CSCE 633 - Machine Learning

Lecture 4

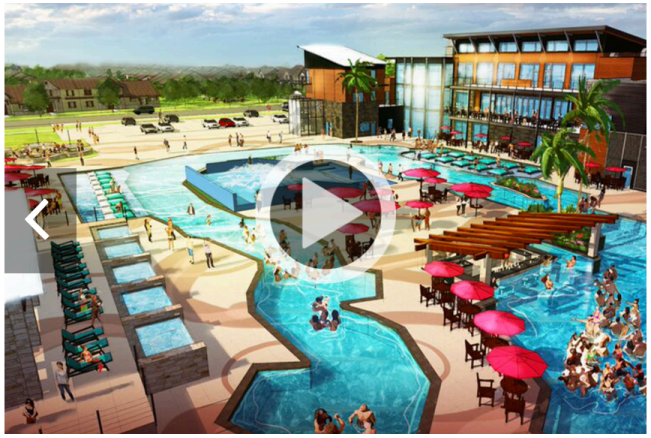
Overview

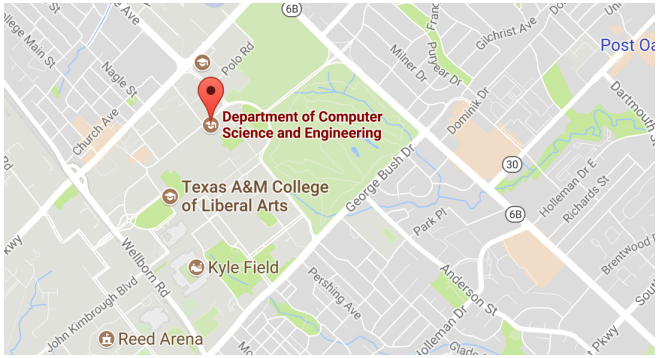
- Linear Regression: Basics
  - Example
  - Representation
  - Analytic Solution: Ordinary Least Squares (OLS) solution
- Linear Regression: Numerical Solution
  - General gradient descent
  - Gradient descent with linear regression (batch, stochastic, minibatch)
- Non-linear basis function for regression

Overview

- Linear Regression: Basics
  - Example
  - Representation
  - Analytic Solution: Ordinary Least Squares (OLS) solution
- Linear Regression: Numerical Solution
  - General gradient descent
  - Gradient descent with linear regression (batch, stochastic, minibatch)
- Non-linear basis function for regression

Linear Regression: Example





Apartment Amenities

Pet Policy

Dogs Allowed: Breed Restrictions May Apply,

- $250 Fee
- 2 Pet Limit

Cats Allowed

- $250 Fee
- 2 Pet Limit

Fitness & Recreation

- Fitness Center
- Pool
- Volleyball Court

Living Space

- Carpet
- Attic
- Crown Molding
- Views
- Walk-In Closets

Outdoor Space

- Balcony
- Patio
- Porch
- Yard
- Barbecue Area
- Barbecue/Grill

Parking

Surface Lot
6 spaces; Assigned Parking.

Services

- Maintenance on site
- Property Manager on Site
- Bilingual
- Courtesy Patrol
- Trash Pickup - Curbside
- Recycling
- Planned Social Activities
- Pet Play Area

Kitchen

- Dishwasher
- Disposal
- Ice Maker
- Granite Countertops
- Kitchen
- Microwave
- Oven
- Refrigerator
- Freezer
- Instant Hot Water

