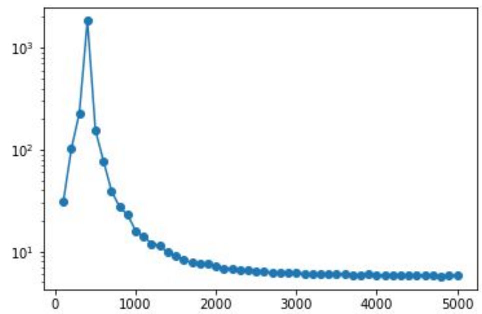
ML2017 report 1　　　　　　　　學號：r05945035 系級： 生醫電資碩一 姓名：陳育青

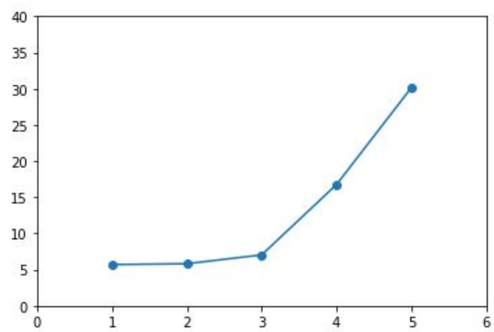
1. 請簡明扼要地闡述你如何抽取模型的輸入特徵 (feature)
答：我使用的特徵是除風向風力外每一個變項本身及其二次方的數值. 風向(θ)及風力(r)的部份是使用$r, \cos\theta, \sin\theta, r\cos\theta, r\sin\theta, r^2\cos^2\theta, r^2\sin^2\theta, r^2\cos\theta\sin\theta$.對於風向這麼仔細(aka 龜毛)好像也沒有比較準，但是我覺得比較有道理，畢竟360度和0度角是同一回事。為什麼沒有$$r^2$$呢？因為$$r^2, r^2\cos^2\theta, r^2\sin^2\theta$$三者線性相關，在線性回歸裡是多餘的。

2.請作圖比較不同訓練資料量對於PM2.5預測準確率的影響
答：如果使用internal validation的RMSE做為依據，訓練資料量100, 200, …., 5000筆資料的和RMSE的圖表為：



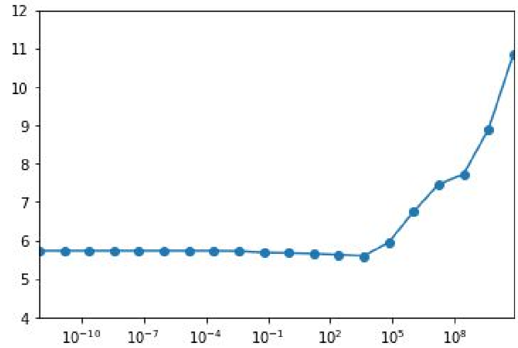
在訓練資料量大的時候RMSE接近5.8看起來資料量愈大愈好；訓練資料線小於1000的時候基本上和亂猜差不多。很特別的是如果訓練資料量小於400，RMSE反而會變小，而且不是因為random variation因為standard error不大，真的不知道要如何解釋。

3. 請比較不同複雜度的模型對於PM2.5預測準確率的影響
答：如果使用internal validation的RMSE做為依據，將取出的特徵做power series的線性回歸，其最大次方(max power)設定為1, 2, 3, 4, 5做回歸得到的RMSE作圖如下，看起來用一次方和二次方差不多，三次方以上會有overfitting的問題

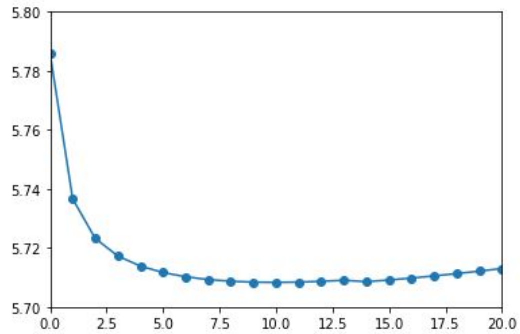


4. 請討論正規化(regularization)對於PM2.5預測準確率的影響
答：我覺得老師在課堂上講的正規化有點太簡化，課堂上的正規化版本是loss function多加一項$\lambda\sum_i w_i^2$，但是這最少有兩個問題，第一是所有的w_i所乘的λ是一樣的，但

是如果我們的測量的某個數值(假使是w_2)的單位如果有變化(例如km $\longleftrightarrow$ miles)但是其它的w_i單位不變，這個正規化就沒有一致，最佳解就會不一樣。換言之，λ沒有一致的單位或因次，如果要解決這個問題就是要先做normalization(去除單位效應)或是乘上Tikhonov matrix；第二是課堂上正規化的緣由是“We believe smoother function is more likely to be correct.”這是加諸先驗知識*a priori* knowledge, 但是不一定正確，如果不正確的話就是「腦補」(引述老師在3/16的課堂上對naïve Bayes as an *a priori* knowledge的評語)。說了那麼多，還是要交作業。以下是不做normalization的正規化RMSE對$\lambda = 16^{-10}, \ldots, 16^{10}$的作圖。看起來在$\lambda = 4096$附近有甜蜜點，超過的話就overregularization了。



如果先做normalization再正規化的RMSE圖如下, 在$$\lambda = 10$$附近有甜蜜點。很特別的是我做了normalized regularized prediction交到kaggle上的評分比沒有normalized還要差, 不知道是為什麼, 我本來以為normalized的方法會比較「有道理」。



這兩張圖X軸有的用log scale有的不用, 其實沒什麼特別原因, 望助教見諒。

5. 在線性回歸問題中，假設有 N 筆訓練資料，每筆訓練資料的特徵 (feature) 為一向量 x^n，其標註(label)為一**存純**量 y^n，模型參數為一向量w (此處忽略偏權值 b)，則線性回歸的損失函數(loss function)為 $\sum_{n=1}^{N} (y^n - w \cdot x^n)^2$ 。若將所有訓練資料的特徵值以矩陣 X = [x^1 x^2 … x^N] 表示，所有訓練資料的標註以向量 y = [y^1 y^2 … y^N]^{T}表示，請以 X 和 y 表示可以最小化損失函數的向量 w 。

答： $w = X^+ y, X^+ = (X^T X)^{-1} X^T$

如果$X^T X$是不可逆的矩陣，代表X有rank deficiency也就是說資料裡有線性相關的成分，此時w不唯一，不過沒關係X^+可以直接套用pseudo inverse的算法算出無限多個可以作為最小化損失函數的w裡的其中一個。