



CS 559: Linear Regression

Lecture 3

In Jang

ijang@stevens.edu



Outline:

- Probability Theory - Continue
- Regression - Supervised Learning
- Linear Regression
 - Example
 - Linear Regression Result Interpretation
 - Linear Regression Limitation



Prior vs. Posterior

- Prior Probability - If we had been asked which box had been chosen before being told the identity of the selected item of fruit, then the most complete information we have available is provided by the probability $P(B)$.
- Posterior Probability - Once we are told that the fruit is an orange, we can then use Bayes' theorem to compute the probability $P(B|F)$, which we shall call the posterior probability because it is the probability obtained after we have observed F.



Conditional Probability

- Conditional probability: Recalculated probability of event A after someone tells you that event B happened.

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$
$$P(A \cap B) = P(A|B)P(B)$$

- Bayes Rule:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$



Bayesian Probability

- Bayesian view: probabilities provide a quantification of uncertainty. Before observing the data, the assumptions about w are captured in the form of a prior probability distribution $P(w)$. The effect of the observed data $D = \{(x_1, y_1), \dots, (x_N, y_N)\}$ is expressed by $P(D|w)$.
- Bayes' theorem:

$$P(w|D) = \frac{P(D|w)P(w)}{P(D)}$$

- Bayes' theorem in words: posterior \propto likelihood \times prior



Continuous Random Variables

- A random variable X is continuous if its sample space X is uncountable.
- In this case, $P(X = x) = 0$ for each x .



Continuous Random Variables

- If $p_x(x)$ is a probability density function for X , then

$$P(a < X < b) = \int_a^b p(x)dx$$

$$P(a < X < a + dx) \approx p(a) \cdot dx$$



- The cumulative distribution function is $F_x(x) = P(X < x)$. We have that $p_x(x) = F'(x)$, and $F(x) = \int_{-\infty}^x p(s)ds$.
- More generally, If A is an event, then

$$P(A) = P(X \in A) = \int_{x \in A} p(x)dx$$
$$P(\Omega) = P(X \in \Omega) = \int_{x \in \Omega} p(x)dx = 1$$



Mean, Variance, and Conditionals

- Mean: $E(x) = \int_x x \cdot p(x)dx$
- Variance: $Var(X) = E(X^2) - E(X)^2$
- If X has pdf $p(x)$, then $X|(X \in A)$ has pdf

$$p_{(x|A)}(x) = \frac{p(x)}{P(A)} = \frac{p(x)}{\int_{x \in A} p(x)dx}$$

- Only makes sense if $P(A) > 0$!



Bivariate Continuous Distributions

- $p_{x,y}(x, y)$, joints probability density function of X and Y
- $\int_x \int_y p(x, y) dx dy = 1$
- Marginal distribution: $p_X(x) = \int_{-\infty}^{\infty} p(x, y) dy$
- Conditional distribution: $p(x|y) = \frac{p(x,y)}{p(y)}$
- Note: $P(Y = y) = 0!$
- Independence: X and Y are independent if $p_{(x,y)}(x, y) = p_x(x)p_y(y)$



The Univariate Gaussian

- Probability density function

$$p(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

- Easy to validate:

$$\int_{-\infty}^{\infty} p(x|\mu, \sigma^2) dx = 1$$

- Expectation

$$E(x) = \int_{-\infty}^{\infty} p(x|\mu, \sigma^2) x dx = \mu$$

- Variance

$$Var(x) = E(x^2) - E(x)^2 = \sigma^2$$



Products of Gaussian pdfs

- Suppose $p_1(x) = p\left(x, \mu_1, \frac{1}{\beta_1}\right)$ and $p_2(x) = p\left(x, \mu_2, \frac{1}{\beta_2}\right)$, then

$$\begin{aligned} p_1(x)p_2(x) &\propto p\left(x, \mu, \frac{1}{\beta}\right) \\ \beta &= \beta_1 + \beta_2 \\ \mu &= \frac{1}{\beta}(\beta_1\mu_1 + \beta_2\mu_2) \end{aligned}$$

- In general,

$$\begin{aligned} p_1(x)p_2(x) \dots p_n(x) &\propto p\left(x, \mu, \frac{1}{\beta}\right) \\ \beta &= \sum \beta_n \\ \mu &= \frac{1}{\beta} \sum_n \mu_n \beta_n \end{aligned}$$

- This is also true for multivariate Gaussians!



ML estimator

- Assuming data points are independent and identically distributed (i.i.d.), the probability of the data set given μ and σ^2 (the likelihood function):

$$P(\mathbf{x}|\mu, \sigma^2) = \prod_{n=1}^N p(x_n|\mu, \sigma^2)$$

- log-likelihood:

$$\log P(\mathbf{x}|\mu, \sigma^2) = -\frac{1}{2\sigma^2} \sum_{n=1}^N (x_n - \mu)^2 - \frac{N}{2} \log \sigma^2 - \frac{N}{2} \log(2\pi)$$

- Maximizing log-likelihood with respect to μ and σ^2 :

$$\mu_{ML} = \frac{1}{N} \sum_{n=1}^N x_n, \sigma_{ML}^2 = \frac{1}{N} \sum_{n=1}^N (x_n - \mu_{ML})^2$$



