CSE 634 - Data Mining Concepts and Techniques

Regression Analysis for Prediction

Professor Anita Wasilewska

Sources

- http://www3.cs.stonybrook.edu/~has/CSE545/Slides/6.11-11.pdf
- http://cs229.stanford.edu/notes/cs229-notes1.pdf
- http://www.mit.edu/~6.s085/notes/lecture3.pdf
- https://lagunita.stanford.edu/c4x/HumanitiesScience/StatLearning/asset/model_selection.pdf
- https://lagunita.stanford.edu/c4x/HumanitiesScience/StatLearning/asset/linear_regression.pdf
- http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5506092
- https://www.youtube.com/watch?v=djRh0Rkqygw
- http://vision.stanford.edu/teaching/cs231n/slides/2015/lecture3.pdf

Overview

- Introduction
- Motivation
- Linear Regression
- Least Sum of Squares
- Ridge Regression
- Lasso Regression
- Research Paper

‘In God we trust, all others must bring data.’
-W. Edwards Deming

Introduction

- A competitor car company comes up with a new model of a car.
- You have multiple theories on which variables will impact the sales - if it will jump or plummet?
- Six weeks later, the sales jump.
- **Regression Analysis** is a way of *mathematically* sorting out which of the variables does indeed have an impact!
- It has a set of **dependent variables** (which we want to predict) and a set of **predictor variables** (which we think might affect the dependent variables).

Motivation

- **We want answers** for correlation problems - how much does one variable affect the other.
- Managers want a method to predict sales, best time to hire people, best promotion strategies, etc.
- Helps aggregate the impact of many independent variables on the outcome of the dependent variable.
- In short - **Prediction is the Motivation**!

Linear Regression

- In Linear Regression, data are modeled to fit a straight line.
- It assumes the dependence of ***Y*** (dependent variable) on $X_1, X_2, \ldots X_p$ (predictor variables) is linear.
- Consider linear regression on advertising data:
- Model:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

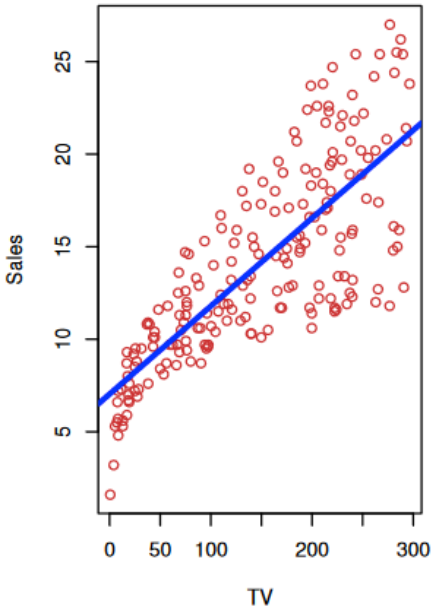


Image Source: https://lagunita.stanford.edu/c4x/HumanitiesScience/StatLearning/asset/linear_regression.pdf

Linear Regression

$$Y = \beta_0 + \beta_1 X + \epsilon$$

- Given some estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ for the model coefficients, we predict future values using:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

- $\hat{y}_i$ represents prediction of ***Y*** on the basis of ***X = x***

Multiple Linear Regression

- For multiple variables, we have:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_m X_{m1} + \epsilon_i$$

- Or in vector notation across all i:

$$Y = X\beta + \epsilon$$

- Weights for multiple linear regression can be calculated using:

$$\beta = (X_i^T X_i)^{-1} X_i^T y$$

Predicted and Residual Values

- Predicted, or fitted, values are values of y predicted by the least-squares regression line obtained by plugging in *x1, x2, ..., xn* into the estimated regression line

