Linear Regression

CSCI 447/547 MACHINE LEARNING

Outline

- Linear Models
- 1D Ordinary Least Squares (OLS)
- Solution of OLS
- Interpretation
- Anscombe's Quartet
- Multivariate OLS
- OLS Pros and Cons

Optional Reading

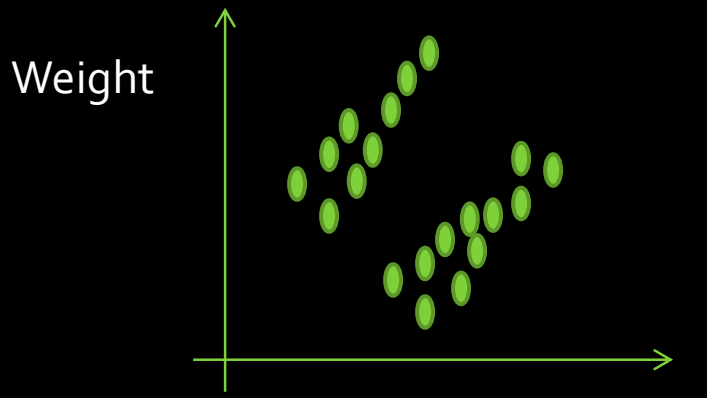
- K. Murphy, Machine Learning: A Probabilistic Perspective
- C. Bishop, Pattern Recognition and Machine Learning
- T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction (available free online)

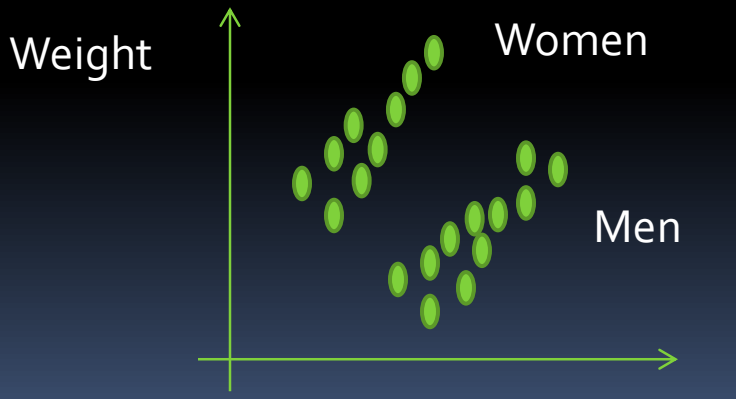
Terminology

- Features (Covariates or predictors)
- Labels (Variates or targets)
- Regression
- Classification

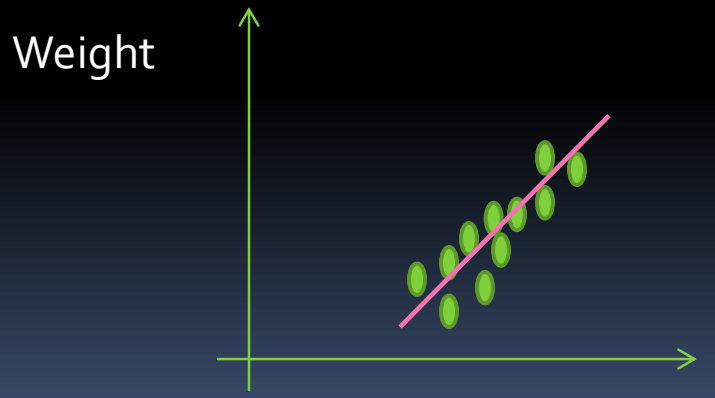
Types of Machine Learning

- Unsupervised
  - Finding structure in data
- Supervised
  - Predict from given data





Classification
categorical output data
Logistic Regression



OLS Regression
(Prediction)
continuous output data

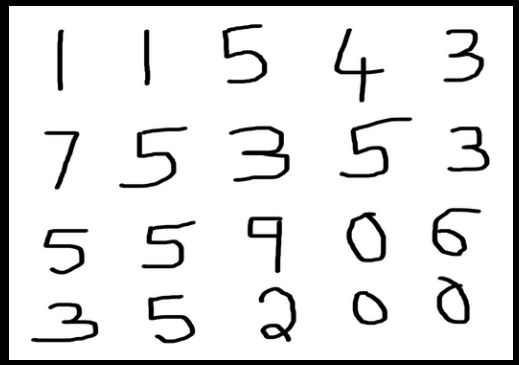
What is a Linear Model?

- Predict Housing Prices
  - Depends on:
    - Area
    - # of bedrooms
    - # of bathrooms
  - Hypothesis is that relationship is linear
    - Price = k_1(Area) + k_2(#bed) + k_3(#bath)
    - $y_i = a_0 + a_1x_1 + a_2x_2 + \ldots$

Why Use Linear Models?