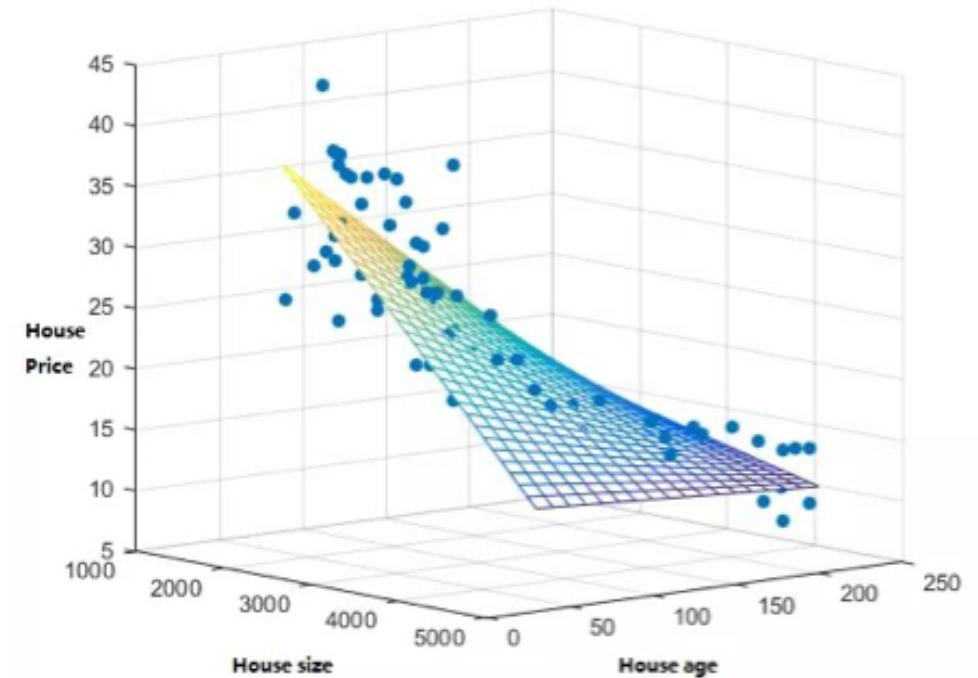
Linear Regression

Regression

- Regression: a ML technique to summarize relationships between continuous variables
 - One is regarded as response, outcome, dependent, or target variable
 - The rest is regarded as predictor(s), explanatory(ies), features(s) or independent variable(s)
- Simple linear regression
 - Simple: study only one predictor variable (one target – one sample)
 - Linear: the relationship is linear and features are independent to each other.
- The ML model to predict the response

$$y \approx H(w, x) = w_0 + w_1 x$$

where y is the observed outcome, e.g., target variable, x is the explanatory variable, $w = [w_0, w_1]^T$ is the model parameters to be determined from ML training, and $H(w,x)$ is the linear regression model that would give the predicted target variable.





Linear Regression

The linear regression model involves a **linear combination** of the input variables.

- The simplest model

Formation:

$$y(\mathbf{x}, \mathbf{w}) = w_0 + w_1x_1 + \cdots + w_nx_n \quad (1)$$

- Assume we have a vector with D features, $\mathbf{x} = (x_1, \dots, x_D)^T$
- The key terms are the parameters $\mathbf{w} = (w_0, w_1, \dots, w_D)$, also known as weight.
- The predicted value function $y(\mathbf{x}, \mathbf{w})$ predicts the target variable, t . Then

$$t = y(\mathbf{x}, \mathbf{w}) + \epsilon$$

where ϵ is a noise that has a zero mean Gaussian random variable with *inverse variance* $\beta, \mathcal{N}(0, \beta^{-1})$.



Linear Regression

Eqn. (1) imposes significant limitations on the model.

Often, we extend (1) by considering linear combinations of fixed nonlinear functions of \mathbf{x} using basis function, $\phi_j(\mathbf{x})$:

$$y(\mathbf{x}, \mathbf{w}) = w_0 + w_1\phi_1 + \cdots + w_D\phi_n = w_0 + \sum_{j=1}^{M-1} w_j\phi_j(\mathbf{x}) \quad (2)$$

- If we let $\phi_0(\mathbf{x}) = 1$, then (2) becomes

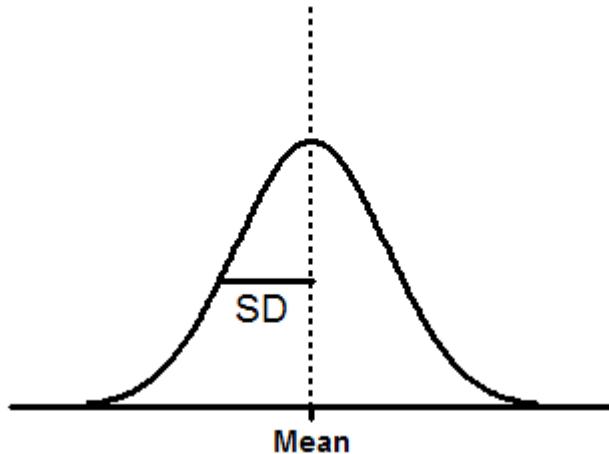
$$y(\mathbf{x}, \mathbf{w}) = \sum_{j=0}^{M-1} w_j\phi_j(\mathbf{x}) = \mathbf{w}^T \boldsymbol{\phi}(\mathbf{x})$$

- Is this function linear or non-linear?
 - Non-linear function ϕ
 - But the function is linear in \mathbf{w} .