$$\hat{y}_1 = \hat{\beta}_0 - \hat{\beta}_1 x_1$$

$$\hat{y}_2 = \hat{\beta}_0 - \hat{\beta}_1 x_2$$

- Residuals are the deviations of observed and predicted values

$$e_1 = y_1 - \hat{y}_1$$

$$e_2 = y_2 - \hat{y}_2$$

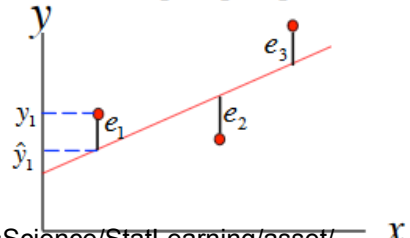


Image Source: https://lagunita.stanford.edu/c4x/HumanitiesScience/StatLearning/asset/linear_regression.pdf

Least Sum of Squares

- If $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ be the prediction of Y based on the ***ith*** value of X.
- Then $e_i = y_i - \hat{y}_i = \hat{y}_i = \hat{\beta}_0 - \hat{\beta}_1 x_i$ is the **ith residual**.
- The residual sum of squares is given by:

  ***RSS* = $e_1^2 + e_2^2 + \ldots + e_n^2$**

- Least Sum of Squares method chooses ***β'$_0$*** and ***β'$_1$*** to minimize the value of

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$ be calculated as:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Least Square Estimation via Gradient Descent

via Gradient Descent

Start with $\hat{\beta}_0 = \hat{\beta}_1 = 0$

Repeat until convergence:

Calculate all $\hat{Y}_i$

$$\hat{\beta}_0 = \hat{\beta}_0 - \alpha(\sum_{i=1}^{n} \hat{Y}_i - Y_i)$$

$$\hat{\beta}_1 = \hat{\beta}_1 - \alpha(\sum_{n} X_i(\hat{Y}_i - Y_i))$$

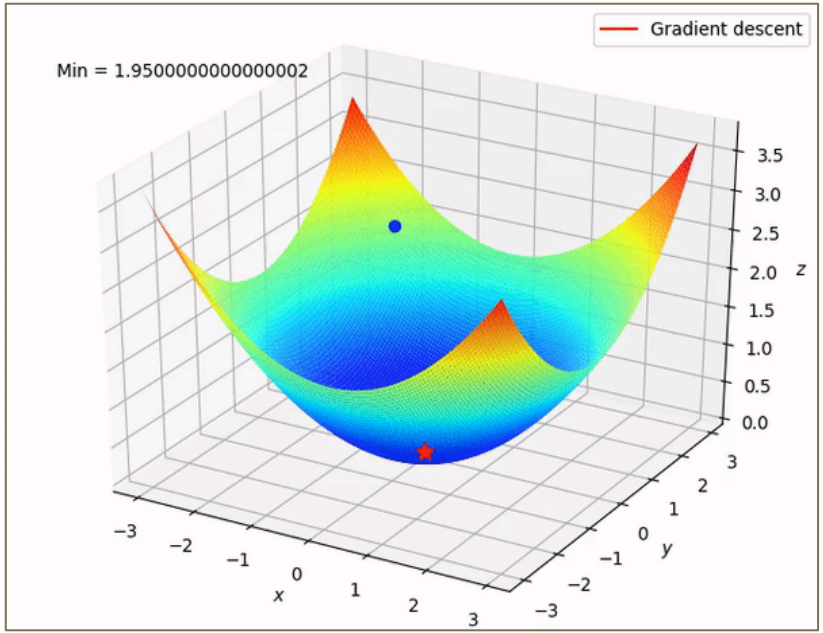


Image source: https://jed-ai.github.io/py1_gd_animation/

Multiple Linear Regression: Predict house price

Living area (feet^2)	#bedrooms	Price (1000$s)
2104	3	400
1600	3	330
2400	3	369
1416	2	232
3000	4	540
⋮	⋮	⋮

Image source: https://www.youtube.com/watch?v=bQI5uDxrFfA&list=PLLssT5z_DsK-h9vYZkQkYNWcItqhlRJLN

