MSCI 546: Advanced Machine Learning

Sirisha Rambhatla

University of Waterloo

Lecture
Linear Regression

Outline

Outline

Regression

Predicting a continuous outcome variable using past observations

- Predicting future temperature
- Predicting the amount of rainfall
- Predicting the demand of a product
- Predicting the sale price of a house
- ...

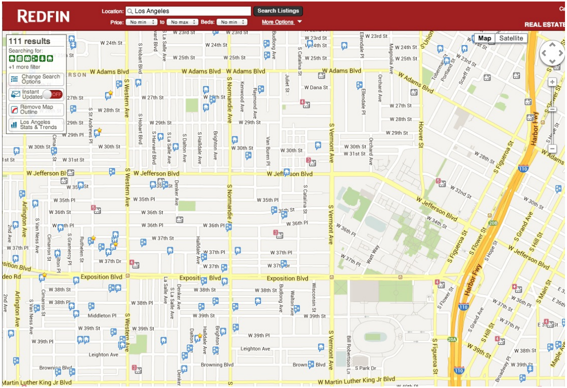
Key difference from classification

- continuous vs discrete
- measure *prediction errors* differently.
- lead to quite different learning algorithms.

Linear Regression: regression with linear models

Ex: Predicting the sale price of a house

Retrieve historical sales records (training data)



Features used to predict

3620 South BUDLONG — Los Angeles, CA 90007 — Status: Closed

$1,510,000 Last Sold Price | **14** Beds | **6** Baths | **4,418** Sq. Ft. $342 / Sq. Ft.

Built: 1956 Lot Size: 9,649 Sq. Ft. Sold On: Jul 26, 2013

Overview | Property Details | Tour Insights | Property History | Public Records | Activity | Schools

1 of 12

Five unit apartment complex within 2 blocks of USC campus, Gate #6. Great for students (most student leases have parents as guarantors). Most USC students live off campus, so housing units like this are always fully leased. Situated on a gated, corner lot, and across from an elementary school, this complex was recently renovated, and has in-unit laundry hook ups, wall -unit AC, and 12 parking spaces. It is within a DPS (Department of Public Safety) and Campus Cruiser patrolled area. This is a great income generating property, not to be missed!

Property Type: Multi-Family
Style: Two Level, Low Rise
Community: Downtown Los Angeles
County: Los Angeles
MLS#: 22176741

Property Details for 3620 South BUDLONG, Los Angeles, CA 90007

Details provided by i-Tech MLS and may not match the public record. Learn More.

Interior Features

Kitchen Information
- Remodeled
- Oven, Range

Laundry Information
- Inside Laundry

Heating & Cooling
- Wall Cooling Unit(s)

Multi-Unit Information

Community Features
- Units in Complex (Total): 5

Multi-Family Information
- # Leased: 5
- # of Buildings: 1
- Owner Pays Water
- Tenant Pays Electricity, Tenant Pays Gas

Unit 1 Information
- # of Beds: 2
- # of Baths: 1
- Unfurnished
- Monthly Rent: $1,700

Unit 2 Information
- # of Beds: 3
- # of Baths: 1
- Unfurnished
- Monthly Rent: $2,250

Unit 3 Information
- Unfurnished

Unit 4 Information
- # of Beds: 3
- # of Baths: 1
- Unfurnished
- Monthly Rent: $2,350

Unit 5 Information
- # of Beds: 3
- # of Baths: 2
- Unfurnished
- Monthly Rent: $2,325

Unit 6 Information
- # of Beds: 3
- # of Baths: 1
- Monthly Rent: $2,250

Property / Lot Details

Property Features
- Automatic Gate, Card/Code Access
- Automatic Gate, Lawn, Sidewalks
- Corner Lot, Near Public Transit

Lot Information
- Lot Size (Sq. Ft.): 9,649
- Lot Size (Acres): 0.2215
- Lot Size Source: Public Records

Property Information
- Updated/Remodeled
- Square Footage Source: Public Records
- Tax Parcel Number: 5040017019

Parking / Garage, Exterior Features, Utilities & Financing

Parking Information
- # of Parking Spaces (Total): 12
- Parking Space
- Gated

Building Information
- Total Floors: 2

Utility Information
- Green Certification Rating: 0.00
- Green Location: Transportation, Walkability
- Green Walk Score: 0
- Green Year Certified: 0

