**从Linear Regression谈起**

这个专题主要起到一个系列的开篇的作用，基本上这个系列是从统计学角度看机器学习方法，我们先从最基本的Linear Regression开始谈起。  
Linear Regression的目标是从一组变量X和其对应响应Y的样本来推测一个X和Y之间的线性关系，假设X是一个N\*d的矩阵表示N个sample input，而Y是一个N\*1的矩阵表示N个sample output，那么我们的目标就是计算一个d\*1的向量beta，使得x\*beta近似等于y：

从Linear Regression谈起X

对于这个问题，我们手上有的数据是样本数据X和Y，那么最基本的Least Square Regression的目标是使得Squared loss尽量小，即：  
从Linear Regression谈起  
求解上式，我们可以解出beta的值：  
从Linear Regression谈起  
我们现在做一个假设，即y是从x的线性组合中抽取的，但是会有一个加性的噪声：  
从Linear Regression谈起  
其中epsilon是一个独立的随机变量且其期望为0，代入上式，可以看出：  
从Linear Regression谈起  
因此我们可以看出beta和期望值就等于最佳的预测，因此Least Square是一个unbiased学习算法，著名的高斯-马尔科夫定理证明了，在所有unbiased线性回归方法中，Least Square的Variance是最小的，我们熟知的Bias-Variance Tradeoff说明了期望的Loss等于Bias的平方加上算法的Variance：  
从Linear Regression谈起  
其中y是算法对新样本的估计值，y\_0是新样本的实际值，可以看出如果我们希望Squared Loss尽量小，我们需要降低算法的Bias，或者降低算法对数据的敏感度（即降低y的Variance）。高斯-马尔科夫定理已经说明了在Unbiased情形下我们没有可能做到更好，这也就为我们指明了道路：去利用较小的Bias换取Variance的降低来达到提高算法性能的目的。  
Subset Selection是一个可能的降低Variance的方式，事实上由于可能样本数据中的维数比较高，学习方法会比较倾向于过拟合数据（Overfitting），Subset Selection则倾向于选择较少的和输出相关度最高的维度，并利用这些维度进行Regression。另外的一个好处则是能更好的解释数据，通过寻找最相关的几个特征，我们能更好地直观感受输出结果受那几个输入影响最大。  
Best-Subset Selection试图寻找全部特征的一个子集，并利用这个子集进行Least Square Regression，通过限制子集的大小，我们可以控制算法的Bias-Variance Tradeoff。子集选择的过程需要一点技巧，因为子集的个数是指数相关于子集大小的，我们利用一点迭代技巧来处理这个问题。我们可以选择从一个空的子集中逐步加入和目标函数相关系数最大的特征（或者说加入这个特征可以使得SquaredLoss减少最多），同样也可以选择从满的特征集合中逐步drop相关性最小的特征，这里用到了统计学中的Z-score，我们不去细表。  
与子集选择非常相似的一个方法则是Stagewise Regression，这里我们每次选择一个特征，然后把它设为最优值（即使得Squared Loss尽量小），每次我们选择的特征应该是和目标相关度最大的特征或者调整后能使Squared Loss减小最多的特征。Stagewise Regression相比于Least Square等方法速度要慢很多，因为它使用了一种迭代性的求解方式而不是解析性的求解，这使得在历史中这个方法并没有收到很大重视，但是最近的研究表明在高维空间中这种'Slow fitting'可以很好地降低Variance。  
Shrinkage Method则是另一类降低Variance的Method，通常我们试图对Least Square追加一项关于系数的penalty，例如著名的Ridge Regression：  
从Linear Regression谈起  
与此等价的另一种形式是：  
从Linear Regression谈起  
我们考虑第一种形式，在这种形式下我们可以解出ridge regression的解：  
从Linear Regression谈起  
可以看出通过追加一个正的项我们可以使得矩阵的逆在数值计算中更加稳定，为了更加深刻得理解Ridge Regression，我们考虑X的SVD分解，基本上SVD分解有如下形式：  
从Linear Regression谈起  
其中U和V都是列向量正交阵，代入Least Square的解可得：  
从Linear Regression谈起  
同理我们可以将SVD分解代入Ridge Regression得到：  
从Linear Regression谈起  
因此Ridge Regression可以视为将y投影在X所张成的平面内，然后按照某一个系数收缩投影的各方向，可以证明X的SVD分解的singular values的平方刚好就是X的PCA的各个分量系数，因此Ridge Regression可以视为将X的较弱分量进行较强收缩，而将X的较强分量维持不变。这里隐含的一个假设是响应Y在X的主分量上变化比较大，而在X的次要分量上变化比较小，在很多情形下这一假设还是比较可取的。  
和Ridge Regression很接近的另一个方法被称为lasso，lasso方法有如下形式：  
从Linear Regression谈起  
通过利用1范数，lasso可以达到求解sparse solution的目的，在很多情形下lasso的解中很多个分量的系数都是0，这是由于1范数在坐标轴上不可导的性质决定的，而Ridge Regression则没有这个性质。部分由于这个原因，有人提出了elastic-net penalty：  
从Linear Regression谈起  
他可以被视为某种介于lasso和ridge regression之间的方法。  
与lasso相关的一个算法被称为Least Angle Regression，LAR是一种计算Least Square的stepwise的迭代方法，它首先寻找一个特征和目标响应相关性最大，和Stagewise Regression不同的是，它并不直接将该特征的系数设成使Squared Loss最小的系数，与此相对的，它逐渐的改变该特征的系数，直到存在一个别的特征的相关性和该特征相等。LAR维护一个Active List，然后同时改变Active List内的所有特征并保持它们和目标响应的相关性相等，并不停地将新的特征加入Active List。如果我们在LAR的过程中将所有系数成为0的特征移出Active List，可以证明这种LAR的修正版本的解和lasso的解的path完全一致，因此LAR也可以用来求解lasso。