Gaussian Distribution

$$\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} \exp\left\{-\frac{1}{2\sigma^2}(x - \mu)^2\right\}$$



Normalized: $\int_{-\infty}^{\infty} \mathcal{N}(x|\mu, \sigma^2) dx = 1$

Average: $\mathbb{E}[x] = \int_{-\infty}^{\infty} \mathcal{N}(x|\mu, \sigma^2) x dx = \mu$

Second order moment: $\mathbb{E}[x^2] = \int_{-\infty}^{\infty} \mathcal{N}(x|\mu, \sigma^2) x^2 dx = \mu^2 + \sigma^2$

Variance: $\mathbb{E}[x^2] - \mathbb{E}[x] = \sigma^2$

Probability: $p(x|\mu, \sigma^2) = \prod_{n=1}^N \mathcal{N}(x_n|\mu, \sigma^2)$

The log likelihood function:

$$\ln p(x|\mu, \sigma^2) = -\frac{1}{2\sigma^2} \sum_{n=1}^N (x_n - \mu)^2 - \frac{N}{2} \ln \sigma^2 - \frac{N}{2} \ln(2\pi) \quad (3)$$

$$\mu_{ML} = \frac{1}{N} \sum_{n=1}^N x_n$$

$$\sigma_{ML}^2 = \frac{1}{N} \sum_{n=1}^N (x_n - \mu_{ML})^2$$



Linear Regression

The target t probability is then

$$p(t|x, w, \beta) = \mathcal{N}(t|y(x, w), \beta^{-1})$$

Consider a data set:

Input: $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$

Target: $\mathbf{t} = \{t_1, \dots, t_N\}$

t is from a single observation and each data points are drawn independently from the distribution.

The likelihood function

$$p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) = \prod_{n=1}^N \mathcal{N}(t_n | \mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1})$$

So (3) simply because (4)

$$\ln p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) = \frac{N}{2} \ln \beta - \frac{N}{2} \ln(2\pi) - \frac{\beta}{2} \sum_{n=1}^N \{t_n - \mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}_n)\}^2$$

where $\beta = \sigma^{-2}$.



Estimation with MLE

$$\mathbf{w} = \operatorname{argmax}_{\mathbf{w}} P((y_1, x_1), \dots, (y_n, x_n) | \mathbf{w}) = \operatorname{argmax}_{\mathbf{w}} \prod_{i=1}^n P(y_i, x_i | \mathbf{w})$$

- Independence & chain rule of probability.
- x_i is independent of \mathbf{w} & $P(x_i)$ is constant and can be dropped.
- $P(y_i, x_i | \mathbf{w}) = P(y_i | x_i, \mathbf{w})P(x_i | \mathbf{w}) = P(y_i | x_i, \mathbf{w})P(x_i) = P(y_i | x_i, \mathbf{w})$

$$= \operatorname{argmax}_{\mathbf{w}} \prod_{i=1}^n P(y_i | x_i, \mathbf{w}) = \operatorname{argmax}_{\mathbf{w}} \sum_{i=1}^n \log[P(y_i | x_i, \mathbf{w})]$$

- Log is a monotonic function

$$= \operatorname{argmax}_{\mathbf{w}} \sum_{i=1}^n \left[\log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) + \log\left(\exp\left(-\frac{(x_i^T \mathbf{w} - y_i)^2}{2\sigma^2}\right)\right) \right]$$

- $P \sim N(x^T \mathbf{w}, \sigma^2)$

$$= \operatorname{argmax}_{\mathbf{w}} -\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i^T \mathbf{w} - y_i)^2 = \operatorname{argmin}_{\mathbf{w}} \frac{1}{n} \sum_{i=1}^n (x_i^T \mathbf{w} - y_i)^2$$

- The 1st term is a constant

- $\frac{1}{n}$ makes the lost interpretable (average squared error).

* This function is also known as the squared loss or Ordinary Least Square (OLS) which can be optimized with gradient descent, Newton's method, or in closed from: $\mathbf{w} = (X X^T)^{-1} X y^T$



Estimation with MAP

- Independence & chain rule of probability.
- x_i is independent of w & $P(\cdot)$ constant and can be dropped.
- $P(y_i, x_i|w) = P(y_i|x_i, w)P(x_i|w) = P(y_i|x_i, w)P(x_i) = P(y_i|x_i, w)$

$$w = \operatorname{argmax}_w P(w|(y_1, x_1), \dots (y_n, x_n)) = \operatorname{argmax}_w \frac{P(w|(y_1, x_1), \dots (y_n, x_n))P(w)}{P((y_1, x_1), \dots (y_n, x_n))}$$

$$= \operatorname{argmax}_w \left[\prod_{i=1}^n P(y_i|x_i, w) \right] P(w) = \operatorname{argmax}_w \sum_{i=1}^n \log[P(y_i|x_i, w)] + \log P(w)$$

$$= \operatorname{argmin}_w \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i^T w - y_i)^2 + \frac{1}{\tau^2} w^T w = \operatorname{argmin}_w \frac{1}{n} \sum_{i=1}^n (x_i^T w - y_i)^2 + \lambda \|w\|^2$$

- $w \sim N(0, \tau^2)$ assumption

- $\lambda = \sigma^2/n\tau^2$

* This objective is known as Ridge Regression (Next lecture topic).



