归纳偏差

- 归纳偏差:
  - 归纳偏差是一个关于机器学习算法的目标函数的假设.
  - 其实这个指的就是目标函数评分的标准.
    我们利用机器学习算法要做的是, 利用一个学习器, 通过学习样本使得学习器对于任意输入(可以不包含在训练数据中)可以产生正确的预测. 那么, 这个假设就决定了在面对未知数据下如何去作出判断. 例如, 在线性回归中, 作出的假设(归纳偏置)是, 输出与输入是线性的.下面是stackoverflow上的一个解释.(顺便吐槽下国内的很多博客但是简单的翻译甚至是复制粘贴, 不知道写成博客有什么用)
    Every machine learning algorithm with any ability to generalize beyond the training data that it sees has some type of inductive bias. This is the assumptions made by the model to learn the target function and to generalize beyond training data.
    For example in linear regression the model assumes that the output or dependent variable is related to independent variable linearly (in the weights).This is inductive bias in the model
- 之所以对这个概念难以理解, 是因为这个名词太没有烟火气, 其实他就是说的是**模型的指导规则**, 可以同时应用于训练和预测的时候的东西, 是一种超参数(因为是模型的一部分).上面举的例子是线性回归, 还有一种是决策树, 决策树的假设就是,
  - 优先选择较短的树而不是较长的。
  - 选择那些信息增益高的属性里根节点较近的树。
    这里利用了两个假设. 相比之下, 神经网络的假设是相当弱的, 举个例子:
    分类神经网络模型：将输入通过非线性函数进行映射的结果,正确的类别具有较高的softmax值.
- 归纳偏差的种类: 最大条件独立性（conditional independence）：如果假说能转成贝叶斯模型架构，则试着使用最大化条件独立性。这是用于朴素贝叶斯分类器（Naive Bayes classifier）的偏置。

- 最小交叉验证误差：当试图在假说中做选择时，挑选那个具有最低交叉验证误差的假说，虽然交叉验证看起来可能无关偏置，但天下没有免费的午餐理论显示交叉验证已是偏置的。
- 最大边界：当要在两个类别间画一道分界线时，试图去最大化边界的宽度。这是用于支持向量机的偏置。这个假设是不同的类别是由宽界线来区分。
- 最小描述长度（Minimum description length）：当构成一个假设时，试图去最小化其假设的描述长度。假设越简单，越可能为真的。见奥卡姆剃刀。
- 最少特征数（Minimum features）：除非有充分的证据显示一个特征是有效用的，否则它应当被删除。这是特征选择（feature selection）算法背后所使用的假设。
-