机器学习入门第一课：从高中课本谈起

高中课本那些事

点连成线

上小学的时候，我们学过平面内，任意两点之间可以连成一条直线，且只能连成一条直线。

上初中的时候，我们学过$y = kx + b$,可以来表示二维空间内的一个平面，还记得每次中考前的模拟题，都会有一道题：

已知两点$A(x_1, y_1)$,$B(x_2, y_2)$,求过这两点A,B的直线方程。

这道题很简单，二元一次方程组秒杀。

尽可能多的点在一条直线上

上高中的时候，我们牛逼闪闪的高中老师，给我们出了一道牛逼闪闪的题，平面内一堆点，找出一条直线，使得尽可能的多的点在这条直线上。两点之间确定一条直线，这若干点，怎能搞，实在不会，老师教了我们一招绝技。

如何让尽可能多的点连在一条线上

这若干点，怎能搞，实在不会，老师教了我们一招绝技。

如果平面内有点$(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$,可用如下表达式来刻画这些点与直线$y = kx + b$的接近程度：

$[y_1 - (b + kx_1)]^2 + [y_2 - (b + kx_2)]^2 + \cdots + [y_n - (b + kx_n)]^2$

使得上式达到最小值的直线$y = kx + b$就是老师让我们求解的直线，老师说这种方法叫最小二乘法。

最后可以求解出$k = \frac{x_1y_1 + x_2y_2 + \cdots + x_ny_n - n\bar{x}\bar{y}}{x_1^2 + x_2^2 + \cdots + x_1^n - n\bar{x}}$, $b = \bar{y} - k\bar{x}$

高中老师没说过的那些事

点在面上

上初中的时候，我们学会了三点可以确定一个平面。

尽可能多的点在同一面上

上大学的时候，为了让我们以后挣钱挣得更多，于是乎自学起了更高级的数学，遇到一个难题，如果使得尽可能多的点在同一平面上。这道题可以折磨我等笨人好几天睡不着觉啊。有一天和女友散步，想起了高中时候老师讲的那道题，让尽可能多的点连在一条直线上。我灵机一动便有了思路。

如何让尽可能多的点连在一条线上

如果空间内有点$(x_1, y_1, z_1), (x_2, y_2, z_2) \cdots (x_n, y_n, z_n)$，可用如下表达式来刻画这些点与平面$z = ax + by + c$($c$为常数)的接近程度:

$$[z_1 - (c + ax_1 + by_1)]^2 + [z_2 - (c + ax_2 + by_2)]^2 + \cdots + [z_n - (c + ax_n + by_n)]^2$$

使得上式达到最小值的平面$z = ax + by + c$，就是这道题的答案。

数学到算法模型转化的步骤与工具

猜

用数学这把锋利的刀来求解未知问题，做到大胆猜想，往往就可以解决问题，从数学到算法模型转化过程中，猜的作用很大。

独立同分布

- 独立

  独立，顾名思义就是事件和事件之间相互不产生影响和作用，比如火星是行星和我是算法工程师之间就是独立的事件，没有相互影响或者彼此之间的作用。

  假设事件$A_1, A_2 \cdots A_n$的概率分别为$P_1, P_2 \cdots P_n$,这些独立事件同时发生的概率为：

$$P = P_1 P_2 \cdots P_n$$

- 同分布

  同分布指的是事件之间，事件的分布是等同的，一致的。

  独立同分布意味着事件独立且分布一致。

数据分布

在算法模型转化过程中，数据分布很重要，在某一派别中认为应该是假设数据分布(也就是猜出数据分布)，然后进行算法模型转化。

每一种数据分布，都有一种对应的概率。

推

先假设数据的分布，然后根据数据分布对应的概率来推导出需要求解的公式。

似然函数

似然函数，这个东西名字上来看上去绕来绕去，简而言之，就是求解参数。例如高中求解的k,b,大学求解的a,b,c。

工具

极大似然估计

极大似然估计，就是求似然函数的极大值点。

实操 线性回归模型

猜

平面内的点，一般满足正态分布，我们来猜这些点满足正态分布，而且是独立同分布的。

正态分布的数据满足，概率：

$$P=\frac{e^{-\frac{(x-\mu)^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma}$$

推

假设事件$A_1,A_2\cdots A_n$的概率分别为$P_1,P_2\cdots P_n$,这些独立事件同时发生的概率为：

$$P=P_1P_2\cdots P_n$$

这些点在尽可能在同一平面上的概率：

$P=\prod\limits_{i=1}^{n}P_i$

$=\prod\limits_{i=1}^{n}\frac{e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma}$

极大似然估计，即P最大，为P_{max}

$$P_{max} = \frac{e^{-max(\sum_{i=1}^{n}\frac{(x-\mu)^2}{2\sigma^2})}}{\sqrt{2\pi}\sigma}$$

即满足$min(\sum_{i=1}^{n}\frac{(x-\mu)^2}{2\sigma^2})$

看，这就是高中到大学的最小二乘法。

机器学习中的那些事 —— 以线性回归举例

任何问题都是有目的的

任何问题都是有目的的，线性回归$y = kx + b$就是目的，我们管这样的函数叫目标函数。

逼近正确就是让失败的情况最差

让失败的情况最差，当失败的情况无限接近最差的时候，就是逼近最正确的时候。最小二乘法就是让失败的情况逼近最差，也就是逼近正确。我们管最小二乘法这样$min(\sum_{i=1}^{n}\frac{(x-\mu)^2}{2\sigma^2})$函数叫损失函数，损失函数最小我们叫经验结构最小。