模型方差过大）。

从病态矩阵的角度：对于原始损失函数，其解析解为： $w=(X^TX)^{-1}X^Ty$ 。事实上这里存在两个问题，一个是 X^TX 可能不可逆，此时说明 X^TX 中有特征值为0，发生场景为X的维度比样本数还大。另一个问题是即使 X^TX 可逆，但是这个矩阵是病态的，也就是说如果y存在很小的波动， 被 $(X^TX)^{-1}$ 乘了以后，结果w都会发生很大的变化。那么我们计算的这个w就非常的不稳定，并不是一个好的模型，发生场景为X的维度比样本数差不多大小。

判断一个矩阵是不是病态矩阵，可以通过计算矩阵的条件数。条件数等于矩阵的最大奇异值和最小奇异值之比。如果矩阵 X^TX 存在很小的奇异值，那么它的逆就存在很大的奇异值，这样对y中的微小变化会放大很多。所以我们的目标就是干掉 X^TX 中极小的奇异值。

加了L2正则项之后，我们的解析解变为： $w=(X^TX+\lambda I)^{-1}X^Ty$。也就是给 X^TX 中的所有奇异值加上一个 λ，可以确保奇异值不会太小，而导致再求逆后，奇异值变的极大。这样有效的解决了病态矩阵的问题。过拟合的实质可以看作由于病态矩阵的存在，如果y有一点波动，整个模型需要大幅度调整。解决了病态矩阵问题，就解决了过拟合。