## 机器学习导论

Machine Learning => Looking for a Function from Data

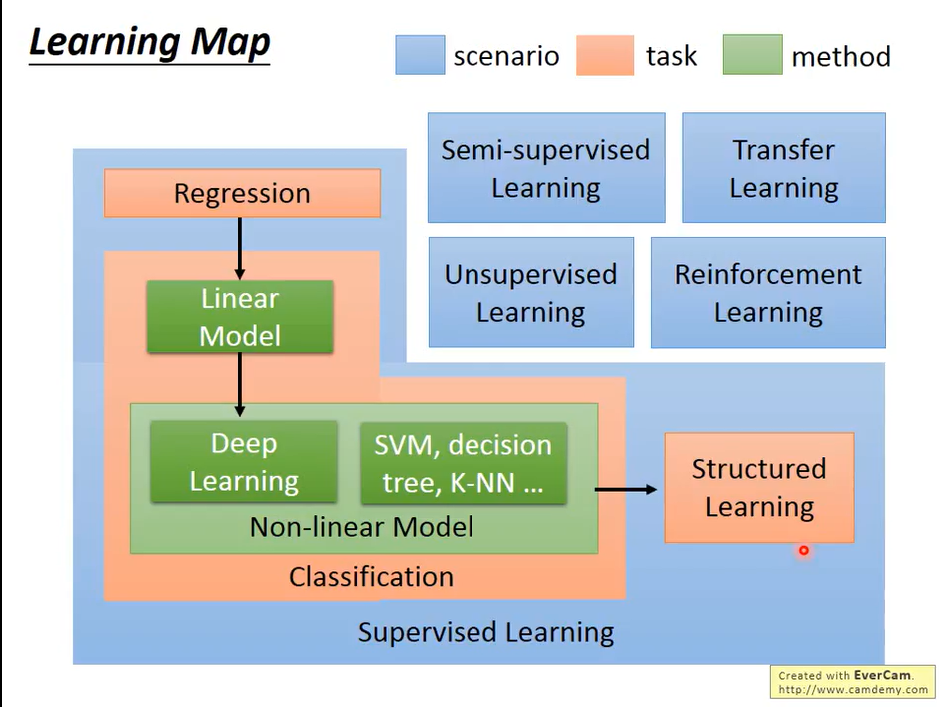
3 Steps:

Step1:Define a set of function

Step2:Goodnessof function

Step3:Pick the best function

Learning Map



## 线性回归

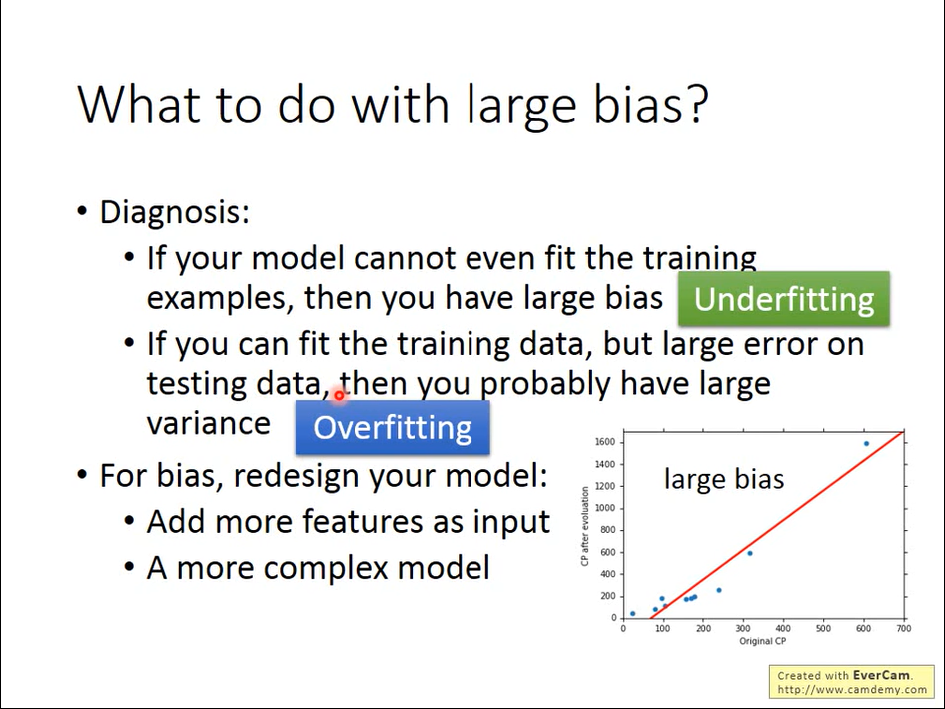
Consider loss function L(w) with one parameter w:

