# Data Science HW01

## 一、 Feature Encoding

### Attribute 1

原為字串型態，先將其月份資訊留存，成為其一 feature。

並且，依照其月份資訊，將其編成介於 -1 到 1 的 cos 函數和 sin 函數，具體實現為將 (月份 / 12 * 2π) 視為 radian，進行運算，依此編碼原因為，退測下雨與否和季節有所關係，而 sin 和 cos 函式合併來看，就具有週期性，因此採用此編碼方式，形成兩個獨立的 feature。

另外，透過其月份再將其推至可能的季節，在這邊做了較為廣泛的猜測將 3 - 5 月設為春天、6 – 8 月為夏天、9 - 11 為秋天、12 - 2 為冬天，期許能在季節上有一定的對應基準。

### Attribute 8 、9

原為表達風向的字串資料，總共分為 16 個方向。

原先想要將其編成介於 0 – 15 的整數資料，分別表示 16 個不同的方向，但是考慮到方向間的相關性低，這樣的編碼方式會隱含著數字大小比較和不同方向有所關係，因此放棄此種編碼。

最後，選擇將方向編為向量，具體實作為，將正北視為 0 度，每 22.5 度計算一次，cos 和 sin 的數值，分別將其存入 x 方向和 y 方向的向量中，由此可知會將一方向拆為兩個 feature作為表示。

### Attribute 20、21

原為表達是否下雨的字串資料，將其轉為 0 和 1，分別表示會下雨和不會下雨。

## 二、缺失值處理

在做完將非數值資訊轉為數值資訊後，將所有資料進行**線性插值**的運算，若處理玩依舊留有缺失值，就將此筆資料予以刪除

## 三、 Feature 的擴展

### 前言

由於對於各 feature 的解釋，能去透過一些運算來表示 feature 基於某特性而衍生的 feature 或 feature 間的關聯性，不但可以幫助在 feature importance 上較低的 feature 可能可以找到新的表達方式使其 importance 有所提升，更有機會增加數據的解釋性

### Feature 間的差距

由於我認為具有時間序的 feature 互相的變化量可以構成新的 feature，因此對於有早晚時間的資料都會將其數據相檢視為新的一筆資料

### Feature 間的交互關係

對於兩 feature 或我猜測兩者具有一定的相關性，舉例來說，這邊我任為方向向量和風速是具有相關度的 feature，因此將兩者相乘，並且作為新的 feature。在這邊做的運算方式都是兩者相乘的形式。

## 三、模型超參數的調整

在這邊引用了 XGBClassifier 進行預估是否下雨，而為了去調整此模型的超參數，在這邊使用了 RandomizedSearchCV 用以找尋相對好的超參數組合。

## 四、Feature 的挑選

使用 RamdoizedSearchCV 後可以觀察各 feature 在過程中所占的 importance 程度，而基於此可以剔除importance 較低的 feature，或是將其用其他方式表示為新的 feature。

## 五、成果和心得

在這過程中我嘗試了各式的編碼方式，其實對於各式編碼和擴展方式的好壞程度，要準確評估有其困難，在我最佳的成績，在 public 應可獲得 0.82 上下的成果，這是未使用額外資料的成果，我相信此可以作為我各項操作上的評分依據。