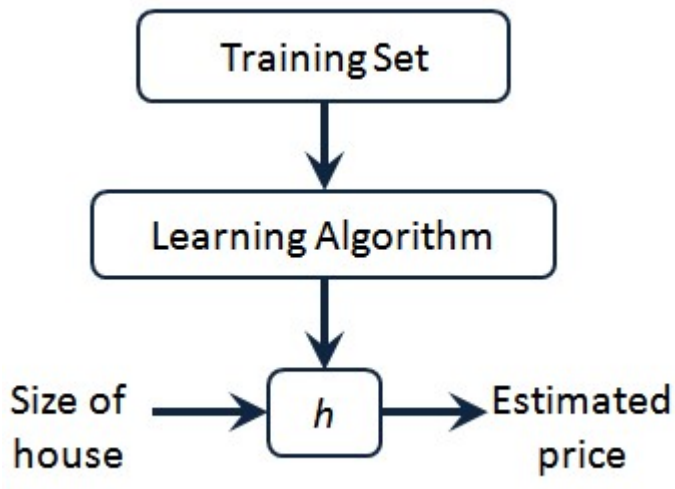
LinearRegressionHousePrice

Ryan

September 12, 2015

Remind of basic ML procedure



Input data/Training Set

```
houseData <- read.csv("Data.csv")
head(houseData)
```

```
##            Address  Price Sq.Feet Beds Baths
## 1 61 Hiawatha Ave. 409000    1293    3     1
## 2    37 Wilmot Rd. 579000    1872    3     2
## 3   25 Marlborough 489900    2040    3     1
## 4   18 Wildwood Ln 499000    1763    4     2
## 5     78 Lake St.  399500    1600    4     2
## 6 47 Pine Vale Rd. 457000    1582    2     2
```

Linear Regression Model (A Family of models, H)

- Independent variable expressed as a linear combination of dependent varibales
- $y = w_0 + w_1 x_1 + ... + w_k x_k$
- let $x_0 = 1$ and rewrite the above formula
- $y = w_0 x_0 + w_1 x_1 + ... + w_k x_k$
- that is $\mathbf{y} = \mathbf{w}^T \mathbf{x}$
- different $\mathbf{w} \implies$ different model(h) in the family(H)

How to determine good model or bad model?

- Commonly, want a model(h) that minimizes MSE of insample predictions
- $MSE = \frac{1}{n-k-1}\Sigma(y^{(i)} - \hat{y^{(i)}})^2$
- where $\hat{y^{(i)}} = \mathbf{w}^T x^{(i)}$
- So the problem, stated in optimization term is:
-

$$\min_{\mathbf{w}} MSE = \frac{1}{n-k-1}\Sigma(y^{(i)} - \mathbf{w}^T x^{(i)})^2 s.t. nothing \qquad really...$$

Sounds complicated....But

- in R, it is as simple as

```
h <- lm(formula = Price ~ Sq.Feet + Beds + Baths,
        data = houseData)
```

- and we got the h now.
- We will talk about how this is solved in futhure.

So... agin how good is the model?

- For now, let's only judge the model with the data we have in hand
- look for the Adjusted R-squared value

```
summary(h)
```

```
##
## Call:
## lm(formula = Price ~ Sq.Feet + Beds + Baths, data = houseData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -127476  -28611   -9346   35080  127963
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 301459.21   51306.06   5.876 3.39e-06 ***
## Sq.Feet         96.91      29.64   3.270  0.00303 **
## Beds         -2762.53   14209.51  -0.194  0.84736
## Baths         8487.69   20886.53   0.406  0.68780
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 56210 on 26 degrees of freedom
## Multiple R-squared:  0.3599, Adjusted R-squared:  0.2861
## F-statistic: 4.873 on 3 and 26 DF,  p-value: 0.008058
```

What else we can do for now?

```
h2 <- lm(formula = Price ~ Sq.Feet + Baths,
         data = houseData)
summary(h2)
```

```
##
## Call:
## lm(formula = Price ~ Sq.Feet + Baths, data = houseData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -124395  -29044   -7703   34082  127162
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 298543.11   48182.33   6.196 1.26e-06 ***
## Sq.Feet         95.21      27.80   3.424  0.00198 **
## Baths         6791.16   18635.03   0.364  0.71838
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 55200 on 27 degrees of freedom
## Multiple R-squared:  0.359,  Adjusted R-squared:  0.3115
## F-statistic:  7.56 on 2 and 27 DF,  p-value: 0.00247
```

What else we can do for now?

```
h3 <- lm(formula = Price ~ Sq.Feet,
         data = houseData)
summary(h3)
```

```
##
## Call:
## lm(formula = Price ~ Sq.Feet, data = houseData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -126676  -30245   -6865   31044  127468
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 303738.88   45306.31   6.704 2.81e-07 ***
## Sq.Feet         99.15      25.21   3.933 0.000503 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 54340 on 28 degrees of freedom
## Multiple R-squared:  0.3558, Adjusted R-squared:  0.3328
## F-statistic: 15.47 on 1 and 28 DF,  p-value: 0.000503
```

What else we can do for now?

```
h4 <- lm(formula = Price ~ Sq.Feet + Beds * Baths,
         data = houseData)
summary(h4)
```

```
##
## Call:
## lm(formula = Price ~ Sq.Feet + Beds * Baths, data = houseData)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -92030 -25823  -3804  21900 112417
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    179.19  117366.72   0.002  0.99879
## Sq.Feet         69.41      28.19   2.462  0.02104 *
## Beds        101344.61   39441.13   2.570  0.01653 *
## Baths       170722.13   61112.68   2.794  0.00986 **
## Beds:Baths  -45815.75   16439.26  -2.787  0.01001 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 50070 on 25 degrees of freedom
## Multiple R-squared:  0.5116, Adjusted R-squared:  0.4335
## F-statistic: 6.548 on 4 and 25 DF,  p-value: 0.000951
```

Try yourself, We can discuss next time

- How good can you get?
- And ... is this "good" guaranteed for outside data?

data <- houseData data$IT1 <- -data$Beds/data$Sq.Feetdata$IT2 <- data$Baths/data$Sq.Feet data$IT3 <- -data$Beds*data$Bathsdata$IT4 <- data$Beds * data$Baths/data$Sq.Feetdata$IT5 <- data$Sq.Feet^2data$Beds <- as.factor(as.character(data$Beds))data$Baths <- as.factor(as.character(data$Baths)) g <- lm(Price~.-Address-Baths,data) summary(g)