(Randomly) Pick an initial value w0

Compute and find the mini loss Function

## 错误分析

Error from two sides: bias and variance



### large bias

1.Redesign the model

Some other hidden factors

2.A more complex model

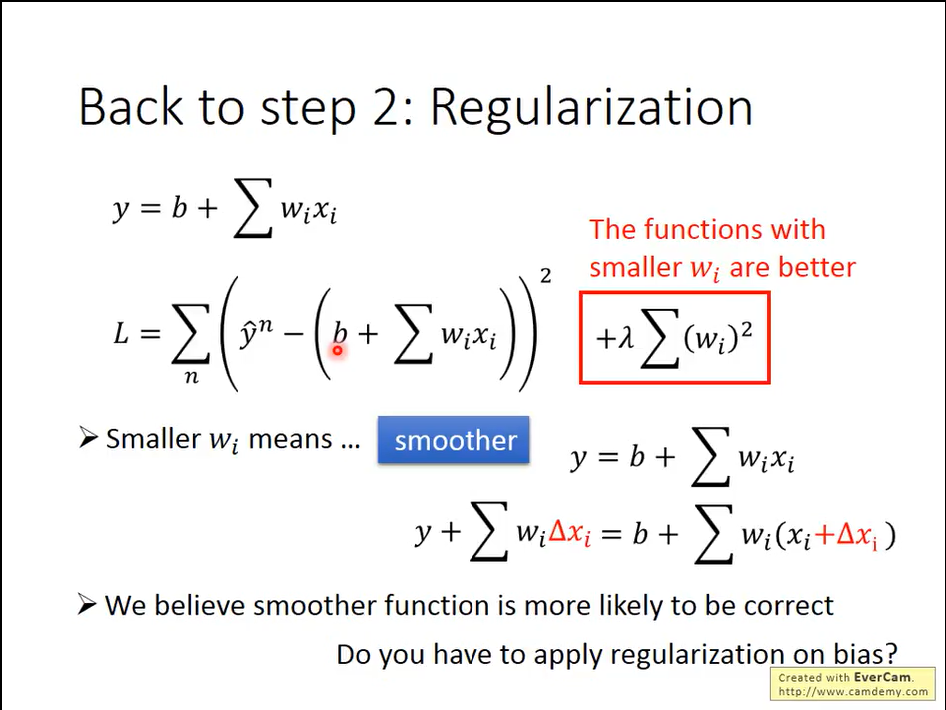
### large variance

Change a new loss function

1.more data

very effective,but not always practical(sometimes can make more data on our own)

2.regularization



More complex model has smaller bias and larger variance .

Underfitting(bias) and Overfitting(variance)

Make function much smoother, smoother function is more likely to be correct.