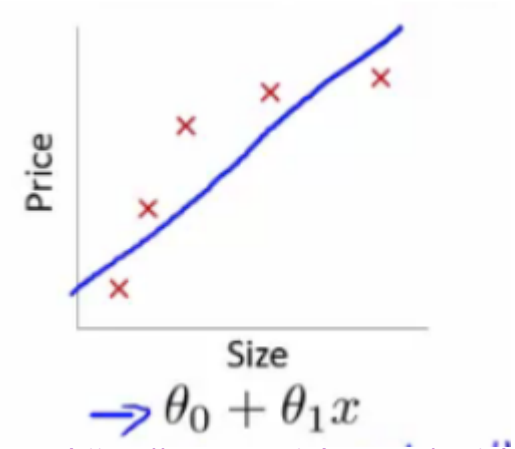
- To perform supervised learning, we decide to approximate y as a linear function

$$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2$$

- Here, the θi ’s are the parameters(also called weights) and Xi’s are feature

Underfitting

- Fit a linear function to the data - not a great model
  This is underfitting - also known as high bias



- Fit a quadratic function

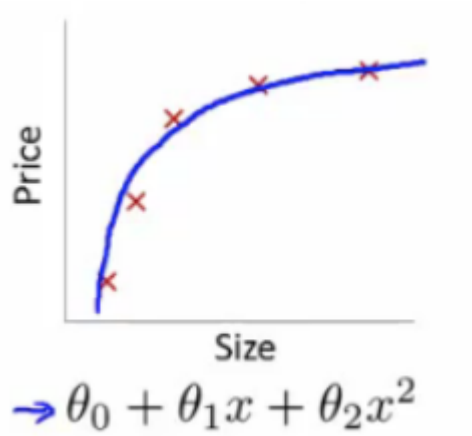


Image source: https://www.youtube.com/watch?v=bQI5uDxrFfA&list=PLLssT5z_DsK-h9vYZkQkYNWcItqhlRJLN

Overfitting

- Fit a 4th order polynomial. Now curve fits through all five examples. Seems to do a good job fitting the training set. But, despite fitting the data we've provided very well, this is actually not such a good model
- This is overfitting - also known as high variance
- Using too many features or a too complex model can often lead to overfitting.

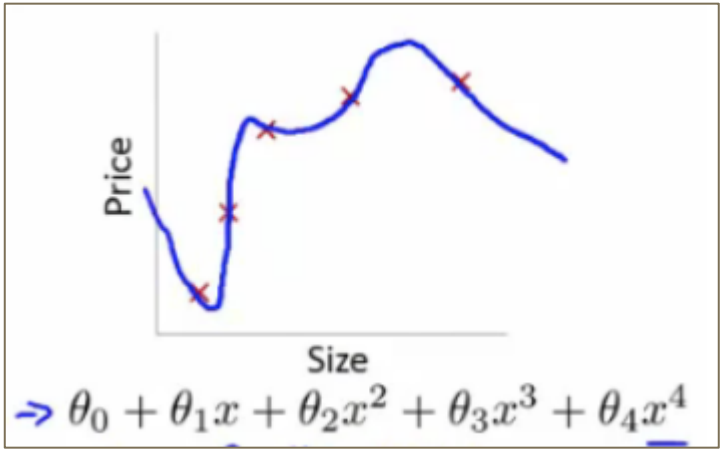


Image source: https://www.youtube.com/watch?v=bQI5uDxrFfA&list=PLLssT5z_DsK-h9vYZkQkYNWcItqhlRJLN

Advantages and Disadvantages of Linear Models

Advantages - Simplicity, interpretability, good predictive performance

Disadvantages-

Prediction Accuracy: especially when p(features/predictors) > n(no. of records), to control the variance.

