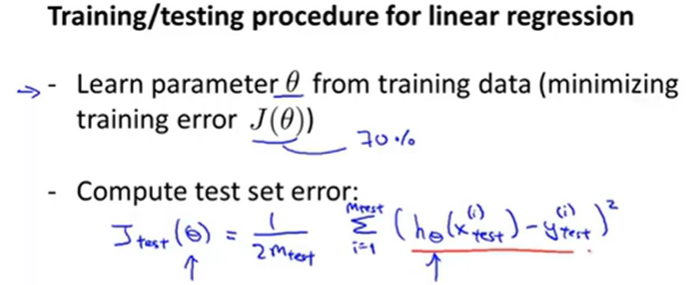
03 机器学习的一些建议

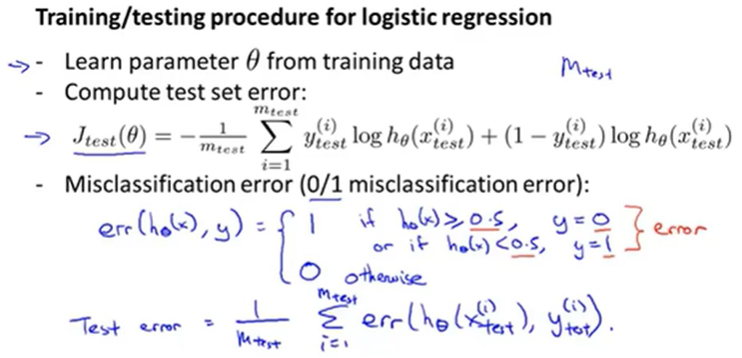
by Bosco Xue 2020.04.22

模型选择与评估，机器学习中会遇到的一些问题及解决办法，以及让在选择哪些改进方法中有法可循。

1. **模型评估与选择**

对于线性回归模型，可以通过代价函数在测试集上的表现来评价我们的假设函数；对于分类问题，除了用代价函数在测试集上的表现评价外还可以用误判率来评价。



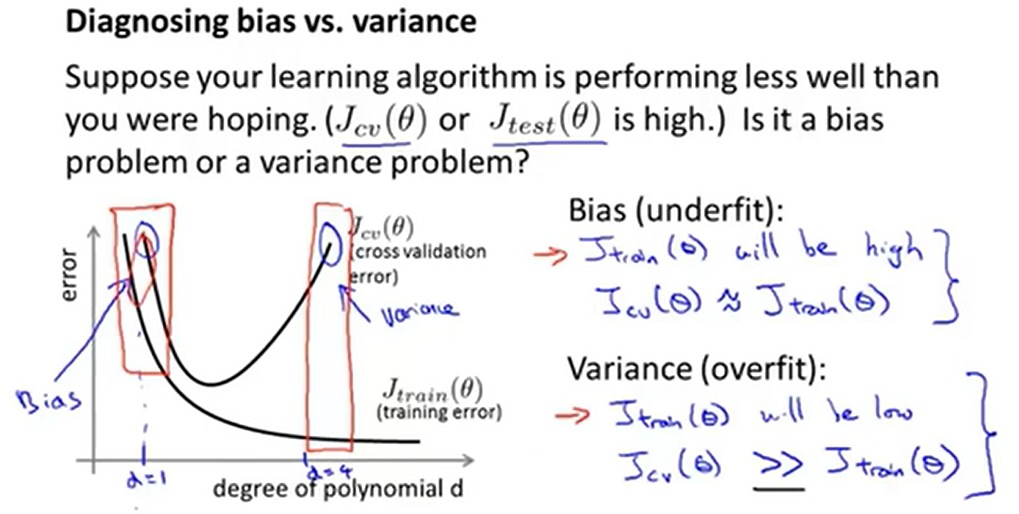


将样本分为训练集、验证集、测试集（通常是3:1:1），通过训练集得到多个模型，通过验证集选择出最好的模型（即选择参数的取值），最后通过测试集看看泛化能力如何。

1. **Bias and Variance**