Bias 是常量，水平线，没有必要regularization

总结：2017/11/1

Overfitting的主要原因：

1.训练集和测试集特征分布不一致（白天鹅黑天鹅）

2.或者模型太过复杂（记住了每道题或者说参数量过多）而样本量不足

所以解决过拟合也从这两方面下手

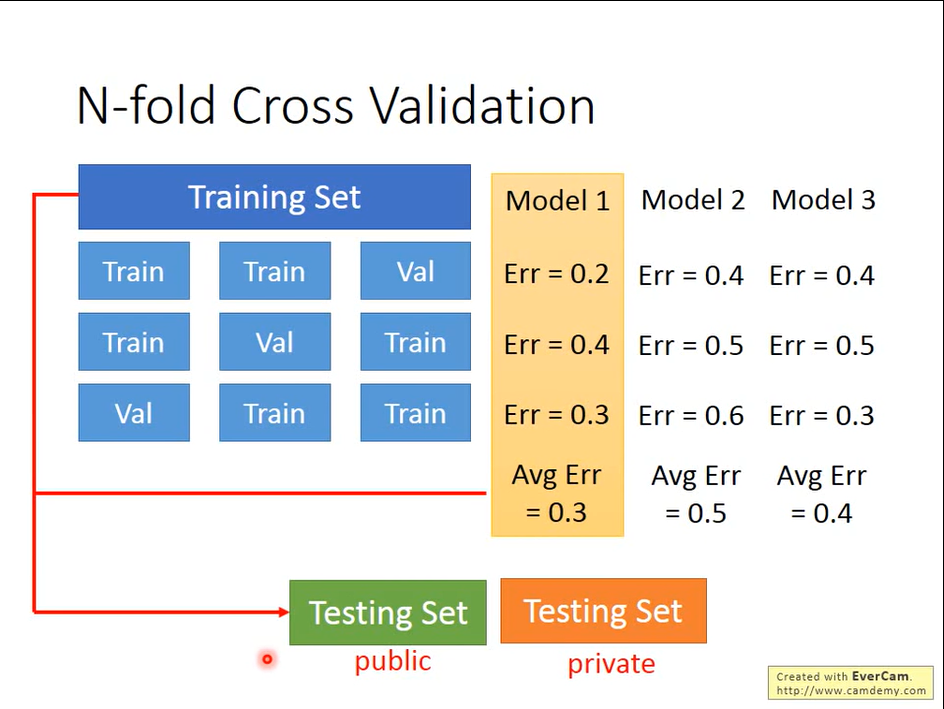
收集多样化的样本，简化模型（减少参数），交叉检验。

### Cross Validation

将training set分为两组，training set和validation set

Training set用来产生model，validation set用来选择model，选择后，可以再使用所有的training set产生model

### N-Cross Validation



N次划分，取平均值

### Adaptive Learning Rates

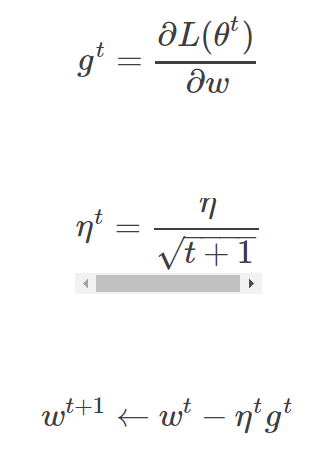
Learning rate cannot be one-size-fits-all

Giving different parameters different learning rates.

**1、Vanilla gradient descent（BGD）**

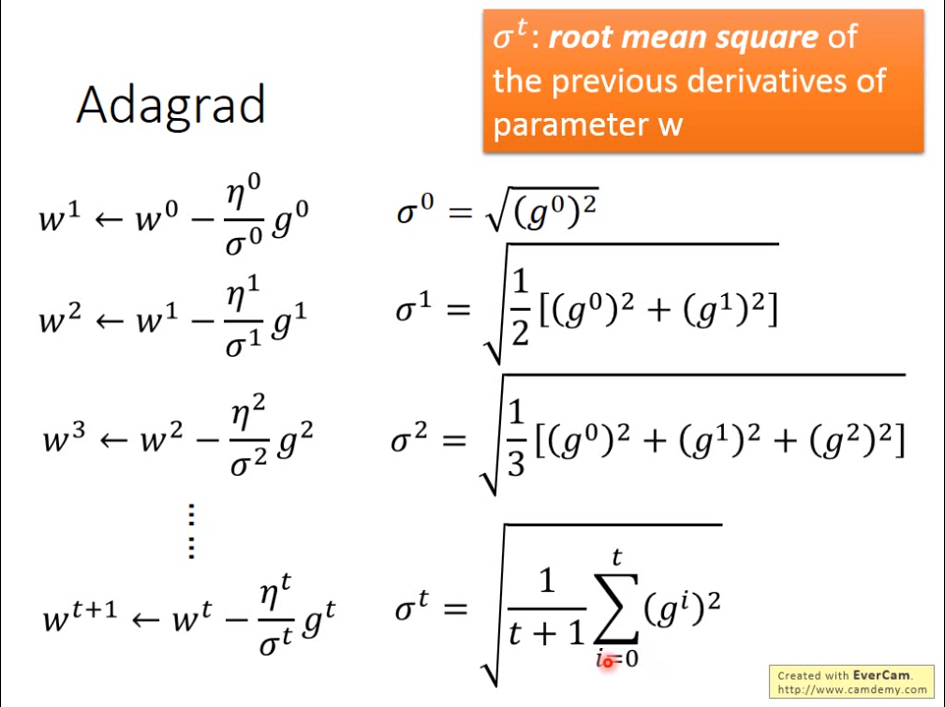
普通梯度下降算法（批梯度算法）

效率差



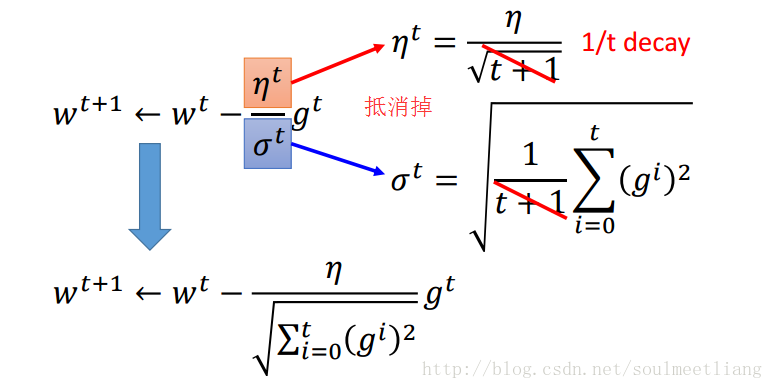
从Vanilla Gradient descent的公式中，我们可以看出，随着参数的更新，步长ηt越来越小，这比较符合直觉，因为随着参数的更新，参数很可能越来越接近极值点，此时我们需要更加小心的更新参数。（每次更新参数是在数据集的基础上计算的）

