大數據分析: 葡萄酒評鑑

組別: 7

組員: 蘇亮羽 Su Liang Yu 7111019015, 王柏皓 Wang Po Hao 4113029014

1. Introduction

葡萄酒曾被視為奢侈品，如今卻越來越受到多元化消費群體的喜愛。葡萄牙作為十大葡萄酒出口國之一，2005年時達到3.17%的全球市場占有率，其中vinho verde的出口量從1997年至2007年增長了36%。為了持續增長，葡萄酒產業正運用先進技術改進生產與銷售過程，並重視認證與品質評估，以防止偽造、保障消費者健康並維持市場信心。品質評估是釀酒改進及高端葡萄酒分類的重要環節，涉及物理化學檢測（如密度、酒精含量、酸鹼值）及由專家進行的感官分析。然而，由於味覺的複雜性及物理化學性質與感官評估之間的關係尚未完全理解，品質評估仍面臨挑戰。

這篇的報告的目的利用Bagging, Boosting, Decision tree, Linear regression四種不同的演算法預測出葡萄酒的偏好分數，並比較四種演算法的結果。

1. Materials and methods

資料集由助教提供，是2009年Paulo Cortez發表，對葡萄酒理化性質推斷偏好的論文中所使用的數據，採用的樣本為在2004/五月至2007/二月間，且依據CVRVV (Comissão de Viticultura da Região dos Vinhos Verdes)協會劃分的特定區域所生產出的葡萄酒，由iLab系統進行採樣，其中紀錄的理化性質包含: Fixed acidity (g(tartaric acid)/dm³), Volatile acidity (g(acetic acid)/dm³), Citric acid (g/dm³), Residual sugar (g/dm³), Chlorides (g(sodium chloride)/dm³), Free sulfur dioxide (mg/dm³), Total sulfur dioxide (mg/dm³), Density (g/cm³), pH, Sulphates (g(potassium sulphate)/dm³), Alcohol (vol.%)。此外在偏好分數的部分，由評估員進行三重複的盲測，對葡萄酒從0(非常差)到10(優秀)的評分。

在原數據中有紅酒(1599筆)及白酒(4898筆)的資料，在這次報告中只使用了白酒的資料進行運算。

1. Expected results

預計在資料進行一些pre-processing和regularization(L1, L2, L1+L2)後，比較regularization的差異，選擇結果佳的帶入演算法做運算。在課堂上老師有提到Random forest及Adaboost的ensemble方法，預計在最後討論的各個演算法的結果差異。