Financial Information
- Capitalization Rate (%): 6.25
- Actual Annual Gross Rent: $128,331
- Gross Rent Multiplier: 11.29

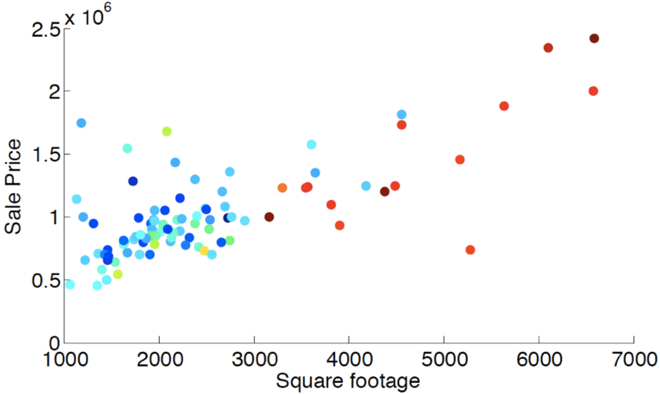
Location Details, Misc. Information & Listing Information

Location Information
- Cross Streets: W 36th Pl

Expense Information
- Operating: $37,664

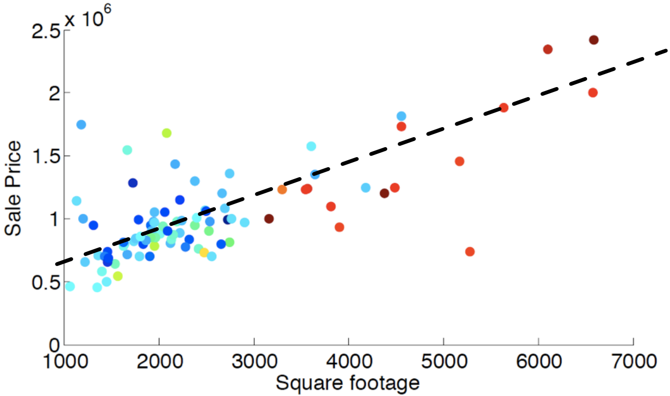
Listing Information
- Listing Terms: Cash, Cash To Existing Loan
- Buyer Financing: Cash

Correlation between square footage and sale price



Possibly linear relationship

Sale price ≈ **price_per_sqft** × square_footage + **fixed_expense**
(*slope*) (*intercept*)



How to learn the unknown parameters?

How to measure error for one prediction?

- The classification error (0-1 loss, i.e. *right* or *wrong*) is *inappropriate* for continuous outcomes.
- We can look at
  - *absolute* error: | prediction - sale price |
  - or *squared* error: (prediction - sale price)^2 (**most common**)

Goal: pick the model (unknown parameters) that minimizes the average/total prediction error, but *on what set*?

- test set, ideal but we *cannot use test set while training*
- training set ✓

Formal setup for linear regression

Input: $\boldsymbol{x} \in \mathbb{R}^{\mathsf{D}}$ (features, covariates, context, predictors, etc)

Output: $y \in \mathbb{R}$ (responses, targets, outcomes, etc)

Training data: $\mathcal{D} = \{(\boldsymbol{x}_n, y_n), n = 1, 2, \ldots, \mathsf{N}\}$

Linear model: $f : \mathbb{R}^{\mathsf{D}} \rightarrow \mathbb{R}$, with $f(\boldsymbol{x}) = w_0 + \sum_{d=1}^{D} w_d x_d = w_0 + \boldsymbol{w}^\top \boldsymbol{x}$ (superscript ^{T} stands for transpose), i.e. a *hyper-plane* parametrized by

- $\boldsymbol{w} = [w_1 \; w_2 \; \cdots \; w_{\mathsf{D}}]^\top$ (weights, weight vector, parameter vector, etc)
- bias w_0

NOTE: for notation convenience, very often we

- append 1 to each x as the first feature: $\tilde{\boldsymbol{x}} = [1 \; x_1 \; x_2 \; \ldots \; x_{\mathsf{D}}]^\top$
- let $\tilde{\boldsymbol{w}} = [w_0 \; w_1 \; w_2 \; \cdots \; w_{\mathsf{D}}]^\top$, a concise representation of all $D+1$ parameters. For ease of notation, we will drop the $\tilde{(\cdot)}$ and use D to subsume the constant term.
- the model becomes simply $f(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x}$ (but don't forget that the bias term is in there!)