**2、Adagrad**



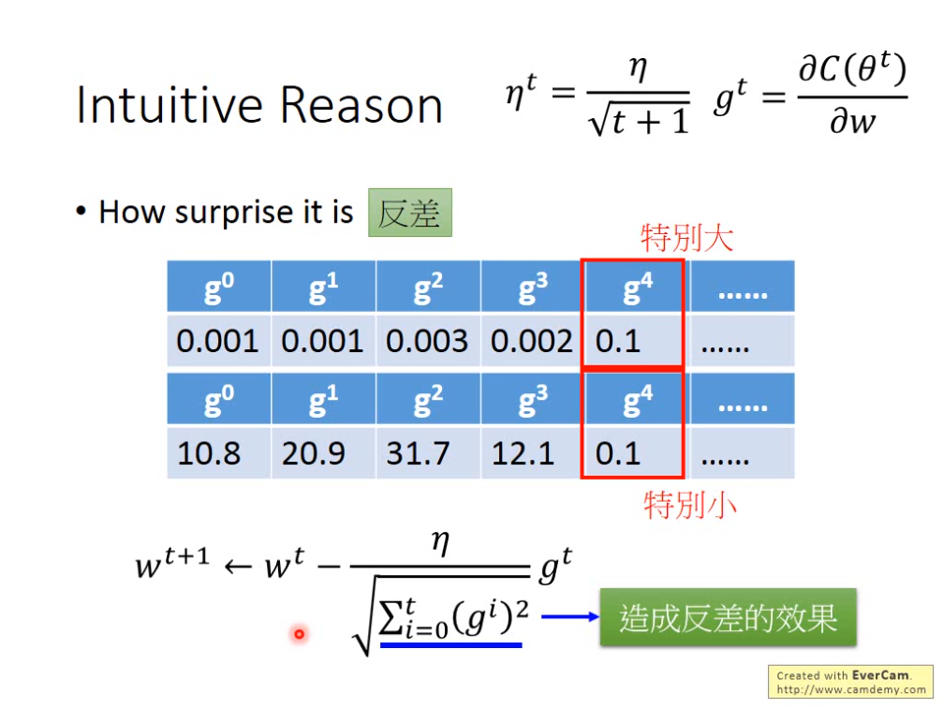
**参数g是gradient，梯度，即偏导**

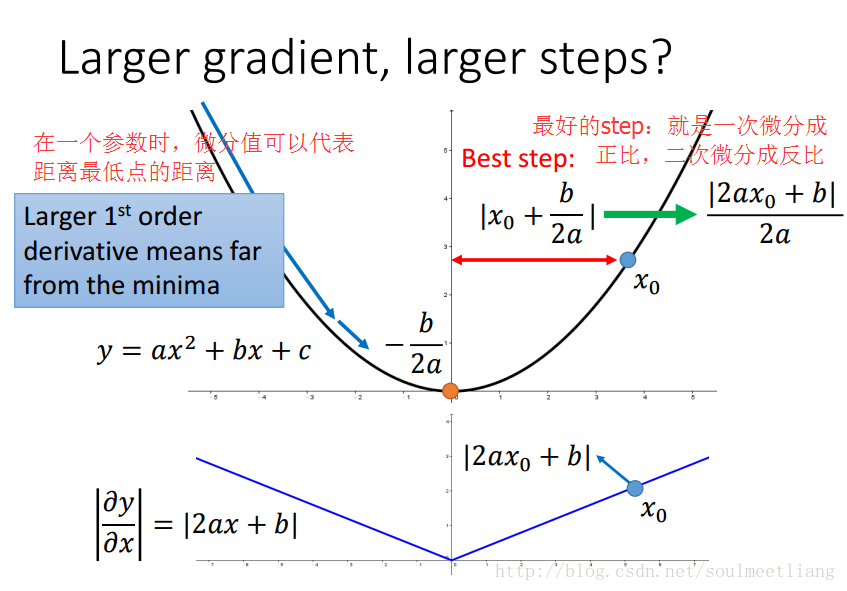
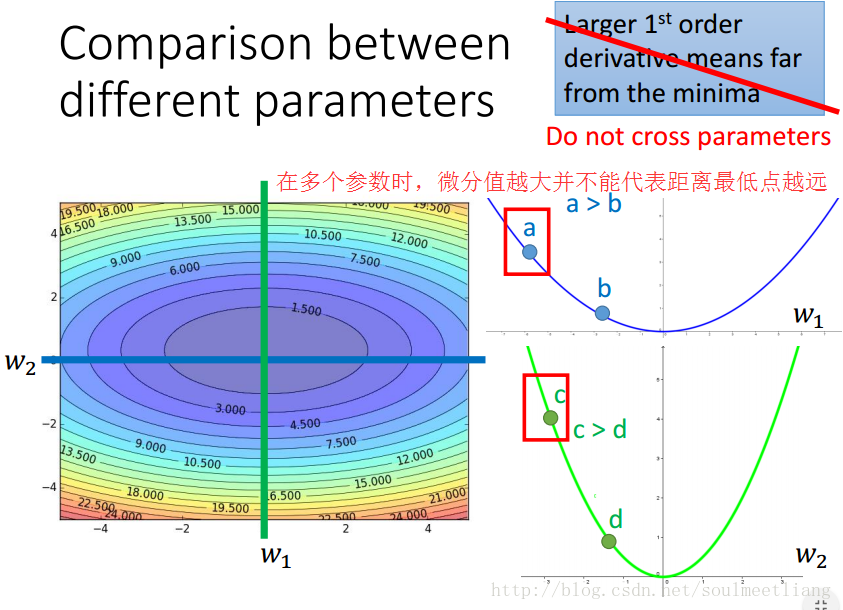
所以用Adagrad后，我们的参数变化要写成这样：