Linear Regression Example

- <https://archive.ics.uci.edu/ml/machine-learning-databases/housing/>
 - 1. CRIM: per capita crime rate by town
 - 2. ZN: proportion of residential land zoned for lots over 25,000 sq.ft.
 - 3. INDUS: proportion of non-retail business acres per town
 - 4. CHAS: Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
 - 5. NOX: nitric oxides concentration (parts per 10 million)
 - 6. RM: average number of rooms per dwelling
 - 7. AGE: proportion of owner-occupied units built prior to 1940
 - 8. DIS: weighted distances to five Boston employment centres
 - 9. RAD: index of accessibility to radial highways
 - 10. TAX: full-value property-tax rate per \$10,000
 - 11. PTRATIO: pupil-teacher ratio by town
 - 12. B: $1000(Bk - 0.63)^2$ where Bk is the proportion of blacks by town
 - 13. LSTAT: % lower status of the population
 - 14. MEDV: Median value of owner-occupied homes in \$1000's



Two simple but often ignored questions before applying any ML algorithms

- First, we need to identify what data feature to predict from the model
 - Assume we're interested in predicting the housing prices

0.00632	18.00	2.310	0	0.5380	6.5750	65.20	4.0900	1	296.0	15.30	396.90	4.98	24.00
0.02731	0.00	7.070	0	0.4690	6.4210	78.90	4.9671	2	242.0	17.80	396.90	9.14	21.60
0.02729	0.00	7.070	0	0.4690	7.1850	61.10	4.9671	2	242.0	17.80	392.83	4.03	34.70
0.03237	0.00	2.180	0	0.4580	6.9980	45.80	6.0622	3	222.0	18.70	394.63	2.94	33.40
0.06905	0.00	2.180	0	0.4580	7.1470	54.20	6.0622	3	222.0	18.70	396.90	5.33	36.20
0.02985	0.00	2.180	0	0.4580	6.4300	58.70	6.0622	3	222.0	18.70	394.12	5.21	28.70
0.08829	12.50	7.870	0	0.5240	6.0120	66.60	5.5605	5	311.0	15.20	395.60	12.43	22.90
0.14455	12.50	7.870	0	0.5240	6.1720	96.10	5.9505	5	311.0	15.20	396.90	19.15	27.10
0.21124	12.50	7.870	0	0.5240	5.6310	100.00	6.0821	5	311.0	15.20	386.63	29.93	16.50

- So the last column MEDV becomes our target variable (and it's numerical)
- Second, what explanatory variables do we use for ML
 - These could be determined/given by the application in hand as well
 - What explanatory variables are easy to get in practice?
 - We can also use whatever data features we have (all other features)
 - Will computation become an issue? Are all of them good quality data?
 - We can explore the data to draw a sensible set of features (e.g., feature selection)



Exploratory data analysis

- Let's visualize pair-wise scatterplots correlations between features

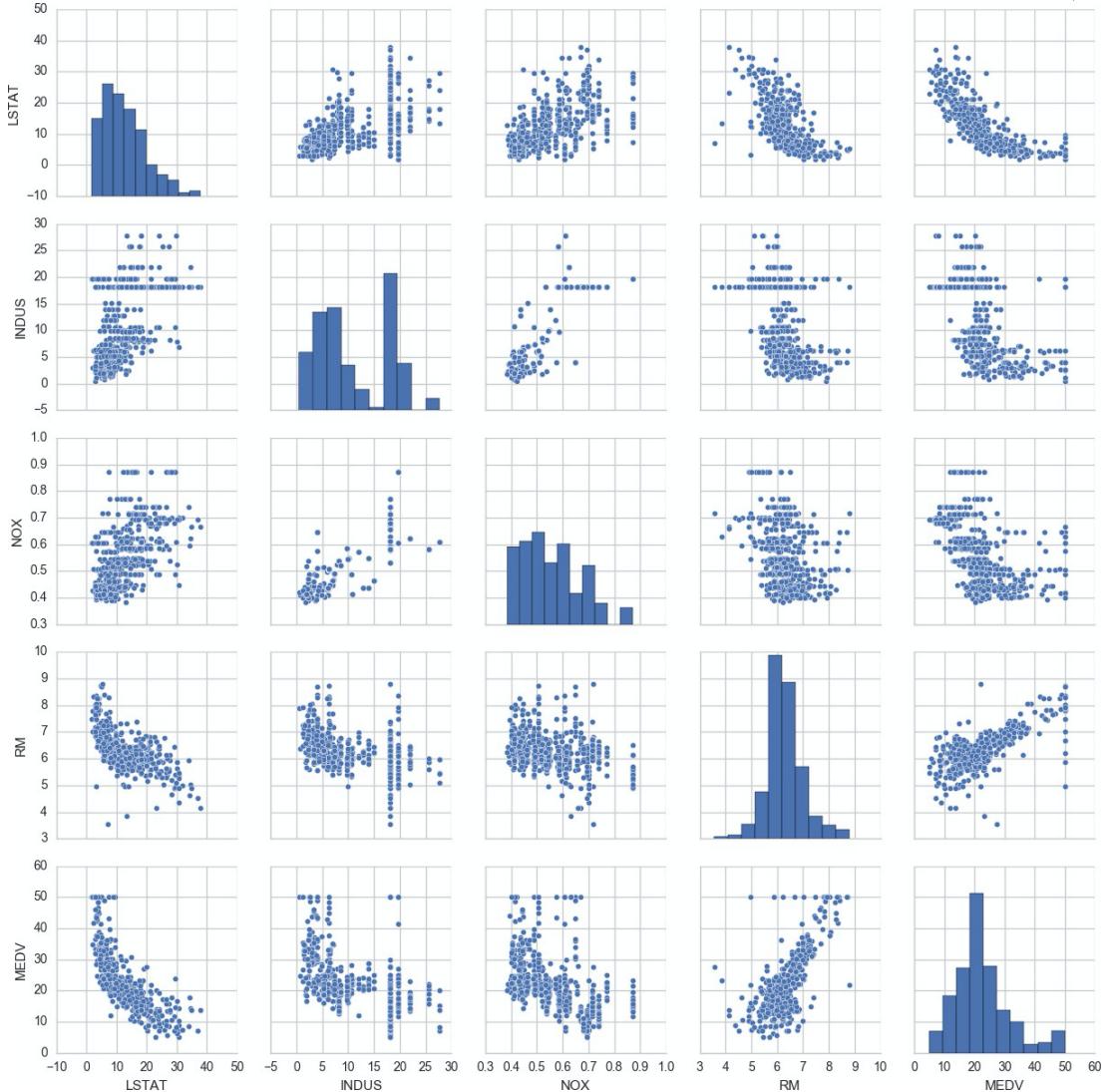
LSTAT: % lower status of the population

INDUS: proportion of non-retail business acres per town

NOX: nitric oxides concentration (parts per 10 million)

