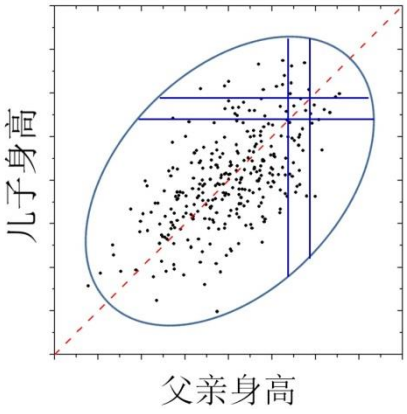
《机器学习及其在化学中的应用》

第一次书面作业题目

1.1“线性回归”是由生物学家兼统计学家高尔顿（达尔文的表弟）在研究人类遗传问题时提出来的。他搜集了大量的父子身高数据，发现其散点图大致呈直线状态，即趋势是父亲较高时儿子也倾向于较高。但是，高尔顿对数据进行深入分析后发现了一个有趣的现象—回归效应：当父亲高于平均身高时，儿子虽然平均会高于平均身高，但平均会矮于父亲。高尔顿最初认为向均值回归是一个因果过程，就像弹簧恢复到平衡长度一样。但后来又发现了一个更令人吃惊的事实：回归的代际顺序可以逆转，也就是说，当儿子高于平均身高时，父亲平均会高于平均身高，但平均会矮于儿子。高尔顿因此放弃了因果解释。



（1）阅读“线性回归的故事.pdf”

（2）记父亲身高为t，儿子身高为s。假设(s,t)服从高斯分布且s与t是对称的：$\langle s\rangle=\langle t\rangle=\mu$，$\langle(s-\langle s\rangle)^2\rangle=\langle(t-\langle t\rangle)^2\rangle=\sigma^2$，$\langle(s-\langle s\rangle)(t-\langle t\rangle)\rangle=r\sigma^2$，其中$0<r<1$。对“儿子身高~父亲身高”（即$s=f(t)=at+b$）进行回归分析，即求$a$与$b$。此时平均而言儿子与父亲哪个更高？

（3）与（2）类似，但对“父亲身高~儿子身高”（即$t=f(s)=as+b$）进行回归分析。此时平均而言儿子与父亲哪个更高？

（4）（免交）思考：为什么“父亲平均矮于儿子”与“儿子平均矮于父亲”能够同时成立？

1.2 口袋里有 50 枚正常铜钱与 50 枚狄青钱（两面都是正面）。从中掏出一枚，连抛 3 次，都是正面朝上，请问这枚铜钱是狄青钱的概率是多少？再抛一次（第 4 次）仍是正面朝上，那这枚铜钱是狄青钱的概率是多少？如果第 4 次的结果是反面朝上，那这枚铜钱是狄青钱的概率又是多少？（扩展阅读：算法之美-贝叶斯法则：预测未来.pdf）