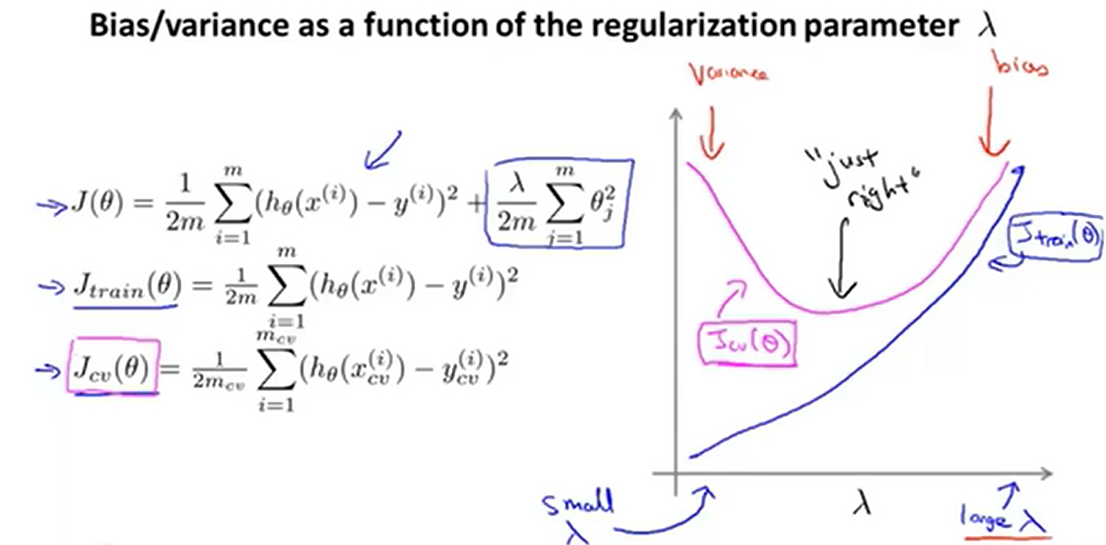
过拟合，模型复杂，则捕捉特征的能力强，容易捕捉到样本中的噪声特征，而增大样本量，模型就“见多识广”，就不会去注意那些噪声特征。

要想改进算法，先判断一下是欠拟合还是过拟合。以模型参数作为横轴，代价函数值或者误差作为纵轴，画出Jtrain和Jcv，进行判断是欠拟合还是过拟合，然后根据欠拟合还是过拟合采取改进办法。



我们知道当模型过拟合时，代价函数里增加正则化项可减小过拟合，但是正则化参数λ过大时，模型还可能修正过渡变成欠拟合，也就是说，λ本身就是一个需要优化的参数，所以λ的选择也需要通过画图的方法确定。

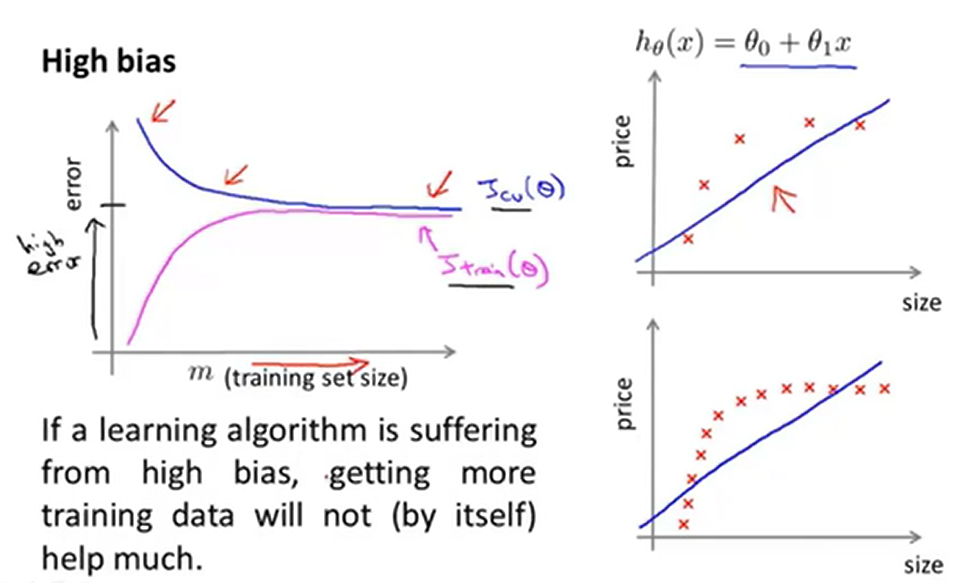
视频里介绍的方法是，先采用带正则化项的代价函数，利用训练集训练出不同λ对应的θ，再把这些θ代入不带正则项的代价函数中，画图判断欠拟合还是过拟合，并根据图确定合适的λ。