RM: average number of rooms per dwelling

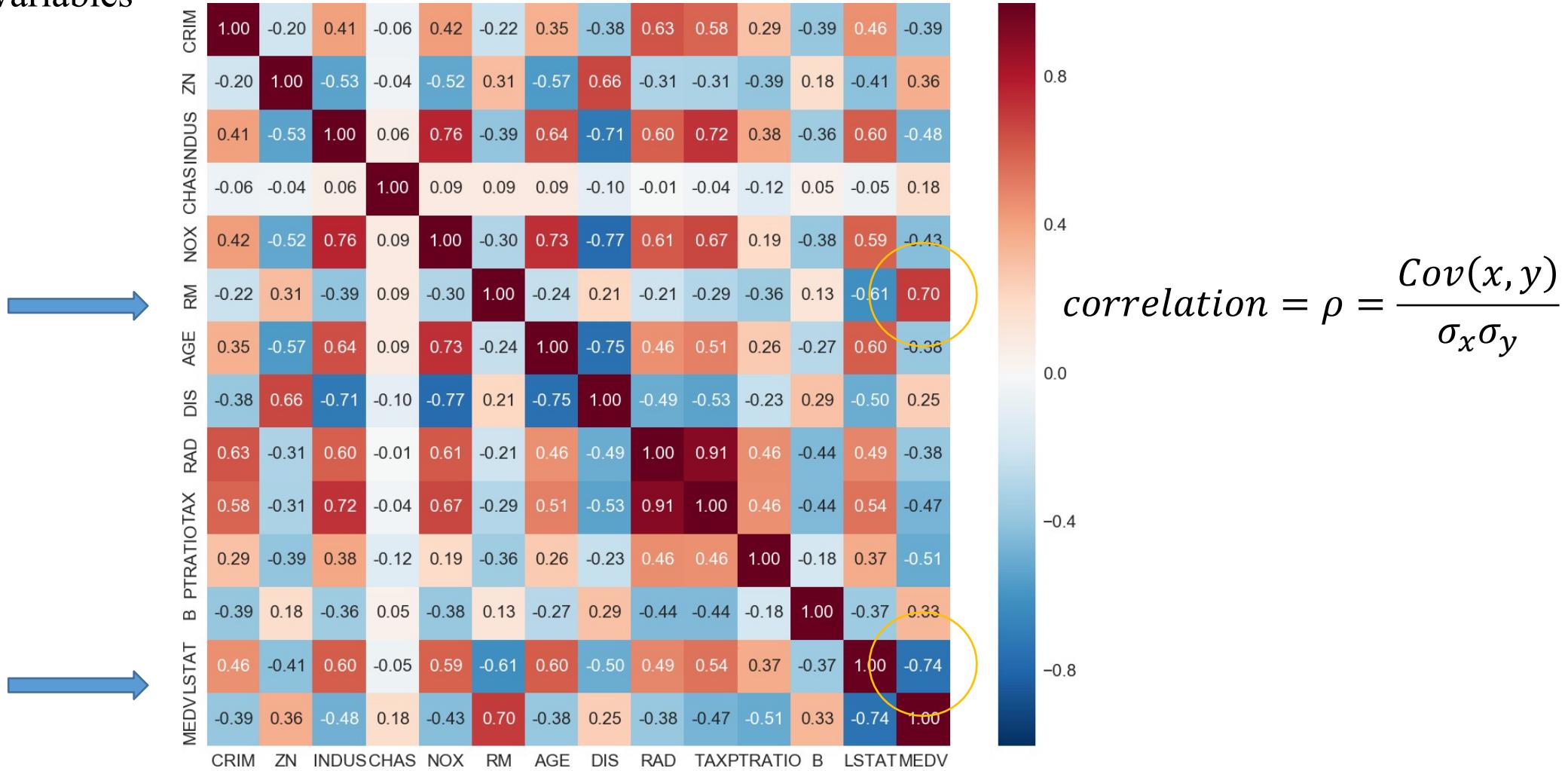
MEDV: Median value of owner-occupied homes in \$1000's





Correlation matrix for the data

- Two features (RM and LSAT) have higher correlation with MEDV, implying high chances of being explanatory variables





Simple linear regression example

- The explanatory variable is chosen as RM (average number of rooms per dwelling)
- The target variable is chosen as MEDV (Median value of owner-occupied homes in \$1000's)
- The ML model to predict the response

$$H(w, x) = w_0 + w_1 x$$

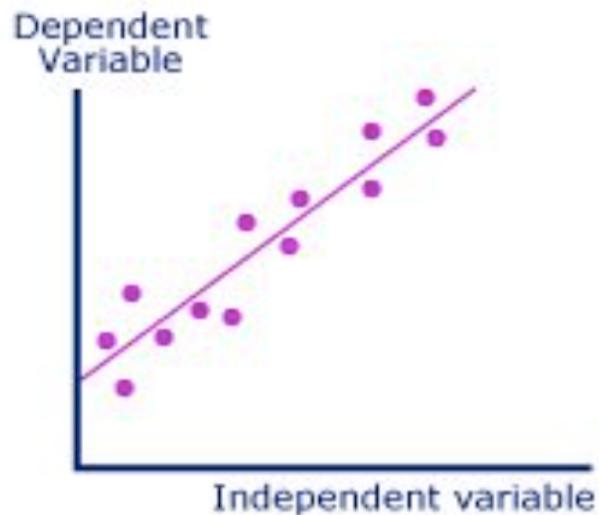
where y is the observed outcome, e.g., target variable, x is the explanatory variable, $w = [w_0, w_1]^T$ is the model parameters to be determined from ML training