- Interpretable
  - Relationships are easy to see
- Low Complexity
  - Prevents overfitting
- Scalable
  - Scale up to more data, larger problems
- Baseline
  - Can benchmark other methods against them

Examples of Use



- Example of Use
  - MNIST dataset – handwritten digits
  - Best performance – neural networks and regularization
    - 99.79% accurate
    - Takes about a day to train
    - More difficult to build
  - Logistic Regression
    - 92.5% accurate
    - Takes seconds to train
    - Can be built with less expertise
- Building Blocks of Later Techniques

Optional Reading

- Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. "Gradient-based learning applied to document recognition." Proceedings of the IEEE, 86(11):2278-2324, November 1998. Dataset available from http://yann.lecun.com/exdb/mnist/

Definition of 1-Dimension OLS

- The Problem Statement
  - i is an observation, we have N of them
  - i = 1…N
  - x is the independent variable (feature)
  - y is dependent variable (output variable)
  - y = ax + b, a,b are constants
  - $\hat{y}_i = ax_i + b$ OR $y_i = ax_i + b + \varepsilon$
  - Two unknowns – want to solve for a and b

The Loss Function

- $L = \sum_{i=1}^{N}(y_i - \hat{y}_i)^2$
- Goal is to minimize this function

- Using $\hat{y}_i = ax_i + b$, the equation becomes:

- $L = \sum_{i=1}^{N}(y_i - ax_i - b)^2$
- So this is the equation we want to minimize

Solution of OLS

- Derivation
  - $L = \sum_{i=1}^{N}(y_i - ax_i - b)^2$
- Want to minimize L
- Take derivative of loss function wrt each variable
  - $\frac{dL}{da} = 0, \ \frac{dL}{db} = 0$
  - $\frac{dL}{da} = 0 \Rightarrow \frac{dL}{da} = \sum_{i=1}^{N} 2(y_i - ax_i - b)(-x_i) = 0$
  - $\Rightarrow \frac{dL}{da} = \sum_{i=1}^{N} x_i y_i - a\sum_{i=1}^{N} x_i^2 - b\sum_{i=1}^{N} x_i = 0$

Solution of OLS

- Derivation
  - $\frac{dL}{db} = 0 \Rightarrow \frac{dL}{db} = \sum_{i=1}^{N} 2(y_i - ax_i - b)(+1) = 0$
  - $\Rightarrow \frac{dL}{db} = \sum_{i=1}^{N} y_i - \sum_{i=1}^{N} x_i - bN = 0$
  - $b = \frac{1}{N}\sum_{i=1}^{N} y_i - \frac{a}{N}\sum_{i=1}^{N} x_i$
  - This is the closed form solution for b

Solution of OLS

- Derivation
  - From first set,
  - $\frac{dL}{da} = \sum_{i=1}^{N} x_i y_i - a\sum_{i=1}^{N} x_i^2 - b\sum_{i=1}^{N} x_i = 0$
  - $\Rightarrow \sum_{i=1}^{N} x_i y_i = a\sum_{i=1}^{N} x_i^2 + \sum_{i=1}^{N} x_i\left(\frac{1}{N}\sum_{i=1}^{N} y_i - \frac{a}{N}\sum_{i=1}^{N} x_i\right)$
  - $a = \frac{\sum_1^N x_i y_i - \frac{1}{N}\sum_1^N x_i y_i}{\sum_1^N x^2{}_i - \frac{1}{N}(\sum_1^N x_i)^2}$
  - This is the closed form solution for a