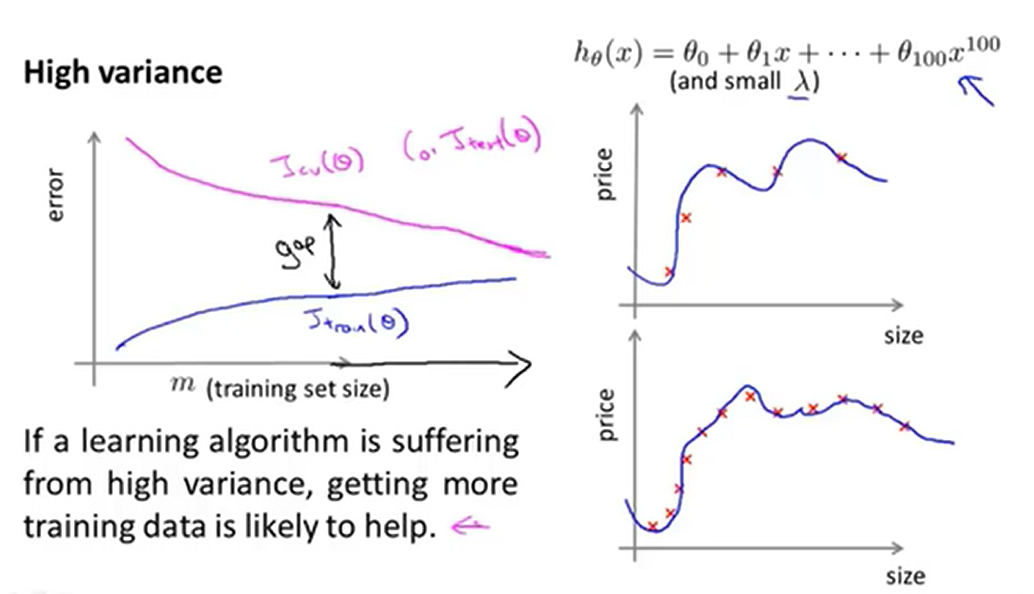


1. **学习曲线Learning Curve**

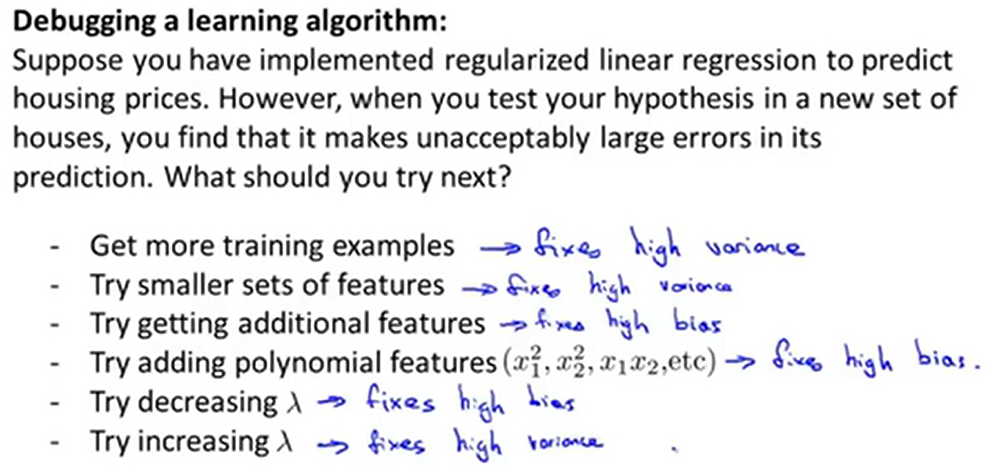
（1）判断欠拟合还是过拟合

样本数量m一般不会去变，肯定是越多越好，但是增加m对欠拟合是没有用的，对减少过拟合有一定效果。因此，画出关于m的学习曲线，有助于分辨模型是欠拟合还是过拟合，也帮助我们判断是否需要去增加更多的样本。