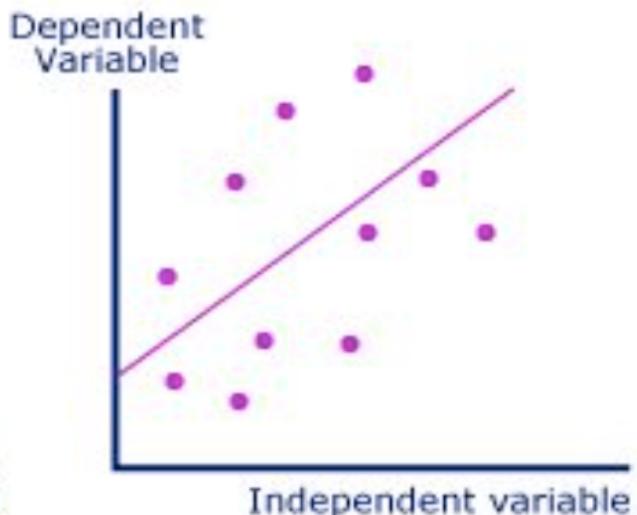
What is the best fit?

Once a linear model is fitted, we assess the overall accuracy of the model.

Two graphs below have fitted with the same model.



Graph A



Graph B



What is the best fit?

The usual way to measure the overall accuracy is to use the coefficient of determination, R^2 .

It measures how well data fits a model or how well the model describes the data:

$$R^2 = 1 - \frac{RSS}{TSS}$$

where TSS is the total sum of squares, which measures the total variance of the output data y and RSS is the residual sum of squares.

Given a dataset, TSS is completely determined and the fitted linear model has the minimum RSS.

$R^2 = 1$ indicates that the regression line perfectly fits the data.

$R^2 = 0$ indicates that the line does not fit the data at all.



What is the best fit?

- Prediction Error

$$e^i = y^i - H(w, x^i)$$

where (x^i, y^i) represents the i-th observed data (RM, MEDV)

- Problem formulation: minimize Sum of Squared Errors (SSE aka Least Square Problem)

$$\text{Minimize}_w f(w) = \frac{1}{2} \sum_{i=1}^n (y^i - H(w, x^i))^2$$

where n is the number of total observations used for training

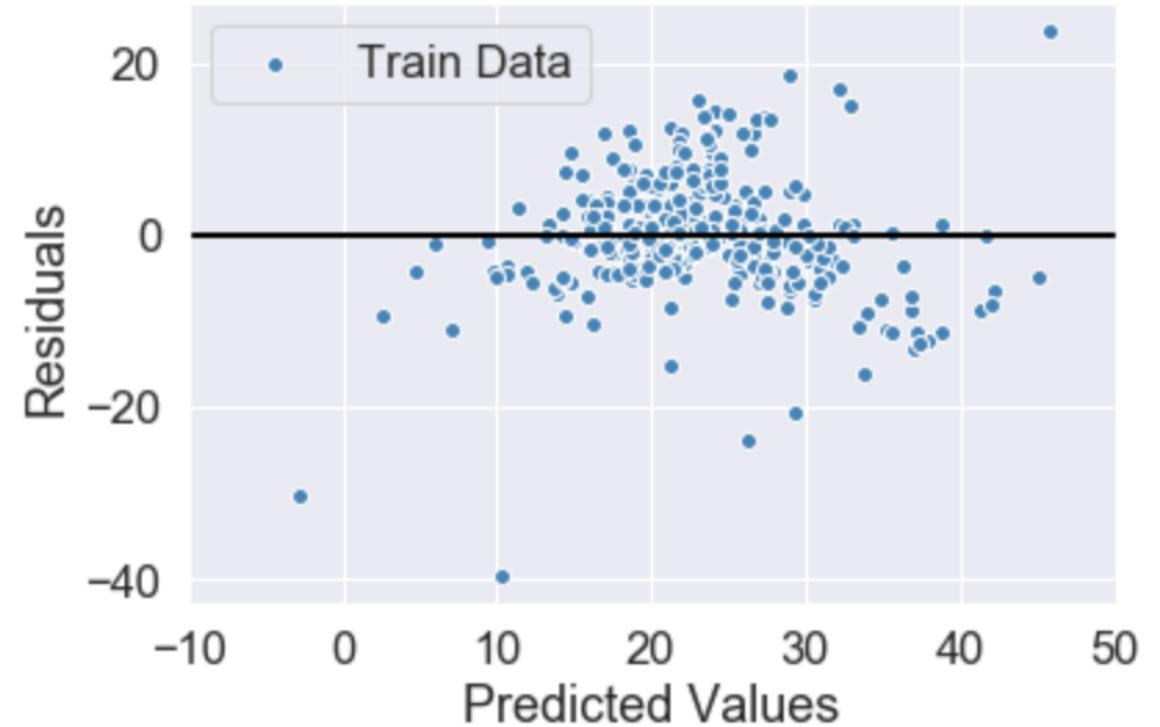
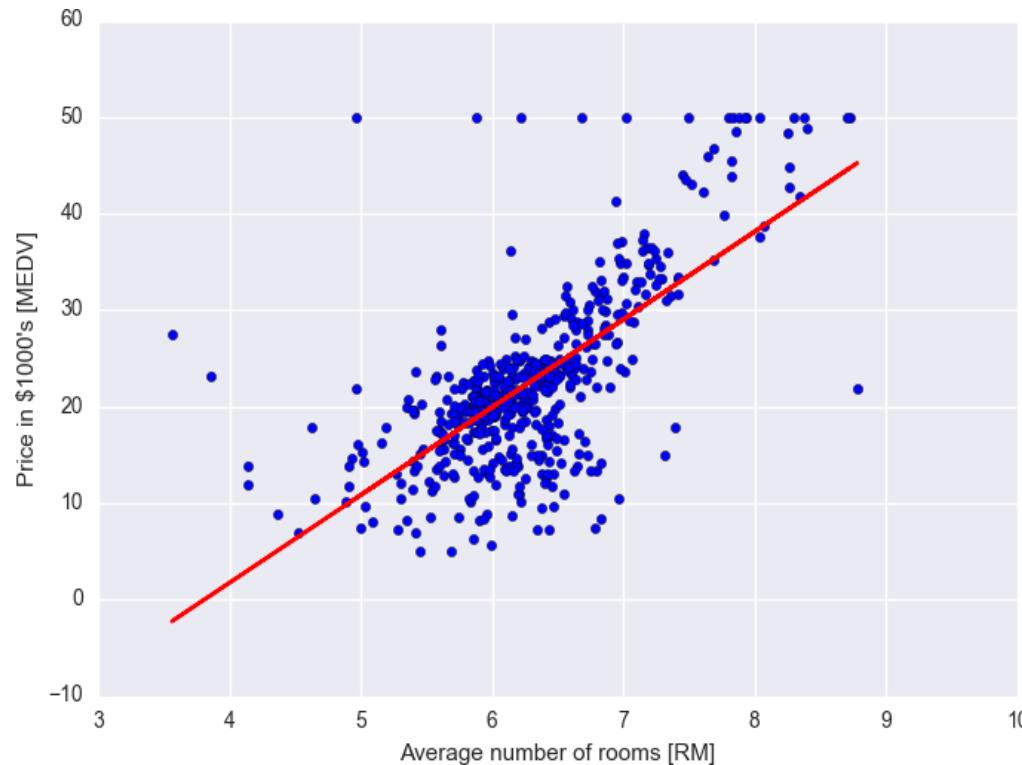
- This is a typical unconstrained optimization problem, more specifically, quadratic programming problem