Property Information

- Built in 2015
- 442 Units/2 Stories

Lease Length

August 1 - July 22

Outdoor Space

- Grill
- Waterfront
- Pond

Features

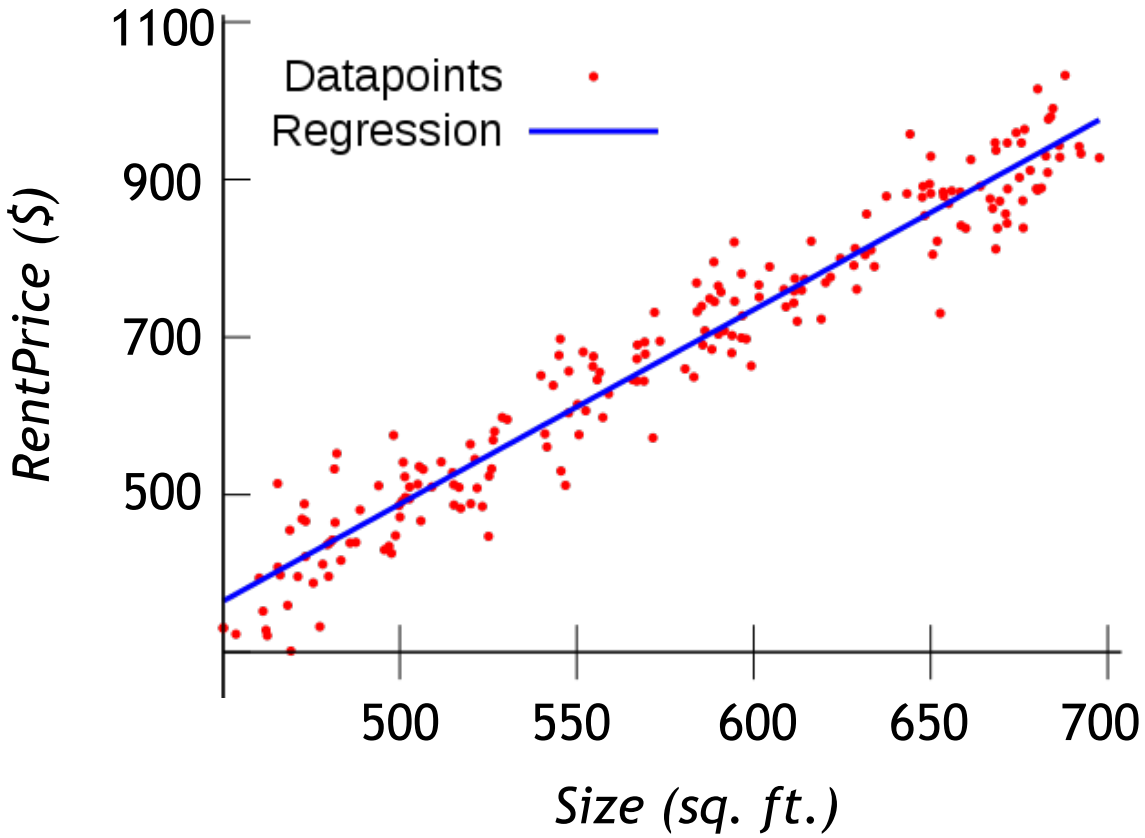
- High Speed Internet Access
- Washer/Dryer
- Air Conditioning
- Heating
- Ceiling Fans
- Smoke Free
- Cable Ready
- Tub/Shower
- Framed Mirrors

Source: apartments.com

$$\text{RentPrice} = w_0 + w_1 \times \text{Size} + w_2 \times \text{DistanceFromCS} + \ldots$$

Linear Regression: Training Data

$$\text{RentPrice} = w_0 + w_1 \times \text{Size} + w_2 \times \text{DistanceFromCS} + \ldots$$



Supervised Learning: Regression

Predict the rent price of an apartment

- Input **x**: apartment attributes (e.g. size, neighborhood, etc.)
- Output y: rent price of apartment
- Model parameters **w**

parametric model

- Linear model

$$y = f(\mathbf{x}|\mathbf{w}) = \mathbf{w}^T\mathbf{x}$$

- Non-linear model (φ: non-linear function)

$$y = f(\mathbf{x}|\mathbf{w}) = \mathbf{w}^T\boldsymbol{\phi}(\mathbf{x})$$

Linear Regression: Training Data

$$\text{RentPrice} = w_0 + w_1 \times \text{Size} + w_2 \times \text{DistanceFromCS} + \ldots$$

How do we find the unknown model parameters $\{w_0, w_1, w_2, \ldots\}$?

We use training data!

