學號：R03922165　　系級：資工碩二　　姓名：李冠穎

以下是基於 AdaGrad，learning rate=1，batch size=32，epoch=1000 所進行的比較

1.

(1) 全部 9 小時污染源

public + private: 6.3918

(2) 全部 9 小時 pm2.5

public + private: 6.4782

可以發現雖然單單靠 pm2.5 即可以達到不錯的成績，但考慮各種污染源以讓準確度再提升一些，事實上透過將 18 種 feature 每一種單獨取 9 小時預測接下來的 pm2.5，可以發現到 pm2.5 本身是影響最大的，而 pm10 其次。

然而雖然其他污染源單單效果並不好，但集合再一起或許還是能提供一些資訊。

2.

(1) 全部 5 小時污染源

public + private: 6.56

全部 5 小時 pm2.5

public + private: 6.6589

由此可發現，事實上 5 小時的歷史資料已經可以達到不錯的效果，feature 若是比較多，頂多就是將多的維度所對應的 weight 設為 0，效果理論上來說應該至少要能一樣。在這邊我們可以發現事實上也大致符合如此。

當然，這件事或許還得考量是否會因為資料量的不足導致實務上不見得如此。

3.

以下 error 為 private, public => (public + private)分數

所有 features

0.1:　　6.10757, 10.09053 => 8.340

