# TIM 245: Final Examination

The final is due on **Thursday, 15** June 2017.

----------------------------------------------------------------------------------------------------------------------------------------------------------------

TIM-245 Course website: https://tim245-spring17-01.courses.soe.ucsc.edu/

----------------------------------------------------------------------------------------------------------------------------------------------------------------

**General Instructions** (applicable to all problems):

1. All work on the solutions to all final problems must be completely your own, except for research that must be properly attributed and cited. You cannot receive any help from anyone or give any help to anyone while doing this examination.
2. As part of your problem-solving approach, make appropriate **assumptions** when necessary, and clearly state the assumptions you make. Be sure to revisit these assumptions when drawing conclusions.
3. Be sure to explain everything you do. To improve readability of your explanations, **structure** your work with appropriate captions/headings.
4. In addition to headings, all explanatory text must be accompanied by **clearly labeled** figures, tables, and diagrams as necessary. Also, key points, results, and conclusions should be **clearly marked** (e.g., highlighted, underlined, etc.)
5. All results, recommendations, and conclusions must be supported by facts (evidence) and/or analysis. To improve readability, include **appendices** at the end of each problem to show relevant details.

# Problem Statement

It is Fall 2017 and the venture capital firm, Redwood Capital, having heard about your revolutionary data-mining work at Xenefits has hired you to help them improve their investment process using data mining. Redwood Capital invests in early stage start-up companies, like Xenefits, in exchange for equity or stock. Typically, these investments are very risky and only a small number, typically 1 in 10, of the companies are successful. Therefore, even small improvements in this process in can result in millions of dollars of profit for the firm.

The firm is hoping that you can help them apply data mining to solve the following three problems:

1. Predicting which companies will be successful (Classification Models).
2. Determining the common characteristics of successful companies (Association Analysis).
3. Identifying promising investment opportunities in new markets (Cluster Analysis).

# Problem 1: Classification Model for Venture Capital Investment (3 hour, 30 points)

Every year Redwood Capital reviews hundreds of start-up companies to potentially invest in. As part of their due diligence process, they perform detailed assessment of each company across a number of different factors including the experience of founding team, the technical feasibility of the product, and the size of the market. A complete description of the factors used in the assessment process is provided below:

|  |  |  |
| --- | --- | --- |
| **Founding Team** | **Product Value Proposition** | **Market Dynamics** |
| **EFF** - Capable of intense sustained effort  **GRO** - Able to evaluate and react to risk  **MKT** - Market Experience  **EXP** - Leadership Experience  **INO** - Able to Innovate  **CPL** - Team Completeness  **TCH** - Quality of Technical Team  **JNT** - Joint Experience of Team | **PTN** – Intellectual Property  **MAC** - Market Acceptance  **VPR** - Value Proposition  **PPL** - Product Execution Plan  **SCH** - Supply Chain Integration  **PRT** - Functioning Prototype | **MGR** - Market Growth Rate  **SOM** - Target Market Share  **MXS** - Targets Existing Market  **CMP** - Competitive Landscape  **MDV** - Market Diversity |

Based on this assessment, they then select a small number of companies to invest in. Eventually these selected companies are either successful and have an Initial Public Offering (IPO) or are shut down. The key metrics that Redwood Capital uses when evaluating an investment is the return on the invested capital, computed as follows: . For example, a 5X return on an investment of $10M would mean that the sale of the stock after the IPO would need to yield $60M in revenue for the firm.

The general guideline used by the firm to rate the quality of an investment is:

* Green = Excellent investment, company provided >5X returns
* Yellow = Fair investment, company provided 2-5X returns
* Red = Poor investment, company shutdown with no return

The firm has assembled a dataset of the historical investments that includes both the initial assessment of the company and final the investment rating (Green, Yellow, Red). You have been asked to perform the following tasks related to creating a classification model that be used to determine if a new company is likely to be a Green, Yellow, or Red investment based on their assessment.

1. Perform your EDA process, developed during your work at Xenefits, on the investment dataset and format the results into a well-structured report. Perform any data pre-processing steps (cleaning, transformation, etc.) necessary for addressing data quality issues that were discovered during the EDA process.
2. Before you start, the firm would like a **written statement of your process** for creating the classification model. *(Hint: The process might include the steps such as experimenting with different learning algorithms or model evaluation).*
3. Next, apply your process to create the classification model.
4. Based on your process, what is the best classification model for predicting an investment rating? What is your assessment of this classification model? Is the performance good enough for the firm to use when making investment decisions?
5. What are some of the issues associated with equally weighting misclassification across the three classes (hint: is a Red misclassified as a Green the same as a Yellow misclassified as a Green?). Describe one possible solution for addressing this problem.

