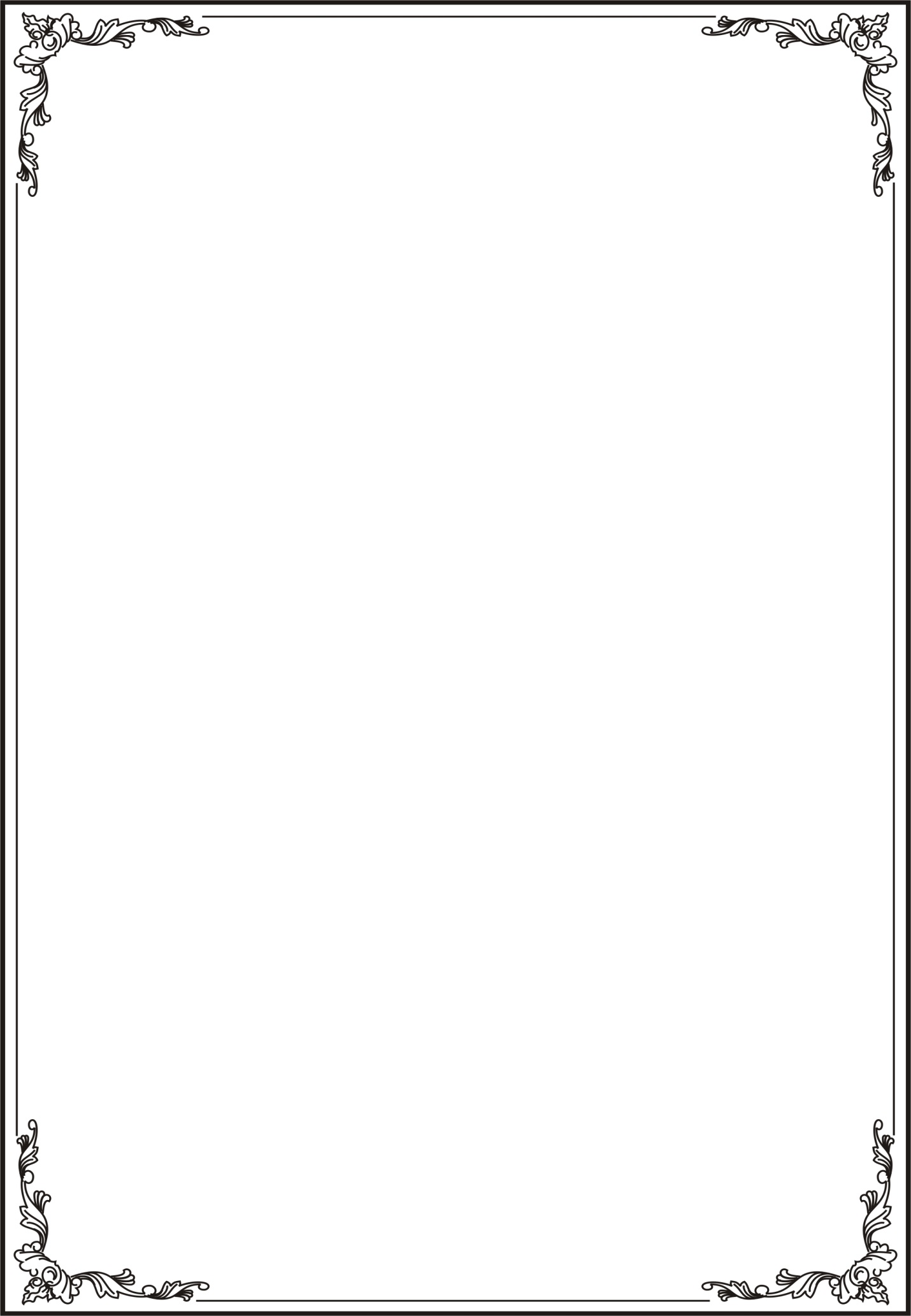
**ĐẠI HỌC QUỐC GIA TP.HCM**

**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**

**KHOA KHOA HỌC MÁY TÍNH**



***BÁO CÁO CASE STUDY 2***

***Đề Tài: Linear Regression – Boston Housing Data***

***Môn Máy học***

***Lớp CS114.J22.KHCL***

***GVHD: Huỳnh Thị Thanh Thương***

**Sinh viên thực hiện: Huỳnh Minh Tuấn – 17521212**

**Nguyễn Thanh Tú – 17521201**

**Hoàng Ngọc Quân – 17520934**

**Ngô Anh Vũ - 17521272**

**TP.HCM, ngày 21 tháng 4 năm 2019**

**MỤC LỤC**

**CHƯƠNG I: GIỚI THIỆU VỀ BÀI TOÁN HỒI QUY TUYẾN TÍNH VỚI BOSTON HOUSING DATASET** 2

I. Giới thiệu 2

II. Phát biểu bài toán 2

III. Cách đánh giá mô hình 2

**CHƯƠNG II: XÂY DỰNG MÔ HÌNH TỪ TẬP HUẤN LUYỆN**  4

I. Mã giả các thuật toán gradient descent và normal equation 4

II. Minh họa thuật toán gradient descent và normal equation 5

**CHƯƠNG III: THỰC NGHIỆM ĐÁNH GIÁ** 5

I. Xử lý bộ dữ liệu từ dataset 5

II. Các thuật giải Gradient Descent 5

III. Các thuật giải Normal Equation 5

**CHƯƠNG IV: LẬP TRÌNH CÀI ĐẶT** 7

I. Tool, thư viện đã sử dụng 5

II. Source code báo cáo 5

**PHỤ LỤC** 5

**CHƯƠNG I: GIỚI THIỆU VỀ BÀI TOÁN HỒI QUY TUYẾN TÍNH VỚI BOSTON HOUSING DATASET**

1. Giới thiệu:

* Hiện nay, với sự phát triển của xã hội, nhu cầu sinh hoạt, chỗ ở hợp lí ngày càng tăng theo. Do đó, nhu cầu mua bán nhà ở ngày càng được quan tâm hơn. Vậy làm thế nào để người mua nhà có thế tìm mua được một căn nhà ưng ý với mình ?
* Xuất phát từ nhu cầu trên, chúng ta phải xây dựng một mô hình tổ chức quản lý dữ liệu lưu trữ thông tin, đặc điểm của những ngôi nhà được mở bán. Trong bài báo cáo này, nhóm chúng em sử dụng dữ liệu của 506 ngôi nhà ở Boston để dự đoán xem tỉ lệ một ngôi nhà có 13 đặc điểm đã cho trước giá có trên 1000$ hay không. Trong bài báo cáo này, nhóm chúng em sử dụng các thuật toán Gradient Descent (Batch, Stochatic và Mini-batch), Normal Equation (dạng basic)

1. Phát biểu bài toán