$$\text{Minimize}_w f(w) = \frac{1}{2} \sum_{i=1}^n (y^i - w_0 - w_1 x^i)^2$$



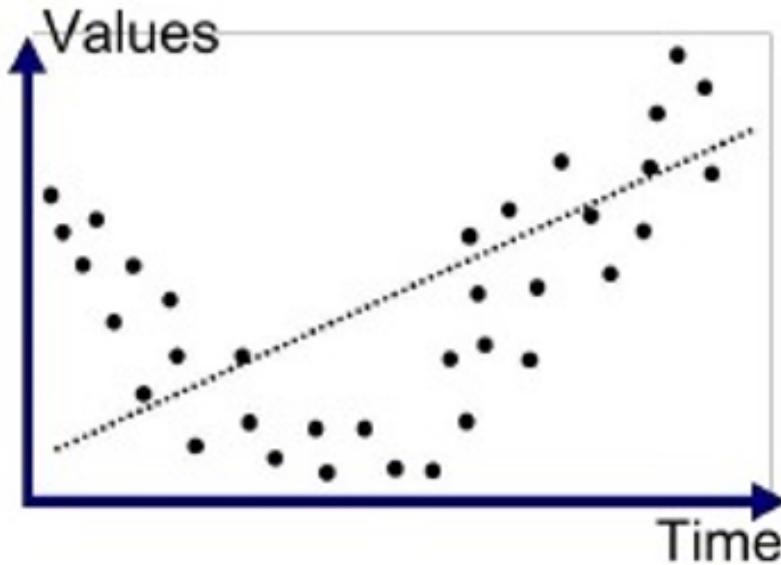
What is the best model?

- Is the model good enough?

