研究生算法课课堂笔记

上课日期： 2016 年 10 月 13 日　　第(2)节课

组长学号及姓名：陶淼　　1601214441
组员学号及姓名：王璐璐　1601214485

内容概要：

1. Missing value 的概念和处理方法
2. Bias and variance
3. Regularization
4. Decision tree regression

详细内容：

1. Missing value(数据缺失)

在实际应用过程中，可能某些样本在某些属性上的数据缺失(missing value)，在决策树中我们应该如何对值缺失的数据进行分类呢？

1.首先考虑一个简单的方法：根据数据集中该属性值的分布情况，选择比例较多的那一类作为缺失数据的值。这种方法是否合理？

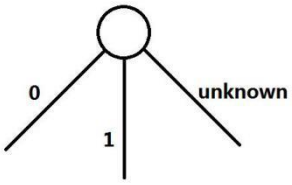
答案是不合理，我们以医院对病人的检查和诊断为例子：假如某个病人怀疑自己是否得了肿瘤，于是去医院检查，医院不会直接就对病人进行穿刺检查，而是先对病人进行一些“小”的初步检查，如果一个病人经初步检查后，医生觉得患肿瘤的可能性很小，那么就不会再对病人进行穿刺，此时医生诊断病人是否患肿瘤时，穿刺的数据就是缺失的。由于只有初步检查患肿瘤可能性很大的病人才会进行穿刺，因此穿刺的结果为阳性的比例比较大，按照这种缺失值的处理方法，

可能就会有较大的比例把缺失值补充为阳性，这显然与实际情况不符。因此这种方法处理缺失值是不合理的。

2.可行的方法：

（1）：把缺失值的属性视为目标属性，进行机器学习，求出缺失的值。

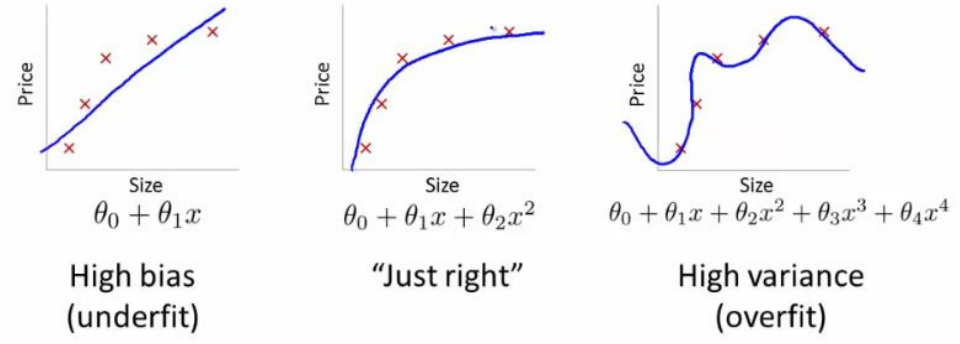
（2）：把缺失的值视为新的一类，将其值设置为 unknown。



3.xgboost 对数据缺失的处理方法：把缺失的值先全都放在第一类，再全都放在第二类，最后比较放在那一类效果好，就将缺失值放在效果好的那一类。

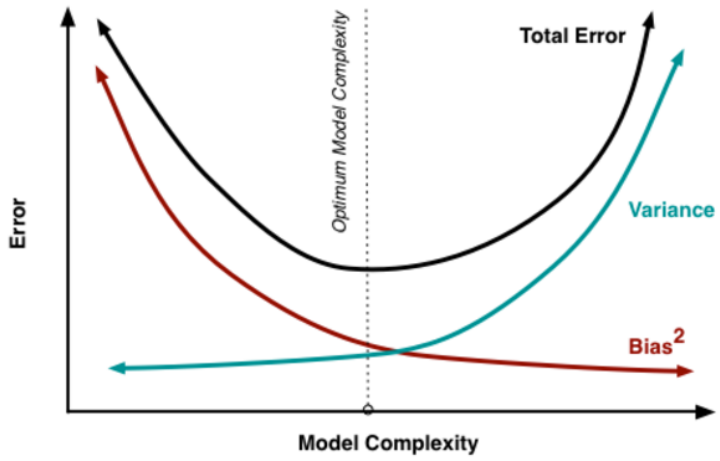
2. bias & variance

High bias 是高偏差，high variance 是高方差，前者指的是欠拟合，后者指过拟合，以多项式回归为例，如图：



造成 high bias 的原因主要是模型的复杂度过低拟合能力不足（比

如只有一层的决策树：决策树桩），造成 high variance 的主要原因是训练数据太少（比如数据量少于维度的数量）或模型的复杂度过高，拟合能力太强（比如：决策树的高度过高）。Bias variance 与模型复杂度的关系如下图所示：



问题：线性回归会出现 high variance 吗？

答案是会，前面讲过，high variance 的原因有两个：数据太少或模型复杂度太高。虽然线性回归是一个非常简单的模型，但是如果数据太少，比如样本属性的维度有 50 维，而样本数据只有 10 个，那么训练时自由度太高，训练出的模型能很好拟合那 10 个数据，但可能与实际分布相差甚远。

3.regularization(正则化)

正则化的目的是控制模型复杂度，防止模型过拟合。不同的模型采用的正则化的方法也不同，比如：线性分类器采用 L1 正则化或 L2 正则化，决策树正则化的主要方法是剪枝。

4.回归决策树：

主要内容：

1. 适用条件

2. 与分类决策树的区别

3. 构造方法&优化方法

详细内容：

1. 适用条件

目标值为连续数值

2. 与分类决策树的区别

（1）选择划分属性标准：按照使得 STD(standard deviation)减小最大的属性（分类树按熵减小最大）。

$$S=\sqrt{\frac{\sum(x-\mu)^2}{n}}$$

STD 的定义.