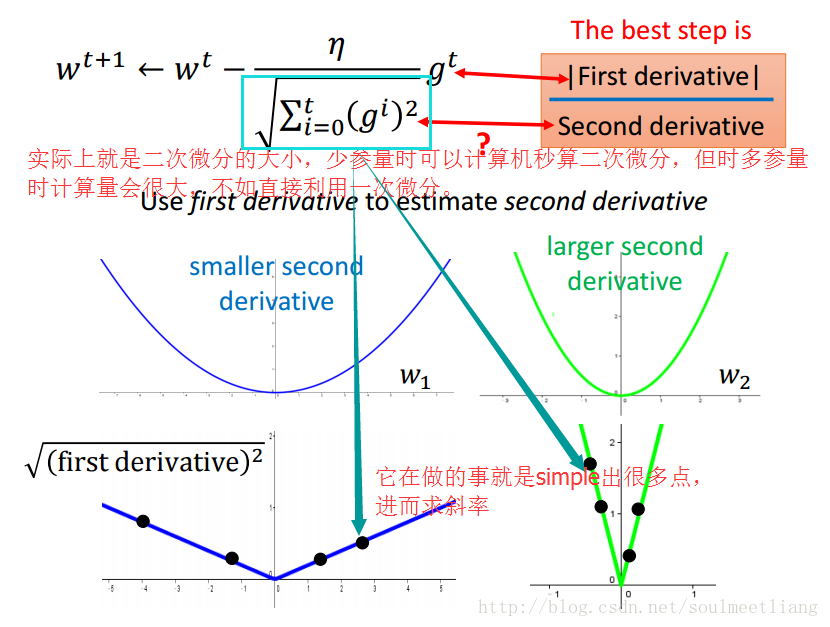


Adagrad越到后面改变会越慢，这是一个正常现象。

Q：有没有奇怪的地方？？是否有所冲突？？



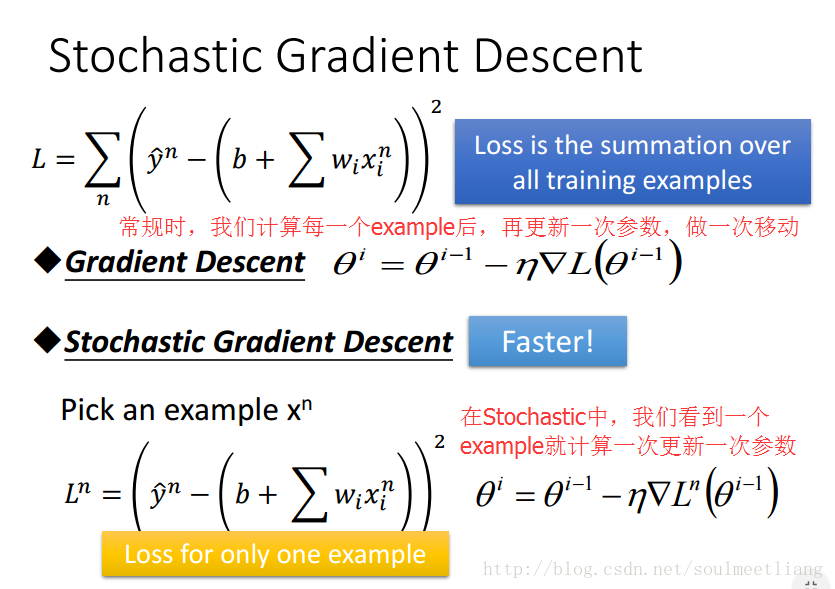
 

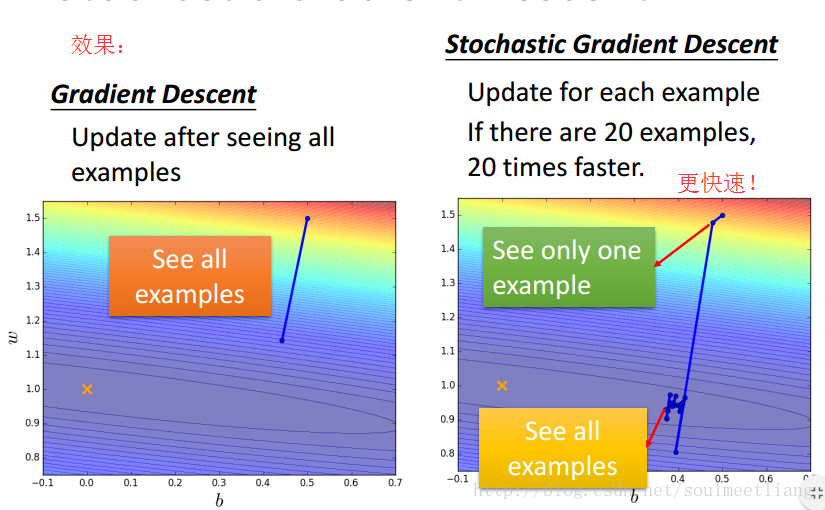


**在adagrad算法中，用一次微分的值去反映二次微分的值，在一个范围内sample多个点，.近似得到一阶导数函数的斜率**

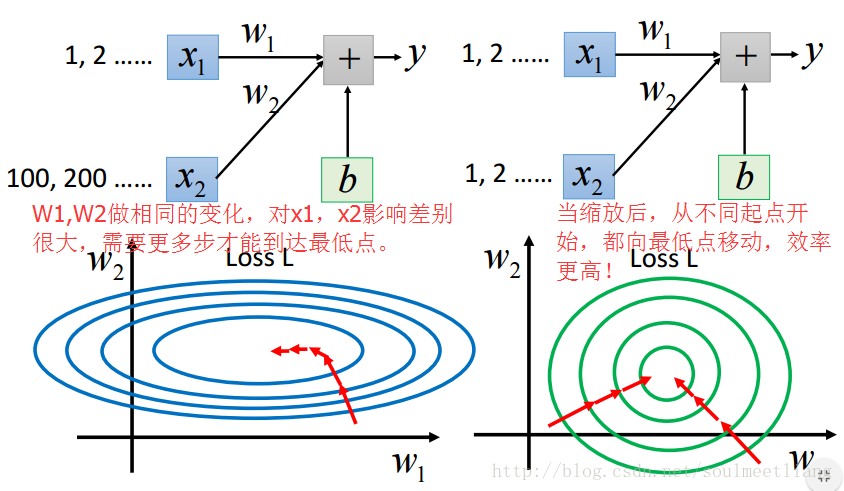
**3、Stochastic Gradient Descent(SGD)**

让你的Training更快一点！





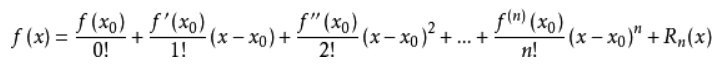