If we have too many features then the learned hypothesis may give a cost function of exactly zero. But this tries too hard to fit the training set fails to provide a general solution i.e. unable to generalize (apply to new examples)

Alternative to Least Square Estimates

- **Subset Selection:** Identify a subset of predictors that we believe to be related to the response. We then fit a model using least squares on the reduced set of variables
- **Dimension Reduction:** We project the p predictors into a M-dimensional subspace, where M < p. Achieved by computing M different linear combinations, or projections, of the variables.
- **Shrinkage**: We fit a model involving all predictors, but the estimated coefficients are shrunken towards zero, relative to the least squares estimates. Also known as regularization, has the effect of reducing variance and can also perform variable selection.

Shrinkage Methods

- The subset selection methods use least squares to fit a linear model that contains a subset of the predictors.

- As an alternative, we can fit a model containing all predictors using a technique that constrains or regularizes the coefficient estimates, or equivalently, that shrinks the coefficient estimates towards zero.

- It may not be immediately obvious why such a constraint should improve the fit, but it turns out that shrinking the coefficient estimates can significantly reduce their variance

Ridge Regression

- Keep all features, but reduce magnitude of parameters θ
- Works well when we have a lot of features, each of which contributes a bit to predicting y
- Penalize and make some of the θ parameters really small e.g. here θ3 and θ4

$$\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2 + 1000\,\theta_3^2 + 1000\,\theta_4^2$$

Ridge Regression

- So here we end up with θ3 and θ4 being close to zero (because the constants are massive). So we're basically left with a quadratic function

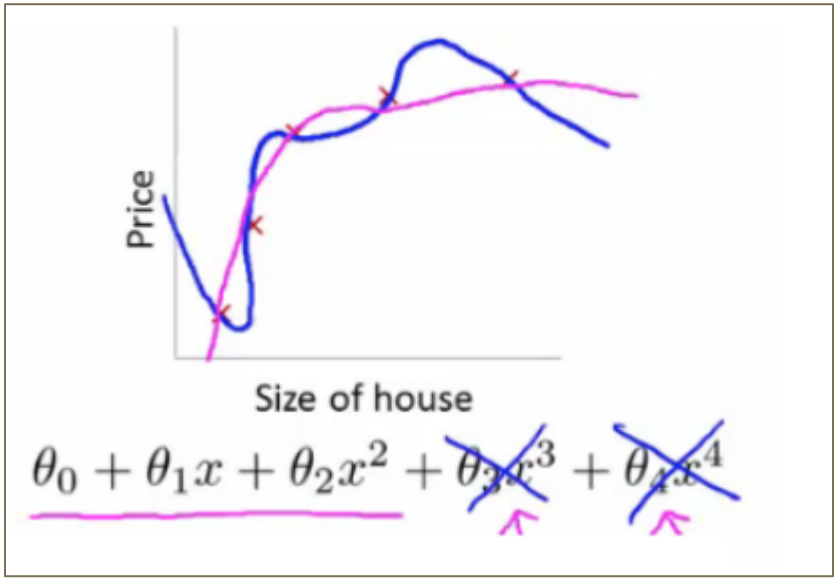


Image source: https://www.youtube.com/watch?v=bQI5uDxrFfA&list=PLLssT5z_DsK-h9vYZkQkYNWcItqhlRJLN

Ridge Regression

- Recall that the least squares fitting procedure estimates β0,β1,...,βp using the values that minimize

$$\mathrm{RSS} = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2$$

- In contrast, the ridge regression coefficient estimates β are the values that minimize

$$\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 = \mathrm{RSS} + \lambda \sum_{j=1}^{p} \beta_j^2,$$

Where λ ≥0 is a tuning parameter, to be determined separately.

