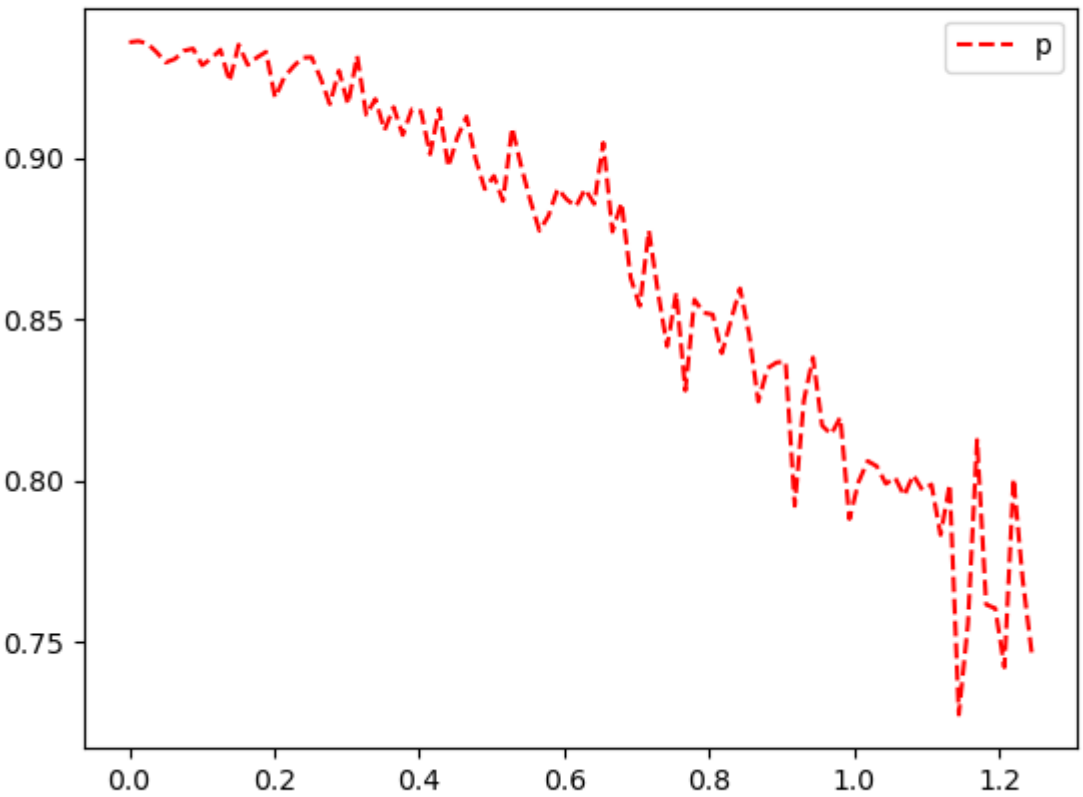
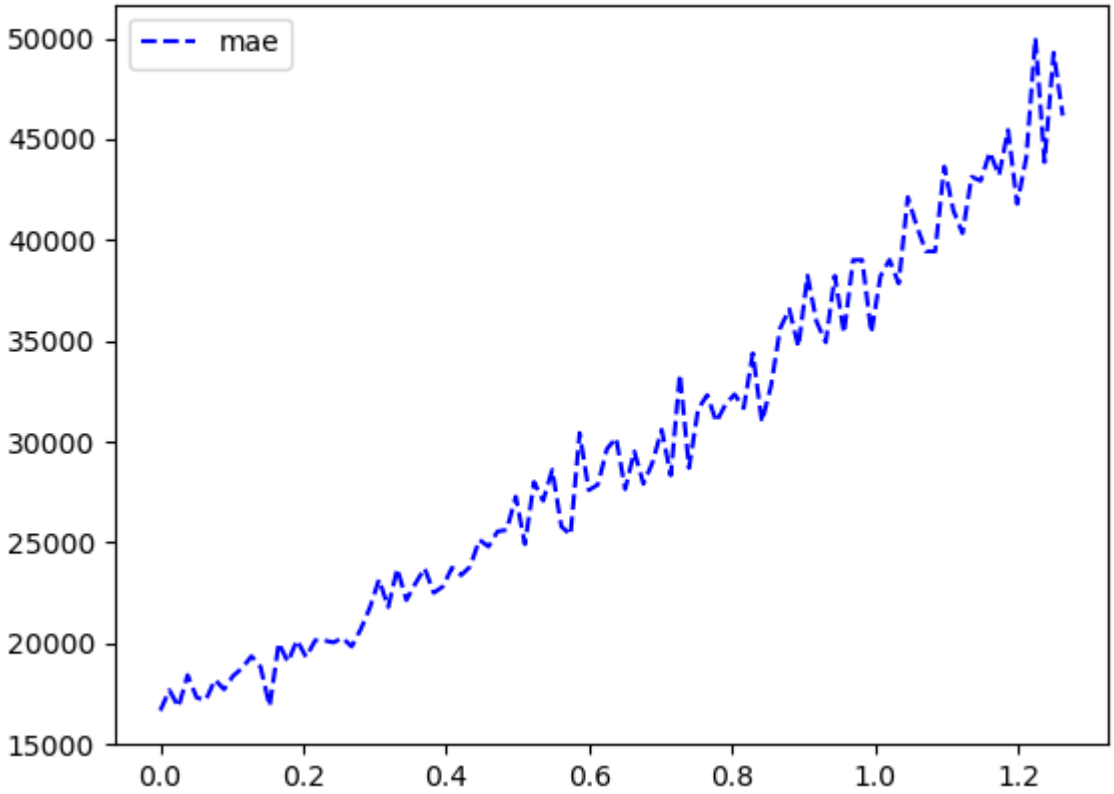
Progress Update 2020.8.2