**4、Feature scaling**

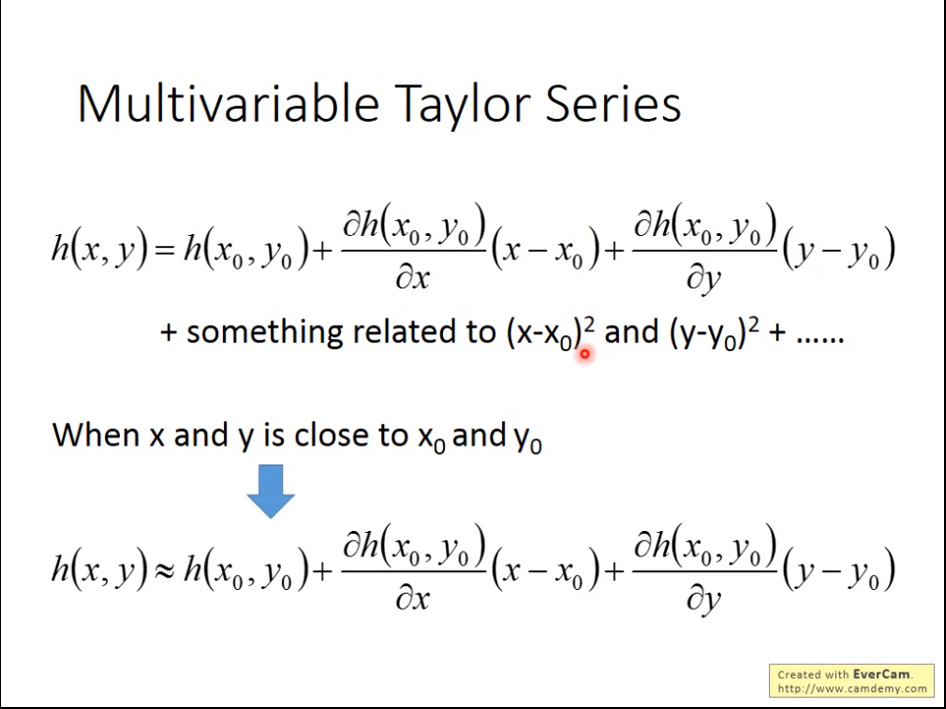


由于一个ranger中input的feature（x1、x2）的值不一样，导致对Loss Function的影响不一样，在Loss Function的三维surface中就体现为椭圆，但若缩放，则会变为正圆，效率更高。

**5、Warning of Math**

Taylor series=>Lagrange's mean value theorem=>finite increment theorem





根据[拉格朗日中值定理](https://baike.baidu.com/item/%E6%8B%89%E6%A0%BC%E6%9C%97%E6%97%A5%E4%B8%AD%E5%80%BC%E5%AE%9A%E7%90%86)导出的有限增量定理有：