* Bài toán dự đoán giá nhà được phát biểu như sau: Cho trước 1 tập dữ liệu gồm 13 thuộc tính D = { d1, d2, d3,…, dn} và tập các kết quả dự đoán giá nhà trên 1000$ C = { c1, c2,…,cn}. Nhiệm vụ của bài toán là với số thuộc tính đã cho trên mỗi dòng dữ liệu thì liệu kết quả dự đoán thực tế có giống hoặc gần giống với kết quả đã cho trước hay không. Hay nói cách khác là ta đi tìm w của hàm E(w) sao cho hàm đạt giá trị nhỏ nhất:

Với:

* E(w): độ mất mát (độ lỗi)
* N: tổng số dữ liệu
* xi: Tổng các giá trị thuộc tính
* yi: Kết quả đầu ra trên lý thuyết
* w: giá trị cần tìm

1. Cách đánh giá mô hình

Đánh giá mô hình Gradient Descent với công thức:

Với:

* Error: độ lỗi của hàm
* : kết quả đầu ra của đề bài cung cấp
* : kết quả đã tính được

**CHƯƠNG II: XÂY DỰNG MÔ HÌNH TỪ TẬP HUẤN LUYỆN**

1. Mã giả các thuật toán gradient descent và normal equation
2. Batch Gradient Descent

* Trong khi thực hiện lặp n lần:
* Tính giá trị hàm mất mát với công thức trên toàn bộ tập dữ liệu đầu vào
* Cập nhật lại w sau mỗi lần tìm được giá trị hàm mất mát: w = w – learning rate dw với dw là giá trị hàm mất mát, learning rate phải chọn sao cho không được quá lớn hoặc quá nhỏ (quá lớn sẽ không hội tụ tại điểm nhỏ nhất trên đồ thị, quá nhỏ sẽ di chuyển về điểm nhỏ nhất chậm).
* Tính độ lỗi như mô hình đánh giá bên trên sau mỗi lần cập nhật w.

1. Stochatic Gradient Descent

* Trong khi thực hiện lặp n lần:
* Thực hiện tính giá trị hàm mất mát trên từng dòng dữ liệu với công thức
* Cập nhật lại w sau mỗi lần tìm được giá trị hàm mất mát: w = w – learning rate dw
* Tính độ lỗi như mô hình đánh giá bên trên sau mỗi lần cập nhật w.

1. Mini-Batch Gradient Descent

* Trong khi thực hiện lặp n lần:
* Tương tự thực hiện tính giá trị hàm mất mát như Stochatic nhưng lần này ta chọn ngẫu nhiên 1 “nhúm” nhỏ trong bộ dữ liệu để tính (thường “nhúm” nhỏ chọn từ 50 đến 100 dòng dữ liệu).
* Cập nhật lại w sau mỗi lần tìm được giá trị hàm mất mát: w = w – learning rate dw
* Tính độ lỗi như mô hình đánh giá bên trên sau mỗi lần cập nhật w.

1. Normal Equation

* Tính trên tất cả các dòng dữ liệu với công thức