Ridge Regression

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}\beta_j^2=\mathrm{RSS}+\lambda\sum_{j=1}^{p}\beta_j^2,$$

- As with least squares, ridge regression seeks coefficient estimates that fit the data well, by making the RSS small.
- However, the second term, λ ∑β^2 , called as shrinkage penalty, is small when B1,...,βp are close to zero, and so it has the effect of shrinking the estimates of Bj towards zero.
- Closed form solution for Ridge Regression:

$$\hat{\beta}^{ridge}=(X^TX+\lambda I)^{-1}X^Ty$$

Ridge Regression

- Advantages:
  - Helps reduce the effect of mutually correlated predictor variables

- Disadvantage:
  - All the predictor variables are present in the final model, thus ridge does not perform subset selection in case of unimportant attributes
    - Solution - LASSO

Lasso Regression

- LASSO - Least Attribute Selection and Shrinkage Operator
- Also known as L1 norm
- Goal: to obtain the subset of predictors that minimize the prediction error
- Shrinkage: Constraints on parameters that shrinks the coefficients towards zero
- Selection: Identifies the most important variables associated with the response variable

Lasso Regression

- The penalty in the below equation $\lambda \sum |\beta_j|$, has the effect of reducing some coefficients to zero

$$\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| = \text{RSS} + \lambda \sum_{j=1}^{p} |\beta_j|$$

Lasso Regression

- Useful when number of observation are small but the predictor variables are many
- Increases interpretability of the model by reducing the coefficients of unimportant variables to zero
- This makes the model sparser and reduces the possibility of the model overfitting

Tuning parameter λ

- Ridge Regression
  - If λ is very large we end up penalizing ALL the parameters (θ1, θ2 etc.) so all the parameters end up being close to zero If this happens, it's like we got rid of all the terms in the hypothesis. This results here is then underfitting

- LASSO
  - The tuning parameter λ - as it increases, more and more attribute coefficients are reduced to zero