噪声影响分析

我一直在想，样本噪声应该广泛存在于任何数据集中，然而并非在每个数据集的任务中都被重视考虑，在这里我在经典回归数据集`House Price`上，用测试label噪声对测试回归模型表现的影响，采取p值和mean absolute error两个指标。

label噪声服从正态分布，均值为0，横坐标为噪声方差与样本label方差的比值





朴素的Sample Weight修正

噪声可以被分为三类，分布由data value决定，由label决定，独立于样本本身。

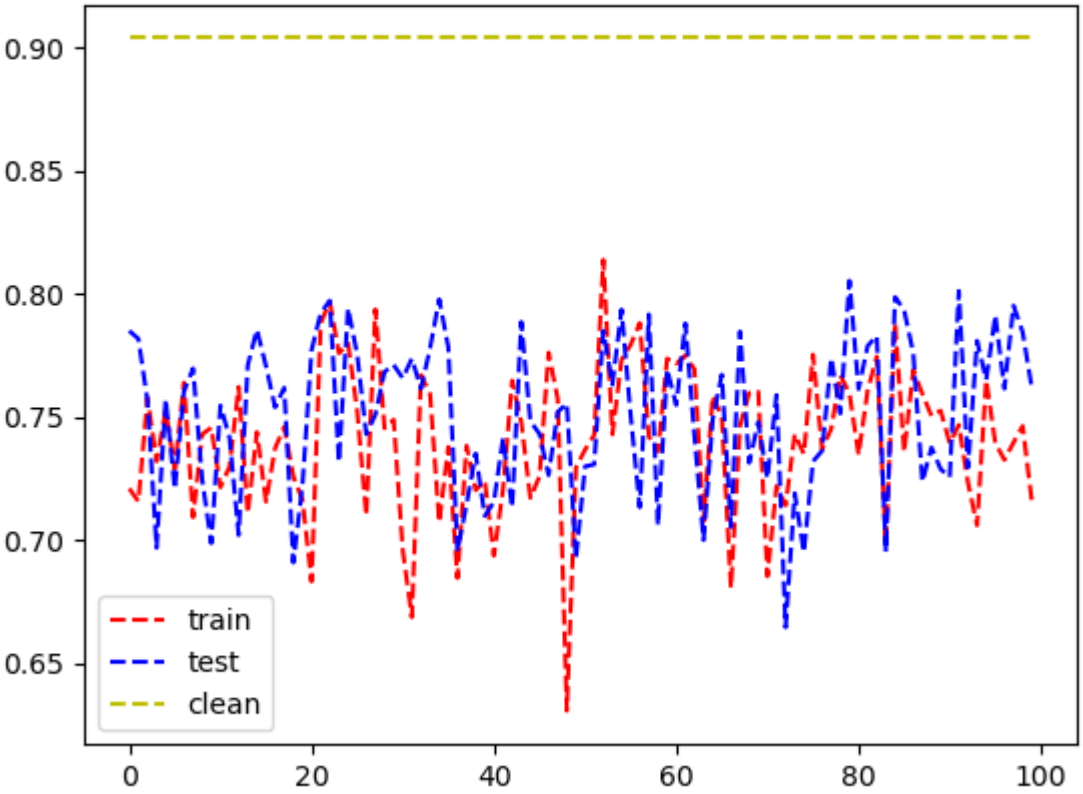
简单的literature review后，发现Learning with Noisy Labels，Classification with Noisy Labels by Importance Reweighting中研究的都是label conditional的噪声。实验误差测量误差应该属于第三者，于前两者着本质的区别。

但是Reweight的方法我认为由借鉴意义，于是设计了一个算法。

- 已知小部分noise free样本(X_{free}, y_{free})，训练回归模型F_{free}
- 对于噪声样本$D_{noise} = (X_{noise}, y_{noise})$，$\forall (X_i, y_i) \in D_{noise}, w_i = \frac{1}{(y_{noise} - F_{free}(X_i))^2 + \epsilon}$，$w_i$为sample weigth
- 对于噪声样本和sample weight的组合训练最终的回归模型

测试结果，噪声为均值为0，方差为1.2样本label方差的正态分布

横坐标为测试轮数，纵坐标为回归p值，`train`代表未reweight的回归模型测试结果，`test`reweight后回归模型的测试结果，`clean`代表在noise free的样本上训练的回归模型的测试结果.



下一步计划

还是需要一个系统的literature review，目前找到的研究回归模型noise影响和改进算法的paper还比较少。

reweight的效果还不明显，还可以调调参。

另外两种类型的噪声也可以进一步研究，事实上这两种只是在实践上有细微差别，本质都是误差分布与样本分布有关。