Training Sample	Size (sq.ft.)	DistanceFromCS (miles)	RentPrice ($)
1	498	11.9	675
2	513	8.6	750
3	621	8.3	800
4	710	3.4	965
…	…	…	…

The three components of learning

Representation	Evaluation	Optimization
Instances	Accuracy/Error rate	Combinatorial optimization
K-nearest neighbor	Precision and recall	Greedy search
Support vector machines	Squared error	Beam search
Hyperplanes	Likelihood	Branch-and-bound
Naive Bayes	Posterior probability	Continuous optimization
Logistic regression	Information gain	Unconstrained
Decision trees	K-L divergence	Gradient descent
Sets of rules	Cost/Utility	Conjugate gradient
Propositional rules	Margin	Quasi-Newton methods
Logic programs		Constrained
Neural networks		Linear programming
Graphical models		Quadratic programming
Bayesian networks		
Conditional random fields		

a learner must be represented in some formal language

an evaluation function assesses the performance of a learner

find the highest-scoring learner

Source: P. Domingos, 2014

Linear Regression: Representation

- **Input**: $\mathbf{x} \in \mathbb{R}^D$ (covariates, predictors, features, etc.)
- **Output**: $y \in \mathbb{R}$ (responses, targets, outcomes, etc.)
- **Training Data**: $\mathcal{D} = \{(\mathbf{x_1}, y_1), \ldots, (\mathbf{x_N}, y_N)\}$
- **Model**: $f : \mathbf{x} \rightarrow y \ , \quad f(\mathbf{x}) = w_0 + \sum_{n=1}^{D} w_d x_d = \mathbf{w}^T \mathbf{x}$

  w_0 (bias term)

  $\mathbf{w} = [w_0 \ w_1 \ \ldots \ w_D]^T$ (parameters)

Linear Regression: Evaluation

- A reasonable thing would be to **minimize prediction error** (also called the residual sum of squares)

$$RSS(\mathbf{w}) = \sum_{n=1}^{N} (y_n - f(\mathbf{x_n}))^2 = \sum_{n=1}^{N} \left[y_n - \left(w_0 + \sum_{d=1}^{D} w_d x_{nd} \right) \right]^2$$

x_{nd} is the d^{th} feature of the n^{th} training sample

N training samples, D features

- An equivalent vector expression is

$$RSS(\mathbf{w}) = (\mathbf{y} - \mathbf{X}\mathbf{w})^T (\mathbf{y} - \mathbf{X}\mathbf{w})$$

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & \dots & x_{1D} \\ \vdots & & & \vdots \\ 1 & x_{N1} & \dots & x_{ND} \end{bmatrix} = \begin{bmatrix} - \tilde{\mathbf{x}}_1^T - \\ \vdots \\ - \tilde{\mathbf{x}}_N^T - \end{bmatrix}$$

$$\mathbf{y} = [y_1, \dots, y_N]^T, \quad \tilde{\mathbf{x}}_\mathbf{n} = [1 \quad \mathbf{x}_\mathbf{n}^T]^T, \quad \mathbf{w} = [w_0 \; w_1 \; \dots \; w_D]^T$$

Linear Regression: Optimization

- We can find an analytical solution

$$\frac{\vartheta RSS(\mathbf{w})}{\vartheta \mathbf{w}} = 0 \quad \Rightarrow \quad \mathbf{w}^* = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

Ordinary least squares solution

(see next slide for derivation)

Linear Regression: Optimization

- Ordinary Least squares solution derivation

$$
\begin{aligned}
RSS(\mathbf{w}) &= (\mathbf{y} - \mathbf{X}\mathbf{w})^T(\mathbf{y} - \mathbf{X}\mathbf{w}) \\
&= \mathbf{y}^T\mathbf{y} - 2(\mathbf{X}\mathbf{w})^T\mathbf{y} + (\mathbf{X}\mathbf{w})^T(\mathbf{X}\mathbf{w}) \\
&= \mathbf{y}^T\mathbf{y} - 2\mathbf{w}^T(\mathbf{X}^T\mathbf{y}) + \mathbf{w}^T(\mathbf{X}^T\mathbf{X})\mathbf{w}
\end{aligned}
$$