- So how do we choose the appropriate λ? We use cross-validation

Lasso and Ridge: The graphical view

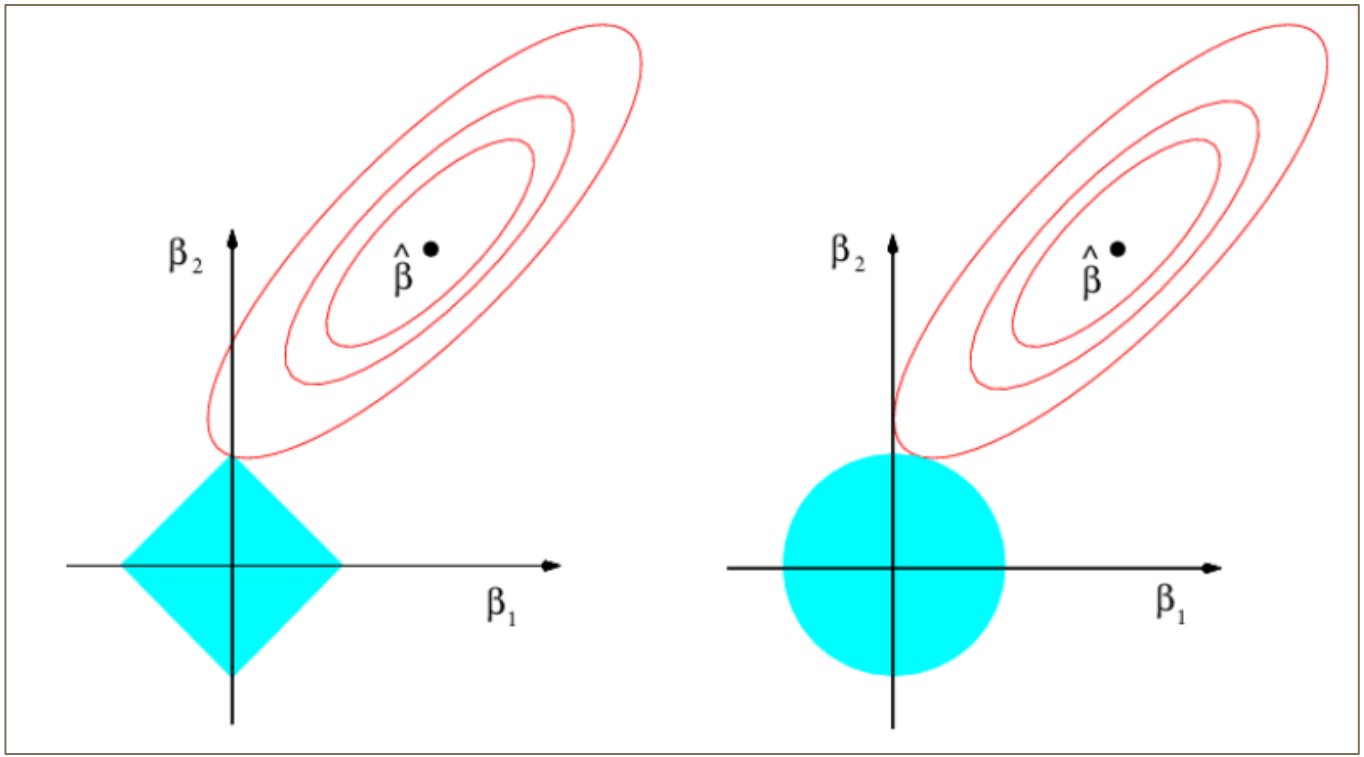


Image source: https://lagunita.stanford.edu/c4x/HumanitiesScience/StatLearning/asset/model_selection.pdf

PART 2

Linear Regression for Face Recognition

Imran Naseem, Roberto Togneri

IEEE Transaction on Pattern Analysis and Machine Intelligence

July 8th, 2010

Total Citations - 703

Introduction

1. Recognises face using features of the image like: RGB values, opacity, saturation, brightness etc.
2. Problem: High Dimensionality
3. Solution: Use Principal Component Analysis(PCA), Independent Component Analysis(ICA) and Linear Discriminant Analysis(LDA).

Linear Regression Classification (LRC)

1. Creates model for each person registered using his/her images.
2. Therefore, each model's weights are calculated using formula