Goal

Minimize total squared error (note that $\boldsymbol{x}_n^\top \boldsymbol{w} = \boldsymbol{w}^\top \boldsymbol{x}_n$)

- **Residual Sum of Squares** (RSS), a function of $\boldsymbol{w}$

$$\mathrm{RSS}(\boldsymbol{w}) = \sum_n (f(\boldsymbol{x}_n) - y_n)^2 = \sum_n (\boldsymbol{x}_n^\top \boldsymbol{w} - y_n)^2 = \|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|_2^2$$

where

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{x}_1^\top \\ \boldsymbol{x}_2^\top \\ \vdots \\ \boldsymbol{x}_\mathsf{N}^\top \end{pmatrix} \in \mathbb{R}^{\mathsf{N} \times D}, \quad \boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_\mathsf{N} \end{pmatrix} \in \mathbb{R}^\mathsf{N}$$

Goal

Minimize total squared error (note that $\boldsymbol{x}_n^\top \boldsymbol{w} = \boldsymbol{w}^\top \boldsymbol{x}_n$)

- **Residual Sum of Squares** (RSS), a function of $\boldsymbol{w}$

$$\mathrm{RSS}(\boldsymbol{w}) = \sum_n (f(\boldsymbol{x}_n) - y_n)^2 = \sum_n (\boldsymbol{x}_n^\top \boldsymbol{w} - y_n)^2 = \|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|_2^2$$

- find $\boldsymbol{w}^* = \operatorname*{argmin}_{\boldsymbol{w} \in \mathbb{R}^{\mathsf{D}}} \mathrm{RSS}(\boldsymbol{w})$, i.e. **least (mean) squares solution** (more generally called **empirical risk minimizer**)

- in principle can apply any optimization algorithm, but linear regression admits a *closed-form solution*

General least square solution

Objective

$$\mathrm{RSS}(\boldsymbol{w}) = \|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|_2^2$$

$$\begin{aligned}
\mathrm{RSS}(\boldsymbol{w}) &= \sum_n (\boldsymbol{x}_n^\top \boldsymbol{w} - y_n)^2 = \|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|_2^2 \\
&= (\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y})^\top (\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}) \\
&= \boldsymbol{w}^\top \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}^\top \boldsymbol{X}\boldsymbol{w} - \boldsymbol{w}^\top \boldsymbol{X}^\top \boldsymbol{y} + \boldsymbol{y}^\top \boldsymbol{y}
\end{aligned}$$

General least square solution

Find stationary points (Matrix Calculus)

$$\begin{aligned}\nabla \mathrm{RSS}(\boldsymbol{w}) &= \frac{\partial}{\partial \boldsymbol{w}} \left(\boldsymbol{w}^\top \boldsymbol{X}^\top \boldsymbol{X} \boldsymbol{w} - \boldsymbol{y}^\top \boldsymbol{X} \boldsymbol{w} - \boldsymbol{w}^\top \boldsymbol{X}^\top \boldsymbol{y} + \boldsymbol{y}^\top \boldsymbol{y} \right) \\ &= \left(\boldsymbol{X}^\top \boldsymbol{X} \boldsymbol{w} + \left(\boldsymbol{w}^\top \boldsymbol{X}^\top \boldsymbol{X} \right)^\top \right) - (\boldsymbol{y}^\top \boldsymbol{X})^\top - \boldsymbol{X}^\top \boldsymbol{y} + 0 \\ &= 2\boldsymbol{X}^\top \boldsymbol{X} \boldsymbol{w} - 2\boldsymbol{X}^\top \boldsymbol{y}\end{aligned}$$

Setting the gradient to zero

$$\begin{aligned}2\boldsymbol{X}^\top \boldsymbol{X} \boldsymbol{w} - 2\boldsymbol{X}^\top \boldsymbol{y} = 0 \implies \boldsymbol{X}^\top \boldsymbol{X} \boldsymbol{w} &= \boldsymbol{X}^\top \boldsymbol{y} \\ \boldsymbol{w}^* &= \left(\boldsymbol{X}^\top \boldsymbol{X} \right)^{-1} \boldsymbol{X}^\top \boldsymbol{y}\end{aligned}$$

assuming $\boldsymbol{X}^\top \boldsymbol{X}$ is invertible for now. By convexity $\boldsymbol{w}^*$ is the minimizer of RSS.