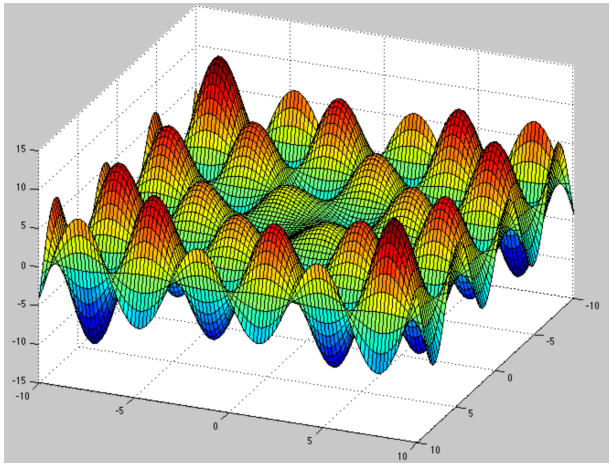
$$
\frac{\vartheta RSS(\mathbf{w})}{\vartheta \mathbf{w}} = 0 \Rightarrow -2\mathbf{X}^T\mathbf{y} + 2\mathbf{X}^T\mathbf{X}\mathbf{w} = 0
$$

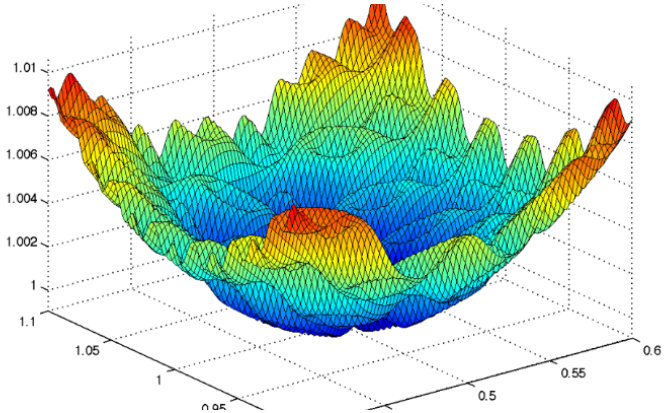
$$
\Rightarrow \mathbf{w} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}
$$

We cannot multiply both sides with $(X^T)^{-1}$ because it is not guaranteed that X^T is invertible.

Linear Regression: Optimization

- Why should convexity be a problem in optimization?





Loss functions might have more than one local optima (minima or maxima)

Convexity of optimization criterion

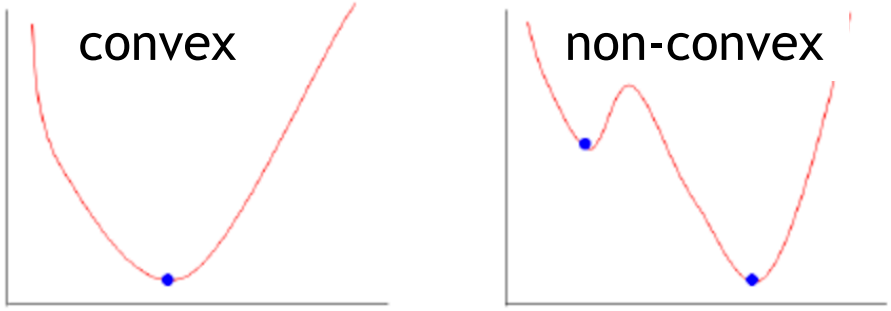
Theorem

Consider an optimization problem

$$\min f(\mathbf{x}) \;\; \text{s.t.} \;\; \mathbf{x} \in \Omega$$

where f is a convex function and Ω is a convex set.

Then any **local** minimum is also a **global** minimum.



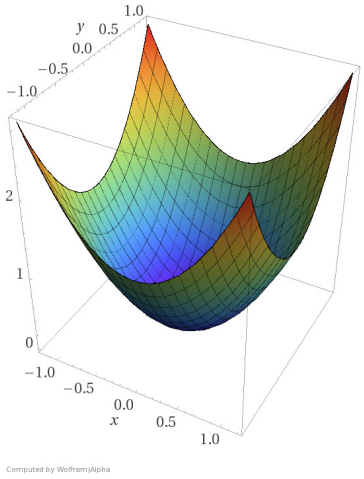
How do we find in f is convex?