Solution of OLS

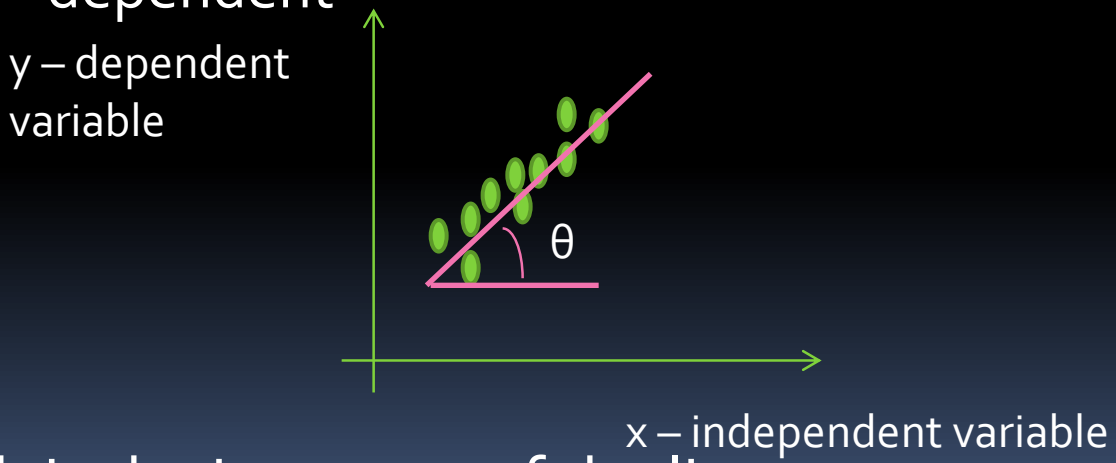
- Optimal Choices

$$a = \frac{\sum_i x_i y_i - \frac{1}{N}(\sum_i x_i)(\sum_i y_i)}{\sum_i x_i^2 - \frac{1}{N}(\sum_i x_i)^2} \qquad \left[= \frac{Cov(X,Y)}{Var(X)}\right]$$

$$b = \frac{1}{N}\sum_i y_i - \frac{a}{N}\sum_i x_i \qquad [= E[Y] - aE[X]]$$

Interpretation

- Interpretation of *a* and *b*
  - *a* is the slope of the line
    - tangent of angle θ
    - the effect of the independent variable on the dependent



  - b is the intercept of the line

Interpretation

- Interpretation of L
  - $L = \sum_{i=1}^{N}(y_i - y_i)^2$
  - Expresses how well the solution captures the variation in the data
  - $R^2 = 1 - MSE/Var(y)$
  - $R^2 \in [0, 1]$

Interpretation

Interpretation of a and b

a – slope of the line (size of relationship between X and Y)
b – intercept

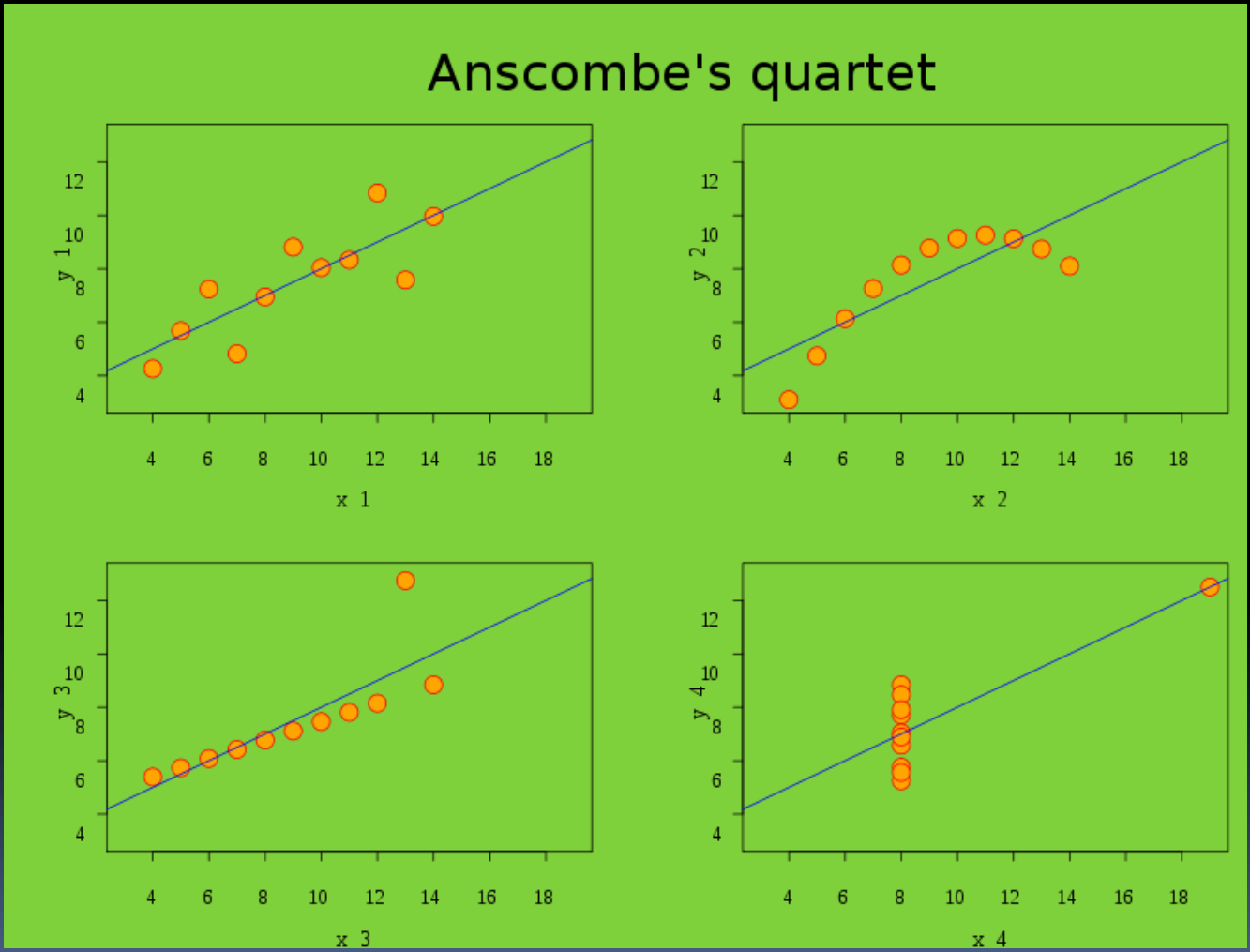
$$i.e.\ \ b = \begin{cases} \hat{y}, & x = 0 \\ 0, & x = \hat{x} \end{cases}$$