0.01:　 5.46006, 7.60733　=> 6.621

0.001:　5.38238, 7.31979　=> 6.642

0.0001: 5.39392, 7.28411　=> 6.409

PM2.5

0.1:　　6.74429, 10.24865 => 8.675

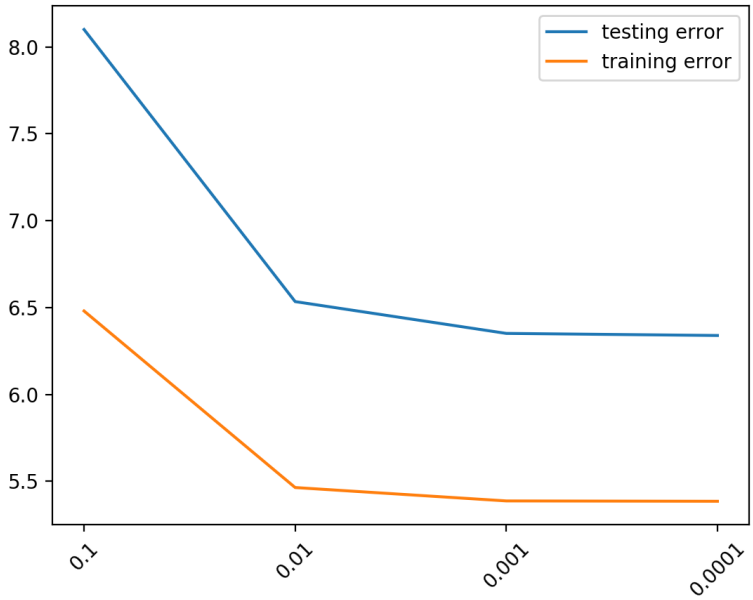
0.01:　 5.66111, 7.67156　=> 6.742

0.001:　5.54089, 7.33408　=> 6.5

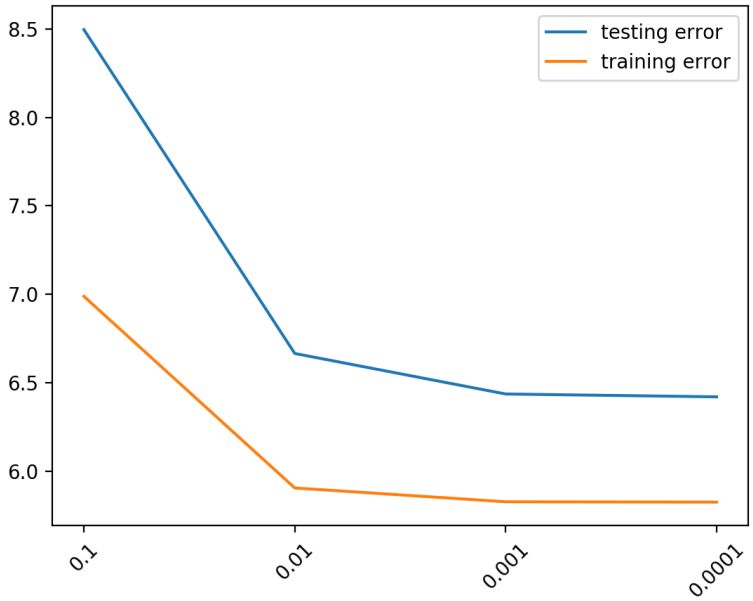
0.0001: 5.55699, 7.28406　=> 6.478

若把兩個 public + private error 平均之後作圖：

All features




PM2.5




可以看到隨著 lambda 變大，training error 也變大。因為曲線變的比較 smooth。

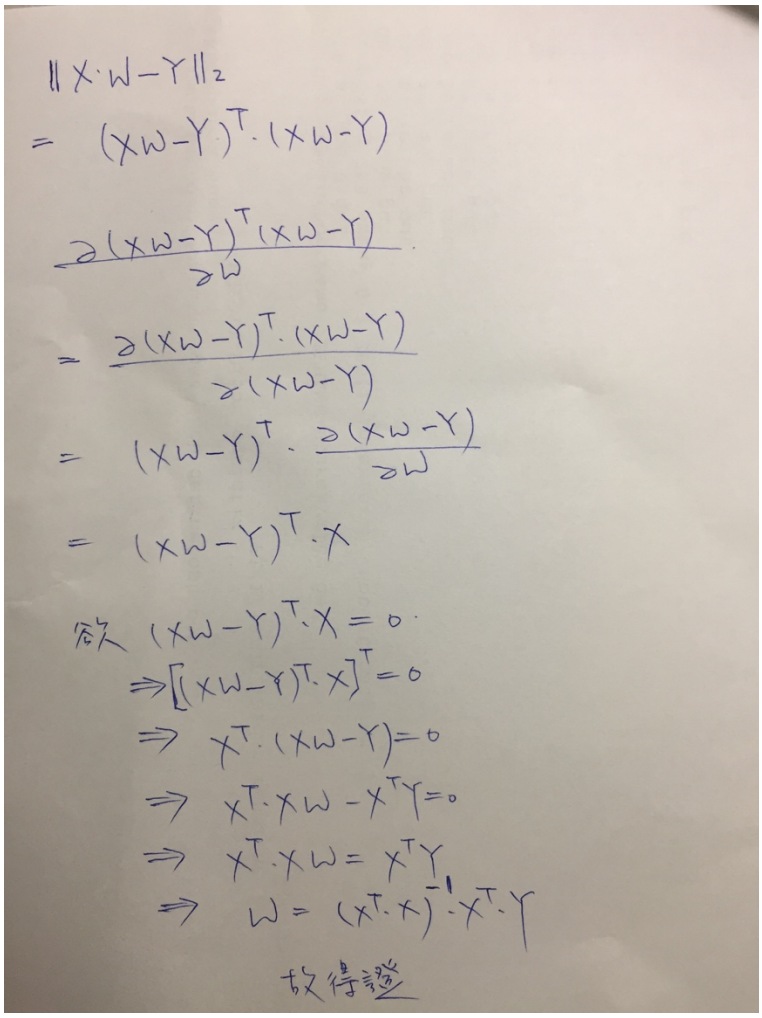
4.

答案為 (C) (XTX)-1XTy

證明如下

可以將 $\sum_{i=1}^{n}(y^n - x^n * w)^2$ 寫成矩陣的形式$||X * W - Y||_2$，也就是$X * W - Y$的 norm 2

學號：R03922165　　系級：資工碩二　　姓名：李冠穎




P.S. 懶得用 Words 打 equation 了…(請助教見諒)