https://gss3.bdstatic.com/-Po3dSag_xI4khGkpoWK1HF6hhy/baike/s%3D265/sign=22c6aa45b151f819f525044cefb54a76/279759ee3d6d55fba92f5a486c224f4a21a4ddf6.jpg

于是：

https://gss1.bdstatic.com/9vo3dSag_xI4khGkpoWK1HF6hhy/baike/s%3D230/sign=4df06440d31b0ef468e89f5dedc551a1/cefc1e178a82b901a3a2ef5d728da9773912ef55.jpg

其中误差α是在Δx→0即x→x0的前提下才趋向于0.

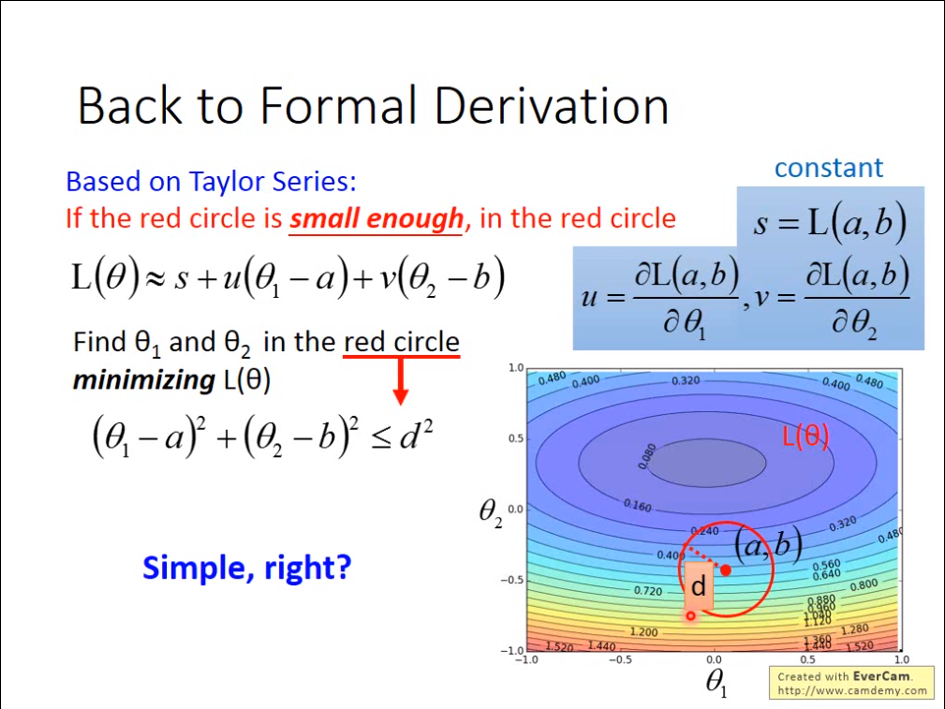
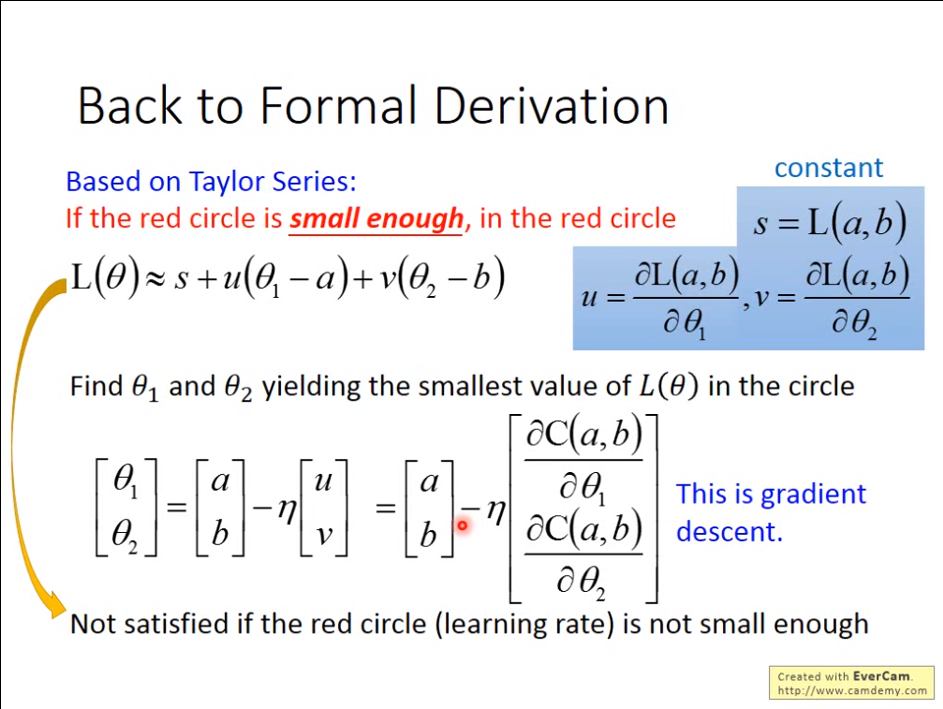
10-27 复习总结：

