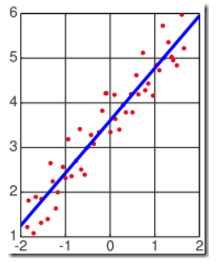
**最大似然与最小二乘**



    学过线性代数的大概都知道经典的最小二乘方法来做线性回归。问题描述是：给定平面上 N 个点，（这里不妨假设我们想用一条直线来拟合这些点——[回归](http://en.wikipedia.org/wiki/Regression_analysis" \t "http://blog.csdn.net/v_july_v/article/details/_blank)可以看作是[拟合](http://en.wikipedia.org/wiki/Curve_fitting" \t "http://blog.csdn.net/v_july_v/article/details/_blank)的特例，即允许误差的拟合），找出一条最佳描述了这些点的直线。

    一个接踵而来的问题就是，我们如何定义最佳？我们设每个点的坐标为 (Xi, Yi) 。如果直线为 y = f(x) 。那么 (Xi, Yi) 跟直线对这个点的“预测”：(Xi, f(Xi)) 就相差了一个 ΔYi = |Yi – f(Xi)| 。最小二乘就是说寻找直线使得 (ΔY1)^2 + (ΔY2)^2 + .. （即误差的平方和）最小，至于为什么是误差的平方和而不是误差的绝对值和，统计学上也没有什么好的解释。然而贝叶斯方法却能对此提供一个完美的解释。

    我们假设直线对于坐标 Xi 给出的预测 f(Xi) 是最靠谱的预测，所有纵坐标偏离 f(Xi) 的那些数据点都含有噪音，是噪音使得它们偏离了完美的一条直线，一个合理的假设就是偏离路线越远的概率越小，具体小多少，可以用一个正态分布曲线来模拟，这个分布曲线以直线对 Xi 给出的预测 f(Xi) 为中心，实际纵坐标为 Yi 的点 (Xi, Yi) 发生的概率就正比于 EXP[-(ΔYi)^2]。（EXP(..) 代表以常数 e 为底的多少次方）。

    现在我们回到问题的贝叶斯方面，我们要想最大化的后验概率是：

P(h|D) ∝ P(h) \* P(D|h)

    又见贝叶斯！这里 h 就是指一条特定的直线，D 就是指这 N 个数据点。我们需要寻找一条直线 h 使得 P(h) \* P(D|h) 最大。很显然，P(h) 这个先验概率是均匀的，因为哪条直线也不比另一条更优越。所以我们只需要看 P(D|h) 这一项，这一项是指这条直线生成这些数据点的概率，刚才说过了，生成数据点 (Xi, Yi) 的概率为 EXP[-(ΔYi)^2] 乘以一个常数。而 P(D|h) = P(d1|h) \* P(d2|h) \* .. 即假设各个数据点是独立生成的，所以可以把每个概率乘起来。于是生成 N 个数据点的概率为 EXP[-(ΔY1)^2] \* EXP[-(ΔY2)^2] \* EXP[-(ΔY3)^2] \* .. = EXP**{-[(ΔY1)^2 + (ΔY2)^2 + (ΔY3)^2 + ..**]} 最大化这个概率就是要最小化 (ΔY1)^2 + (ΔY2)^2 + (ΔY3)^2 + .. 。 熟悉这个式子吗？

    除了以上所介绍的之外，贝叶斯还在词义消岐，语言模型的平滑方法中都有一定应用。下节，咱们再来简单看下朴素贝叶斯方法。