（2）叶子节点取值：通常取平均值（分类树选比例最大的）

（3）停止条件：属性用完，到达限制树高，到达限制节点数，STD<e（分类树目标属性都一样）

3. 构造方法

step1-pre:

选取属性。对连续型属性值，枚举切分点，选取在目标值出现变化处的切分点值作为回归树的属性。

Step 1: The standard deviation of the target is calculated.

Standard deviation (Hours Played) = 9.32

Step 2: The dataset is then split on the different attributes. The standard deviation for each branch is calculated. The resulting standard deviation is subtracted from the standard deviation before the split. The result is the standard deviation reduction.

		Hours Played (StDev)
Outlook	Overcast	3.49
	Rainy	7.78
	Sunny	10.87
SDR=1.66		

		Hours Played (StDev)
Temp.	Cool	10.51
	Hot	8.95
	Mild	7.65
SDR=0.17		

		Hours Played (StDev)
Humidity	High	9.36
	Normal	8.37
SDR=0.28		

		Hours Played (StDev)
Windy	False	7.87
	True	10.59
SDR=0.29		

$$SDR(T,X) = S(T) - S(T,X)$$

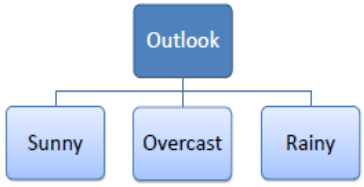
SDR(Hours , Outlook) = **S**(Hours) – **S**(Hours, Outlook)

= 9.32 – 7.66 = 1.66

Step 3: The attribute with the largest standard deviation reduction is chosen for the decision node.

★		Hours Played (StDev)
Outlook	Overcast	3.49
	Rainy	7.78
	Sunny	10.87
SDR=1.66		

Step 4a: Dataset is divided based on the values of the selected attribute.



Outlook	Temp	Humidity	Windy	Hours Played
Sunny	Mild	High	FALSE	45
Sunny	Cool	Normal	FALSE	52
Sunny	Cool	Normal	TRUE	23
Sunny	Mild	Normal	FALSE	46
Sunny	Mild	High	TRUE	30
Rainy	Hot	High	FALSE	25
Rainy	Hot	High	TRUE	30
Rainy	Mild	High	FALSE	35
Rainy	Cool	Normal	FALSE	38
Rainy	Mild	Normal	TRUE	48
Overcast	Hot	High	FALSE	46
Overcast	Cool	Normal	TRUE	43
Overcast	Mild	High	TRUE	52
Overcast	Hot	Normal	FALSE	44

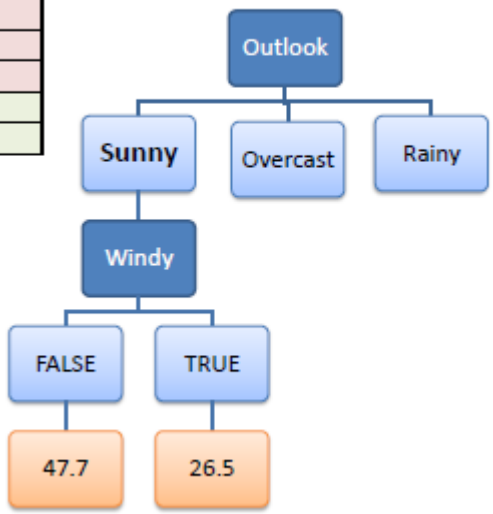
Step 4b: A branch set with standard deviation more than 0 needs further splitting.

In practice, we need some termination criteria. For example, when standard deviation for the branch becomes smaller than a certain fraction (e.g., 5%) of standard deviation for the full dataset OR when too few instances remain in the branch (e.g., 3).

Temp	Humidity	Windy	Hours Played
Mild	High	FALSE	45
Cool	Normal	FALSE	52
Mild	Normal	FALSE	46
Cool	Normal	TRUE	23
Mild	High	TRUE	30

★		Hours Played (StDev)
Windy	False	3.09
	True	3.50
SDR= 7.62		

SDR = 10.87-((3/5)*3.09 + (2/5)*3.5)



Step 5: The process is run recursively on the non-leaf branches, until all data is processed.

When the number of instances is more than one at a leaf node we calculate the average as the final value for the target.

step4-post:

计算叶子节点的平均值。

测试回归树：用 squared loss 值来衡量回归树的效果。squared loss 越小越好。