Computational complexity

Bottleneck of computing

$$\boldsymbol{w}^* = \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{y}$$

is to invert the matrix $\boldsymbol{X}^\top \boldsymbol{X} \in \mathbb{R}^{\mathsf{D} \times \mathsf{D}}$

- aka *pseudo-inverse** denoted by $(\cdot)^\dagger$, i.e. $\boldsymbol{X}^\dagger = \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top$
- naively need $O(\mathsf{D}^3)$ time
- there are many faster approaches

*see `https://en.wikipedia.org/wiki/Moore-Penrose_inverse`

What if $\boldsymbol{X}^\top \boldsymbol{X}$ is not invertible

What does that imply?

Recall $\left(\boldsymbol{X}^\top \boldsymbol{X}\right) \boldsymbol{w}^* = \boldsymbol{X}^\top \boldsymbol{y}$. If $\boldsymbol{X}^\top \boldsymbol{X}$ not invertible, this equation aka *Normal Equations* has

- infinitely many solutions ($\Rightarrow$ infinitely many minimizers)
- This is because *Normal Equations* are always *consistent*[†] meaning a solution *always* exists! It may not be unique though.

[†] See `https://sites.math.washington.edu/~burke/crs/308/LeastSquares.pdf`

How to resolve this issue?

Intuition: what does inverting $\boldsymbol{X}^\top \boldsymbol{X}$ do?

eigendecomposition: $$\boldsymbol{X}^\top \boldsymbol{X} = \boldsymbol{U}^\top \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & \cdots & \lambda_{\mathsf{D}} \end{bmatrix} \boldsymbol{U}$$

where $\lambda_1 \geq \lambda_2 \geq \cdots \lambda_{\mathsf{D}+1} \geq 0$ are **eigenvalues**.

inverse: $$(\boldsymbol{X}^\top \boldsymbol{X})^{-1} = \boldsymbol{U}^\top \begin{bmatrix} \frac{1}{\lambda_1} & 0 & \cdots & 0 \\ 0 & \frac{1}{\lambda_2} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & \cdots & \frac{1}{\lambda_{\mathsf{D}}} \end{bmatrix} \boldsymbol{U}$$

i.e. just inverse of the eigenvalues

How to solve this problem?

Non-invertible $\Rightarrow$ some eigenvalues are 0.

One natural fix: add something positive

$$\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I} = \boldsymbol{U}^\top \begin{bmatrix} \lambda_1 + \lambda & 0 & \cdots & 0 \\ 0 & \lambda_2 + \lambda & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & \cdots & \lambda_{\mathsf{D}} + \lambda \end{bmatrix} \boldsymbol{U}$$

where $\lambda > 0$ and $\boldsymbol{I}$ is the identity matrix. Now it is invertible:

$$(\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I})^{-1} = \boldsymbol{U}^\top \begin{bmatrix} \frac{1}{\lambda_1 + \lambda} & 0 & \cdots & 0 \\ 0 & \frac{1}{\lambda_2 + \lambda} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & \cdots & \frac{1}{\lambda_{\mathsf{D}} + \lambda} \end{bmatrix} \boldsymbol{U}$$

Fix the problem

The solution becomes

$$\boldsymbol{w}^* = \left(\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{y}$$

- not a minimizer of the original RSS

This in fact comes from minimizing **regularized** RSS aka **Ridge Regression**!

$$\min_{\boldsymbol{w}} \|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|_2^2 + \lambda \|\boldsymbol{w}\|_2^2$$

λ is a *hyper-parameter*, can be tuned by cross-validation.

Outline

Reading

PiML1

- Chapter 11 – 11.1, 11.2 (11.2.1, 11.2.2), 11.3 (11.3.1)