Interpretation of L

In general, $R^2 \in [0,1]$
Let's observe the model performance with the help of R^2 which is defined as. Keep in mind, L = loss function.

$$R^2 = 1 - \frac{Loss\ Function}{Var\ (y)} = 1 - \frac{MSE}{\bar{y}}$$

Anscombe’s Quartet



Anscombe's Quartet

- Same values for mean, variance and best fit line
- R^2 values are the same for each example
- But … linear regression may not be the best for the last three examples

Multivariable OLS

- Definition of Model
  - Data Matrix
  - The Loss Function

Mutivariable OLS

- i = an observation
- N = number of observations
- i = 1…N
- M = number of features
- $x_i = [x_{i1}, x_{i2}, \ldots, x_{iM}]$
- y_i - dependent variable
- Data matrix: $X = \begin{matrix} x_{11} & x_{12} \ldots & x_{1M} \\ \ldots & \ldots & \ldots \\ x_{N1} & X_{N2} \ldots & X_{NM} \end{matrix}$

Mutivariable OLS

- Data matrix: $X = \begin{matrix} x_{11} & x_{12} \dots & x_{1M} \\ \dots & \dots & \dots \\ x_{N1} & X_{N2} \dots & X_{NM} \end{matrix}$
- $y = ax + b(1)$

- Add a column of all 1's to left of data matrix to get bias term included
- $\hat{y}_i = B_0 + B_1 x_{i1} + B_2 x_{i2} + \dots + B_M x_{iM}$
- $\overline{x_i} \cdot B$, $B = \begin{matrix} B_0 \\ \dots, \\ B_M \end{matrix}$ $\overline{y} = XB$

Multivariable OLS

- Loss Function
  - $L = \sum_{i=1}^{N}(y_i - \hat{y}_i)^2$
- Still want to minimize L
  - $L = \sum_{i=1}^{N}(y_i - (B_0 + B_1 x_{i1} + \ldots + B_M x_{iM}))^2$
  - $L = \sum_{i=1}^{N}(y_i - \overline{x_i}\overline{B})^2$
  - Norm manner – L2 norm of the vector
  - $L = \|\overline{y} - XB\|_2^2$
  - $L = (y - XB)^T(y - XB)$

Optimization

- A Few Facts from Matrix Calculus

$$\frac{\partial[X\vec{\beta}]}{\partial\vec{\beta}} = X^T$$

$$\frac{\partial[\vec{\beta}^T A\vec{\beta}]}{\partial\vec{\beta}} = 2A^T\vec{\beta}$$

- $\frac{d(ax)}{dx} = a$
- $\frac{d(ax^2)}{dx} = 2ax$

Optimization

- Minimizing the Loss
  - $L = (y - XB)^T(y - XB)$
  - $\frac{dL}{dB} = 0$
    - $\frac{d(y - XB)^T(y - XB)}{dB} = 0$
    - $\frac{d(y^Ty - y^TXB - B^TX^Ty + B^TX^TXB)}{dB} = 0$ ($(XY)^T = Y^TX^T$)
    - $-(X^Ty) - (X^Ty) + 2(X^TX)B = 0$
    - $X^Ty = (X^TX)B$
    - $B = (X^TX)^{-1}X^Ty$

(assuming X^TX is invertible, which is true if X is a full rank matrix, that is none of its columns are linearly dependent)

OLS Pros and Cons

- OLS
  - Pros
    - Efficient to compute
    - Unique minimum
    - Stable under perturbation of data
    - Easy to interpret
  - Cons
    - Influenced by outliers
    - $(X^TX)^{-1}$ may not exist
      - Features may not be linearly independent

Summary

- Linear Models
- 1D Ordinary Least Squares (OLS)
- Solution of OLS
- Interpretation
- Anscombe's Quartet
- Multivariate OLS
- OLS Pros and Cons