Convexity of optimization criterion

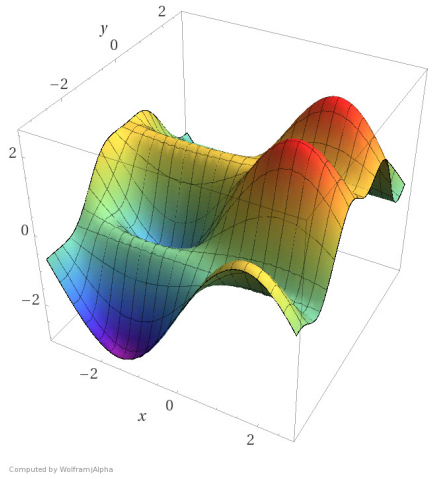
The second-derivative test

If the Hessian matrix $\mathbf{H_f}$ of f is positive semi-definite, then f is convex.

i.e. $\mathbf{u}^T \mathbf{H_f} \mathbf{u} \geq 0 \ , \quad \forall \mathbf{u}$



convex



non-convex

Convexity of optimization criterion

- The residual sum of squares $RSS(\mathbf{w})$ is a convex function wrt $\mathbf{w}$

$$RSS(\mathbf{w}) = (\mathbf{y} - \mathbf{X}\mathbf{w})^T(\mathbf{y} - \mathbf{X}\mathbf{w})$$

$$\frac{\vartheta RSS(\mathbf{w})}{\vartheta \mathbf{w}} = 2\mathbf{X}^T\mathbf{X}\mathbf{w} - 2\mathbf{X}^T\mathbf{y}$$

$$\frac{\vartheta^2 RSS(\mathbf{w})}{\vartheta \mathbf{w}^2} = 2\mathbf{X}^T\mathbf{X}$$

- For every $\mathbf{u} \in \mathbb{R}^D$ we have:

$$\mathbf{u}^T\mathbf{H}_{J(\mathbf{w})}\mathbf{u} = 2\mathbf{u}^T\mathbf{X}^T\mathbf{X}\mathbf{u} = 2(\mathbf{X}\mathbf{u})^T(\mathbf{X}\mathbf{u}) = 2\|\mathbf{X}\mathbf{u}\|_2^2 \geq 0$$

- Therefore the solution $\mathbf{w}^*$ is a global minimum of the error function.

Convexity of optimization criterion

Question: Assume the following non-linear regression model. Which if the following is true?

$$y = w_0 + w_1 x + w_2 x^2$$

$$\mathcal{D}^{train} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$$

$$RSS(w_0, w_1, w_2) = \sum_{n=1}^{N} \left(y_n - \left(w_0 + w_1 x_n + w_2 x_n^2\right)\right)^2$$

A) We don't know if RSS has a global minimum with respect to $[w_0, w_1, w_2]^T$

B) RSS has a single local minimum w.r.t. $[w_0, w_1, w_2]^T$, which is also global

C) It depends on the training data whether RSS has a minimum

Convexity of optimization criterion

The correct answer is B.

$$y = \mathbf{w}^T \mathbf{z}$$

$$\mathbf{w} = [w_0 \;\; w_1 \;\; w_2]^T \;, \;\; \mathbf{z} = [1 \;\; x \;\; x^2]^T$$

$$RSS(w_0, w_1, w_2) = \sum_{n=1}^{N} \left(y_n - \left(w_0 + w_1 x_n + w_2 x_n^2\right)\right)^2$$

$$= \sum_{n=1}^{N} \left(y_n - \mathbf{w}^T \mathbf{z}\right)^2 = (\mathbf{y} - \mathbf{Z}\mathbf{w})^T (\mathbf{y} - \mathbf{Z}\mathbf{w})$$

RSS is a convex function w.r.t. **w**, because the only thing that has changed in the previous expression (slide) is the data matrix **Z** (instead of **X**)

Linear Regression: What have we learnt so far

- **Representation:** linear combination of features

$$f : \mathbf{x} \rightarrow y, \text{ with } f(\mathbf{x}) = w_0 + \mathbf{w}^T\mathbf{x}$$

- **Evaluation:** Minimizing the residual sum of squares

$$\max_{\mathbf{w}} RSS(\mathbf{w}), \quad RSS(\tilde{\mathbf{w}}) = (\mathbf{y} - \mathbf{X}\mathbf{w})^T(\mathbf{y} - \mathbf{X}\mathbf{w})$$

- **Optimization:** Ordinary least squares solution (LSS)

$$\frac{\vartheta RSS(\mathbf{w})}{\vartheta \mathbf{w}} = 0 \quad \Rightarrow \quad \mathbf{w}^* = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

- **Readings:** Alpaydin Ch 2, Abu-Mostafa Ch 3.2