（2）相应的改进措施



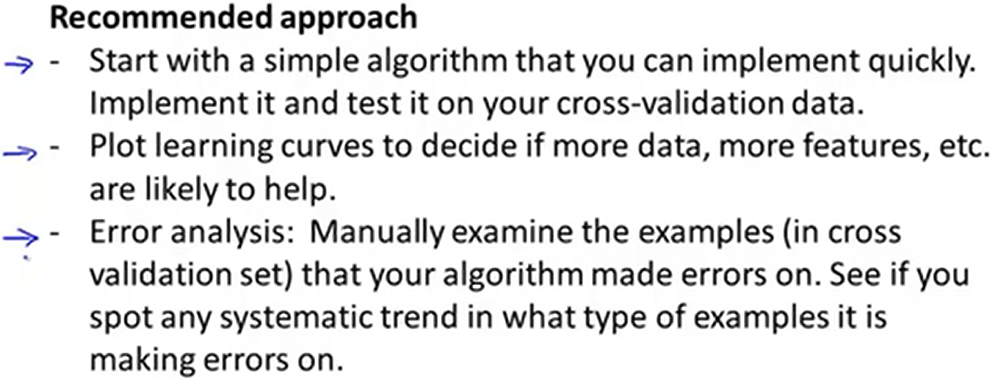
1. **构建机器学习系统**

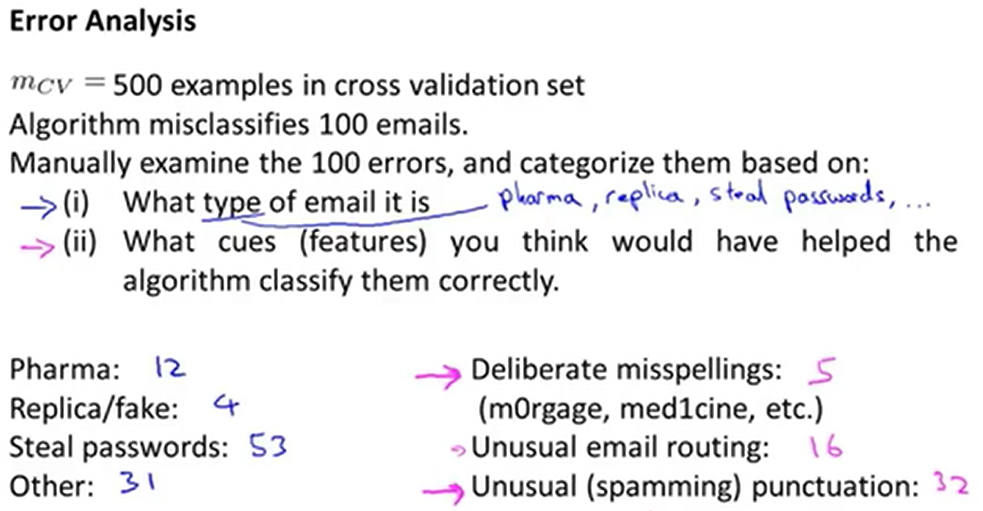
以构件一个垃圾邮件分类系统为例展开。

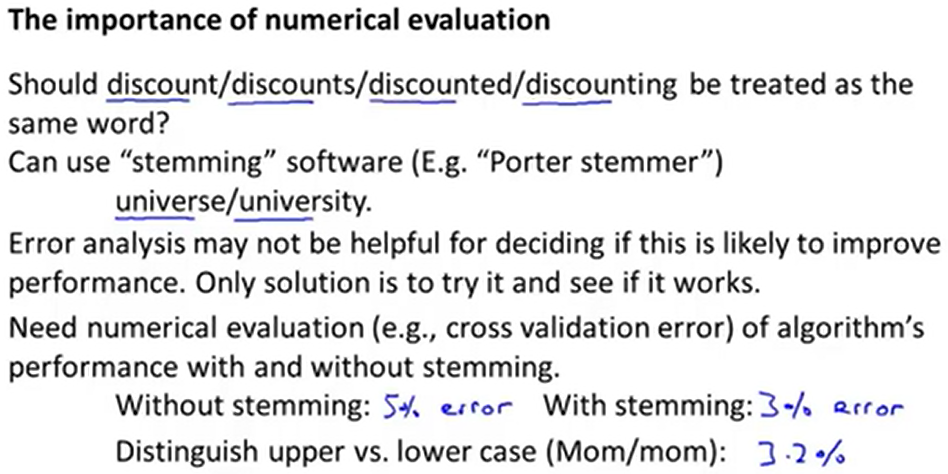
（1）误差分析 error analysis

先实现简单的算法模型，然后画出学习曲线分析一下，接着对错误分类邮件进行统计分析，看看有没有启发。因为，在刚开始构件模型时，很多问题都不出在所选择的算法模型上，所以误差分析就可以先把这些问题给解决或改善。

最后，通过一些数值指标（如验证集上的错误率）验证前面的那些改进想法是否有效。





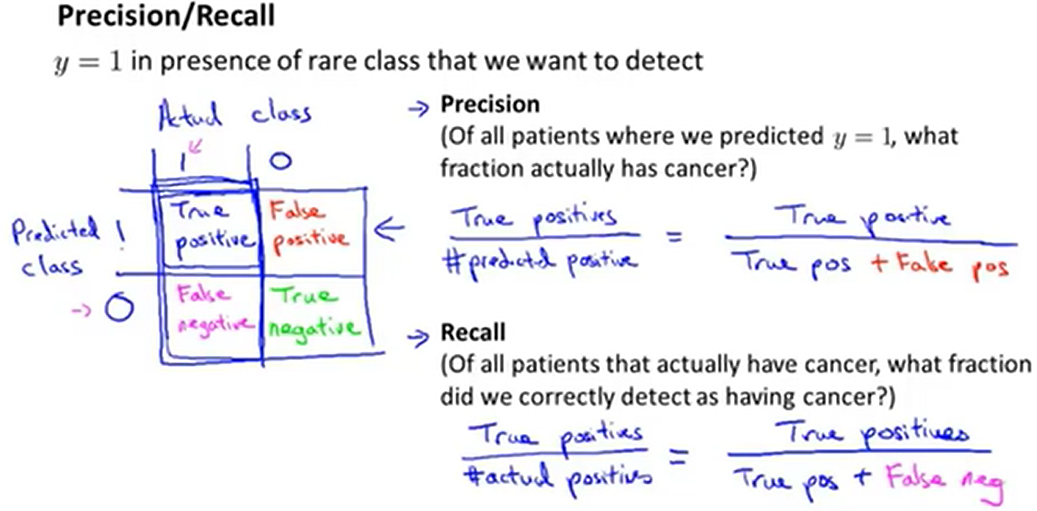


（2）不平衡类别问题 skewed classes

对于不平衡分类问题，比如y=1的样本占了0.5%，那么只要全部预测为y=0就有99.5%的正确率，模型提高空间很小，这个时候很难评价模型的提高是你所采取的一些措施奏效的。针对这种情况，采用准确率（查准率）和召回率（查全率）评价。

查准率和召回率此消彼长，可以用F1 Score（调和平均值，2/F1=1/P+1/R）评价。此外，可以画出P-R曲线，越靠近右上角的模型越好。

通常把稀少的那一类当做y=1的类。



（3）是否需要大量的数据

如果假设的模型太简单，欠拟合，再多的数据也没有用。但是，如果选择的模型带有大量参数，能拟合很复杂的函数，那么他的训练集误差就会很小，数据越大，训练误差与测试误差之间的差距也越来越小。所以，只要模型足够复杂，数据越多越好。