$$\beta_j = (X_i^T X_i)^{-1} X_i^T y$$

3. After which, y is predicted for each β_j

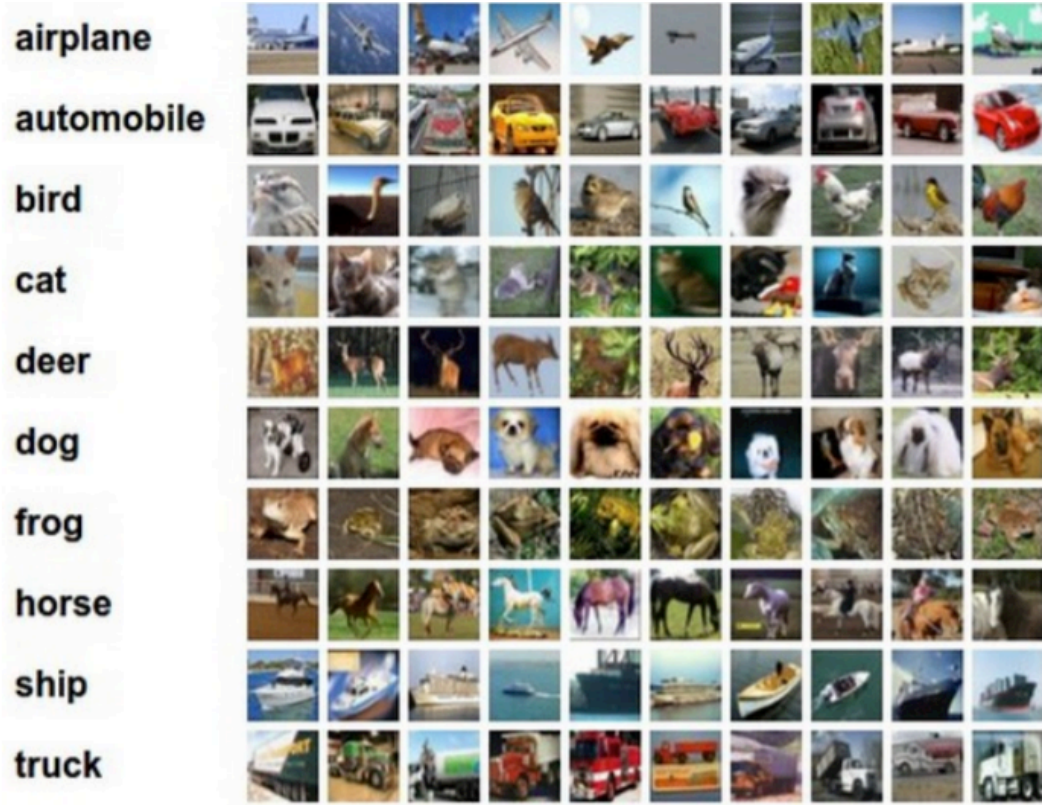
$$y_j = X_i \beta_j$$

Where *j* represents the different models and *i* represents different points

Linear Regression Classification (LRC)

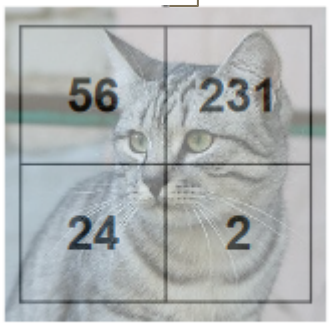
- The class with the highest value of y, gets predicted as the class of X_i
- Example

Linear Regression Classification (LRC)



Source: http://vision.stanford.edu/teaching/cs231n/slides/2015/lecture3.pdf Slide 3

Linear Regression Classification (LRC)



Input image

0.2	-0.5	0.1	2.0
1.5	1.3	2.1	0.2
0.1	0.25	0.2	-0.3

β

56
231
24
2

x_i

496.2	Cat Score
235.1	Dog Score
-96.8	Ship Score

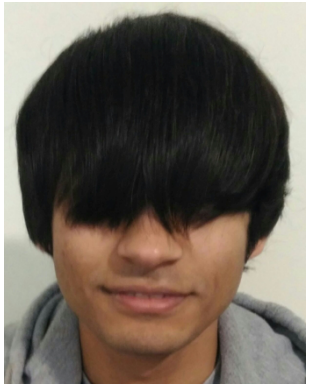
y

Predicted Class: **Cat**

Source: http://vision.stanford.edu/teaching/cs231n/slides/2015/lecture3.pdf Slide 8

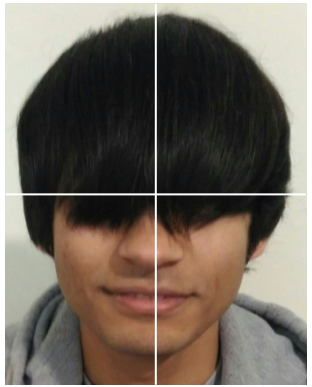
Occluded Images

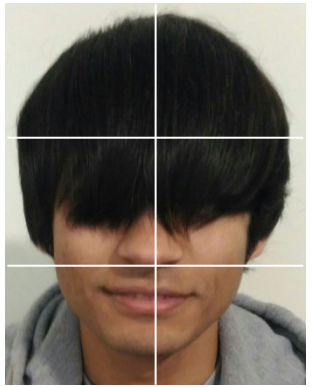


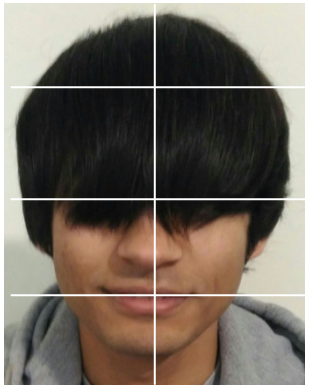



Occluded Images

- Break them into sections
- Class is decided by majority voting
- **Disadvantages:**
  - Gives equal weightage to clean and contaminated sections







D	J
D	D

Majority Vote: Class D

Thank You!