

# The Machine Learning Canvas (v0.4)











Designed for:

Bike Sharing Dataset Data Set

Designed by:

R09631007 吳乙澤

Date: 2021/11

<div>機器學習任務</div> <div></div> <div>預測的輸入和輸出是什麼？ 機器學習任務的類型是什麼？ 可選的演算法模型是什麼？</div> <div>本機器學習任務的輸入資料是天氣狀況 (weathersit)、環境溫度(temp)、體感溫度 (atemp)、濕度 (hum)、風速 (windspeed)、時間(hr)等，輸出則是當天的腳踏車租賃總數(cnt)。</div> <div>另外，本專案的類型為”預測(回歸)”，須根據輸入資料預測輸出。</div> <div>演算法模型的部分，Regression 和 NN 兩種演算法都能做預測。但是，助教有指定本次作業要用 Regression，因此不能使用 NN 去做預測。</div>	<div>決策行動</div> <div></div> <div>模型預測如何變成決策行動？</div> <div>可以透過模型來預測腳踏車租賃總數 (cnt)，藉此幫助廠商決定當天所有客戶需要的腳踏車數量，並規劃相對應的貨車司機，以便將腳踏車分配到腳踏車樁上。</div>	<div>問題定義</div> <div></div> <div>預測系統會為終端使用者帶來什麼價值？我們選擇什麼指標來解決問題？</div> <div>該預測系統的最終用戶是腳踏車租賃商。本系統可以預測腳踏車租賃總數 (cnt)，幫助廠商決定需要的腳踏車數量和運送司機。</div> <div>若模型有一定的準度，則腳踏車運送量和派出的司機會在適中的程度。如此一來，不僅能應付當天租借腳踏車的人流，也能節省貨車的燃油和司機的人事成本。</div>	<div>數據來源</div> <div></div> <div>我們可以使用哪些原始數據？</div> <div>本專案利用 UCI Machine Learning Repository 網站所提供的開放資料集，Bike Sharing Dataset Data Set。資料集的網站連結為 <a href="http://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset">http://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset</a>。</div>	<div>數據輸入輸出</div> <div></div> <div>來源的哪些資料作為訓練？哪些作為測試？</div> <div>訓練資料已經由助教從 UCI Machine Learning Repository 下載下來並放在 NTU COOL 裡。並且，訓練集和測試集已經被切割好，不須自行切割。</div> <div>打開並檢視 Bike_Sharing_test.csv 和 Bike_Sharing_train.csv 兩個檔案，可以發現測試集數量為 1000 筆，訓練集數量為 1500 筆。</div>
<div>線上預測</div> <div></div> <div>我們什麼時候會對輸入做出預測？多久做一次預測？</div> <div>本專案為資料礦掘的期中報告，因此，不須做線上預測。若要實際上線的話，我覺得至少一天做一次預測。具體預測的時間點，是在當天腳踏車租賃開始前，根據當天時間和感測器獲得的天氣數據，來預測腳踏車的租賃總數，以供後續的決策行動。</div>	<div>離線評估</div> <div></div> <div>部署之前，用什麼方法和指標來評估預測系統？</div> <div>可以利用 Bike_Sharing_test.csv 來評估預測系統。也就是說，觀察訓練出來的 Regression 模型，預測 Bike_Sharing_test.csv 的能力。透過比較 train 和 test 的 MAE 誤差，可以得知模型是否有 Overfitting 或 Underfitting 的情形。</div>	<div>特徵工程</div> <div></div> <div>從原始數據中提取什麼特徵(feature)？如何處理這些特徵？</div> <div>一開始，我不做任何前處理，將所有特徵都拿去做訓練。接著，視訓練結果，再利用 InAnalysis 的離群值過濾功能，對特徵做前處理，藉此來改善訓練結果。</div>		<div>建立模型</div> <div></div> <div>如何建立模型？用什麼演算法來訓練？</div> <div>利用張瑞益教授團隊所開發的 InAnalysis 線上系統，對 Bike Sharing Dataset 做前處理和機器學習模型的訓練。InAnalysis 有提供三種不同的 Regression 演算法供我們訓練用，這些演算法為 Linear Regression、Ridge Regression 和 Lasso Regression 三種。</div>
<div>即時評估和監測</div> <div></div> <div>部署之後，用什麼方法和指標來評估預測系統？</div> <div>如何量化它帶來的價值？</div>		<div>本專案為資料礦掘的期中報告，因此，不需要實際去部署模型。若有實際上線的話，我想以 MAE 誤差函數來評估預測系統。藉由比較預測系統的 MAE 及訓練集、測試集的 MAE，決定訓練出來的模型是否具備實用性。</div>		