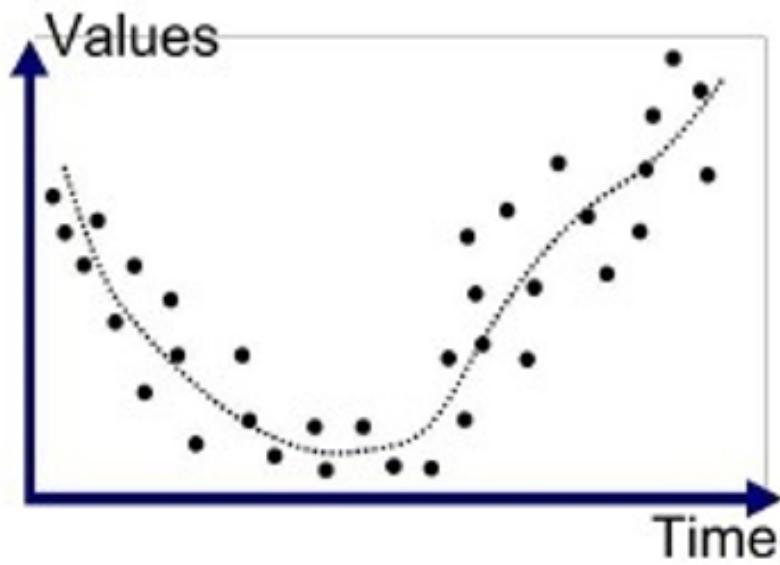




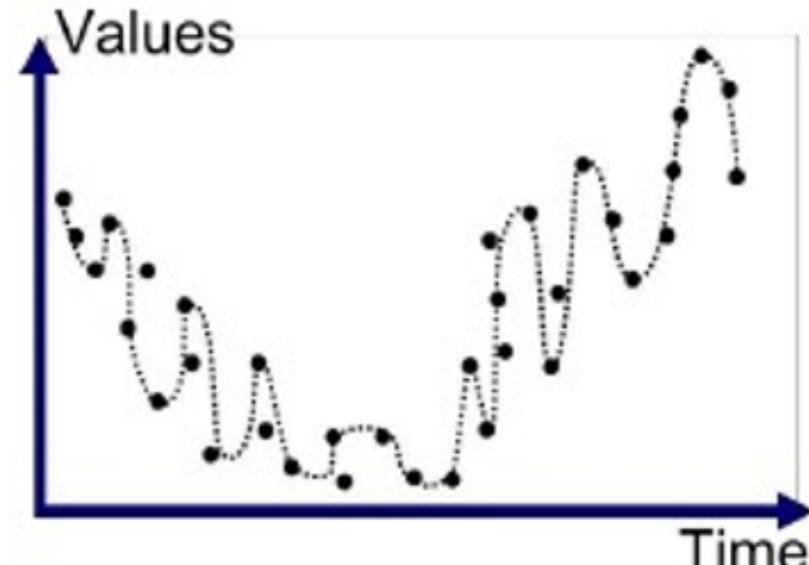
Overfitting vs. Underfitting



Underfitted



Good Fit/Robust



Overfitted



Extensions of linear regression formulations

- Multiple linear regression model

$$H(w, x) = w_0 + \sum w_i x_i$$

- Polynomial regression

$$H(x, w) = w_0 + w_1 x + w_2 x^2 + \cdots + w_d x^d$$

- Nonlinear transformation of some data features

- For example, log transform the LSAT data

$$H(w, x) = w_0 + w_1 x_1 + w_2 \log x_2$$

- Since the unknown model parameters are linear, they all end up with the same quadratic-like formulation



Nonlinear regression

- So far our regression models are all linear w.r.t. model parameters

$$H(\mathbf{w}, \mathbf{x}) = w_0 + w_1 x_1 + \cdots + w_m x_m$$

$$H(\mathbf{w}, \mathbf{x}) = w_0 + w_1 x + w_2 x^2 \dots + w_d x^d$$

$$H(\mathbf{w}, \mathbf{x}) = w_0 + w_1 x_1 + w_2 \log(x_2) + w_3 \log^2(x_2)$$

- Nonlinear regression is simply a replacement of the model with a nonlinear function w.r.t. parameters

$$H(\mathbf{w}, \mathbf{x}) = e^{w_0 + w_1 x_1} + w_2 x_2$$

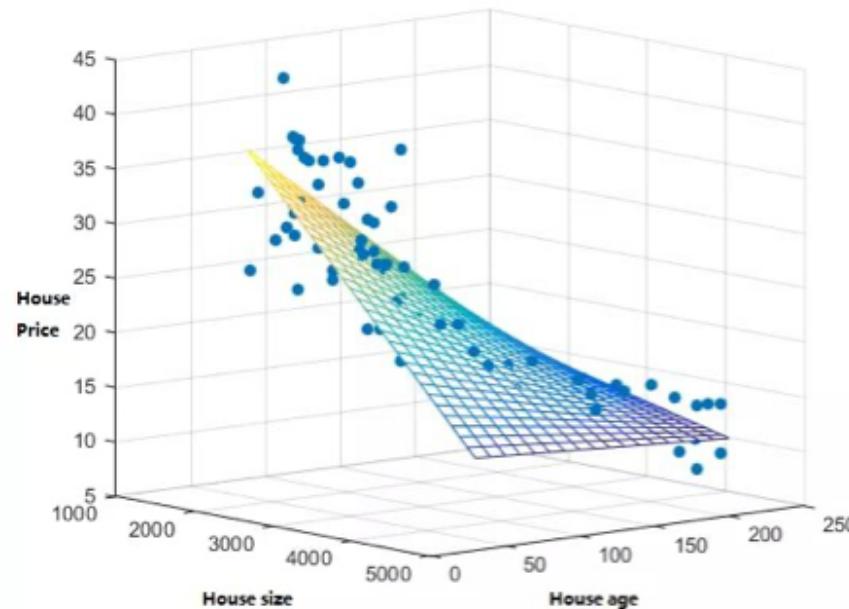
$$H(\mathbf{w}, \mathbf{x}) = \frac{w_0 + w_1 x_1}{1 + w_2 x_2}$$

- Note some nonlinear form can be transformed to be a linear problem

$$\mathbf{y} = H(\mathbf{w}, \mathbf{x}) = e^{w_0 + w_1 x_1} \rightarrow \ln(\mathbf{y}) = w_0 + w_1 x_1$$



Linear Regression Assumptions and Limitations



Assumptions:

No **multicollinearity** between the predictor variables.

- The relationship among independent variables.
- Redundancies within predictors.
- Difficulties in making inferences between predictors and target.
- The standard errors of estimates will be inflated.
- The power and reliability of coefficients will decrease.
- The needs of a larger sample size will be necessary.

Ways to overcome?

- Find multicollinearity and do model selection.
- We can do stepwise procedures. (Is this the best way?)
 - Regularization
 - Bias-Variance Tradeoff
 - Model Selection



- Next Lecture
 - Regularization
 - Gradient Descent
 - Classification