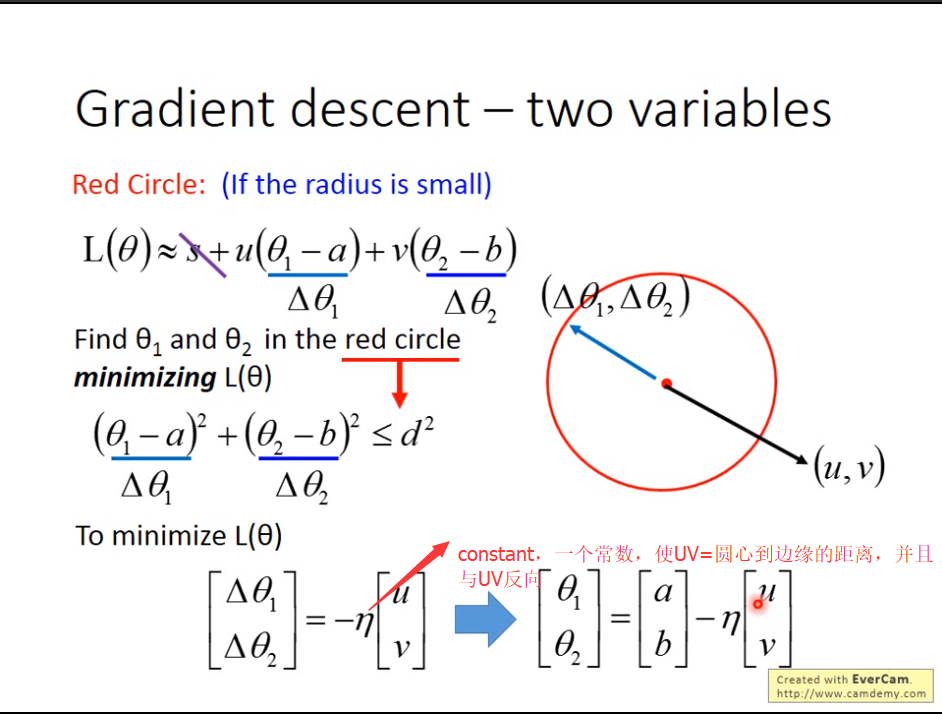
Gradient Descent是用来优化模型，Pick the best function的

但vanilla gradient descnet、adagrad、stochastic是用来adjust learning rates的

其中stochastic也用来pick the best function

Feature Scaling也是为了更快挑选出最佳函数





常数，与θ无关，所以不用管