**Extra Credit: Implement your solution for addressing the misclassification problem. Compare and contrast the results.**

# Problem 2: Association Analysis of Successful Companies (2 hour, 20 points)

Based on the results of the investment classifier, the firm would like to understand if there are any association rules for the attributes of a successful company. For example, are successful companies that have a good leadership team more likely to have better market positioning? These rules would allow them to better understand patterns across their investments and make recommendations for improvement.

To generate these rules, you have been asked to perform the following tasks:

1. Create a subset of the original investment dataset that only contains the successful companies, i.e. green label.
2. Explain the difference between support, confidence, and lift. How will changing the support and confidence change the discovered patterns?
3. Experiment with different values for support and confidence. What threshold do you recommend using for support and confidence?
4. Identify 2-5 interesting rules generated using your selected support and confidence threshold. What is the interpretation of the rule? What is underlying rationale or reason for the rule, e.g. the diapers -> beer rule was because young fathers were sent to the store to buy diapers.
5. What are some other potential applications of the generated rules? For example, would these rules be useful to the founder of a start-up company?

**Extra Credit**: **Repeat the rule generation process using only the unsuccessful companies, i.e. Red label. Compare and contrast the results to the association rules for the successful companies.**

# Problem 3: Clustering Crowdfunded Projects on Kickstarter (4 hours, 40 points)

Crowdfunding is an alternative source of funding for start-up companies where small investments, e.g. $100, are made by thousands of independent investors. These investments are typically managed through a crowdfunding portal, such as Kickstarter and Indie-Go-Go, where companies, or individuals, can post projects that they would like people to fund.

The crowdfunding market is potentially a new emerging investment opportunity for Redwood Capital. In addition to making several multi-million dollar investments every year, they could make small investments in hundreds of companies to diversify their portfolio and reduce risk. However, not all of the projects on crowdfunding sites are good candidates for commercialization. For example, projects often include humanitarian projects and art projects which would not be suitable for profit motivated investment firm like Redwood Capital.

The firm would like to further investigate these crowd-funding sites to see what kinds of projects are being funded and if they are a viable investment opportunity. To this end, they have collected a dataset of 4,000 projects from the popular crowd-funding site Kickstarter.

The dataset contains the following attributes:

* Project Title
* Description
* Amount Funded
* Project Goal
* Project Category
* Type of Currency
* Location
* Number of Backers

You have been asked to perform the following tasks related to understanding the different groups of projects that are being funded on Kickstarter:

1. First, remove the Project Title and Description fields from the dataset. Perform your EDA process on the remaining attributes and format the results into a well-structured report. Perform any data pre-processing steps (cleaning, transformation, etc.) necessary for addressing data quality issues that were discovered during the EDA process.
2. The firm has asked you to explain the differences between the following popular clustering algorithms: K-Means, DB-Scan, Hierarchical Clustering, and Mixture Models (EM). In particular, how do these algorithms with respect to computational complexity, types of clusters that they can find, and interpretability?
3. Before you start, the firm would like a **written statement of your process** for cluster analysis. *(Hint: The process might include the steps such as visualizing the data or reviewing clusters with domain experts)*
4. Then, they would like you to apply your cluster analysis process to the collected data-set and format the results into a well-structured report.
5. Identify 1-2 clusters of projects that would be good candidates for the firm to investigate further. Provide a brief summary description for each candidate cluster, e.g. video game projects developed in San Jose.

**Extra Credit**: **Create an additional set of attributes based on the project description and repeat your process for cluster analysis.**

# Problem 4: Brainstorming for Financial Investment Data Mining (1 hour, 10 points)

The firm has asked you for new ideas about how data mining can be used to improve their investment strategy. Apply a structured brainstorming process to generate 3-5 possible ideas for using data mining to improve financial investments.

For each generated idea, provide a brief description of:

1. What type of investment problem is being addressed, e.g. identifying promising new start-up companies or determining how much money to invest in a particular company.
2. What type of data mining task is involved: classification, prediction, cluster analysis, or association analysis.
3. What data-sets would need to be collected for the data mining task.
4. How could Redwood Capital use the resulting model (supervised learning) or patterns (unsupervised learning)?