CS145: INTRODUCTION TO DATA MINING

2: Vector Data: Prediction

Instructor: Yizhou Sun

yzsun@cs.ucla.edu

January 6, 2019

Methods to Learn

	Vector Data	Set Data	Sequence Data	Text Data
Classification	Logistic Regression; Decision Tree; KNN SVM; NN			Naïve Bayes for Text
Clustering	K-means; hierarchical clustering; DBSCAN; Mixture Models			PLSA
Prediction	**Linear Regression** GLM*			
Frequent Pattern Mining		Apriori; FP growth	GSP; PrefixSpan	
Similarity Search			DTW	

How to learn these algorithms?

- Three levels
  - When it is applicable?
    - Input, output, strengths, weaknesses, time complexity
  - How it works?
    - Pseudo-code, work flows, major steps
    - Can work out a toy problem by pen and paper
  - Why it works?
    - Intuition, philosophy, objective, derivation, proof

Vector Data: Prediction

- Vector Data
- Linear Regression Model
- Model Evaluation and Selection
- Summary

Example

	Sex	Race	Height	Income	Marital Status	Years of Educ.	Liberal-ness
R1001	M	1	70	50	1	12	1.73
R1002	M	2	72	100	2	20	4.53
R1003	F	1	55	250	1	16	2.99
R1004	M	2	65	20	2	16	1.13
R1005	F	1	60	10	3	12	3.81
R1006	M	1	68	30	1	9	4.76
R1007	F	5	66	25	2	21	2.01
R1008	F	4	61	43	1	18	1.27
R1009	M	1	69	67	1	12	3.25

A matrix of $\mathrm{n} \times p$:

- n data objects / points
- p attributes / dimensions

$$\begin{bmatrix} x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{nf} & \cdots & x_{np} \end{bmatrix}$$

Attribute Type

- Numerical
  - E.g., height, income
- Categorical / discrete
  - E.g., Sex, Race

Categorical Attribute Types

- **Nominal:** categories, states, or "names of things"
  - *Hair_color* = {*auburn, black, blond, brown, grey, red, white*}
  - marital status, occupation, ID numbers, zip codes
- **Binary**
  - Nominal attribute with only 2 states (0 and 1)
  - Symmetric binary: both outcomes equally important
    - e.g., gender
  - Asymmetric binary: outcomes not equally important.
    - e.g., medical test (positive vs. negative)
    - Convention: assign 1 to most important outcome (e.g., HIV positive)
- **Ordinal**
  - Values have a meaningful order (ranking) but magnitude between successive values is not known.
  - *Size* = {*small, medium, large*}, grades, army rankings

Basic Statistical Descriptions of Data

- Central Tendency
- Dispersion of the Data
- Graphic Displays

Measuring the Central Tendency

- Mean (algebraic measure) (sample vs. population):

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \mu = \frac{\sum x}{N}$$

  Note: *n* is sample size and *N* is population size.

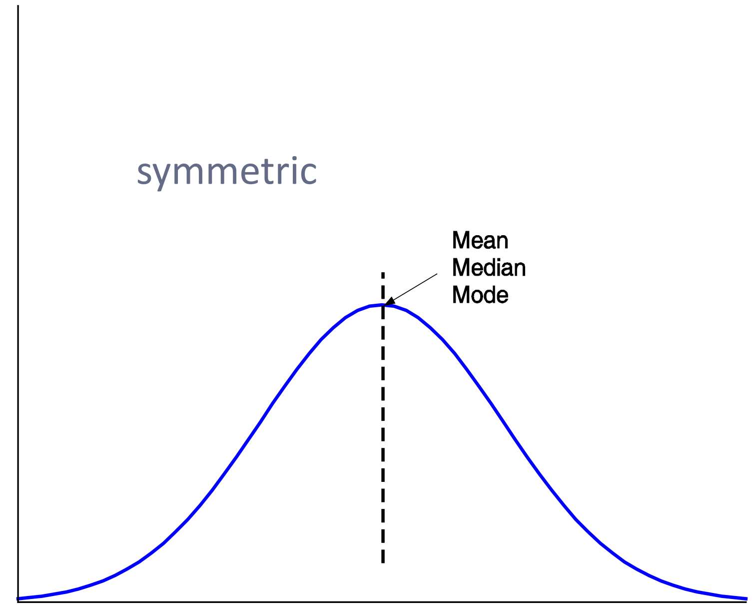
  - Weighted arithmetic mean:

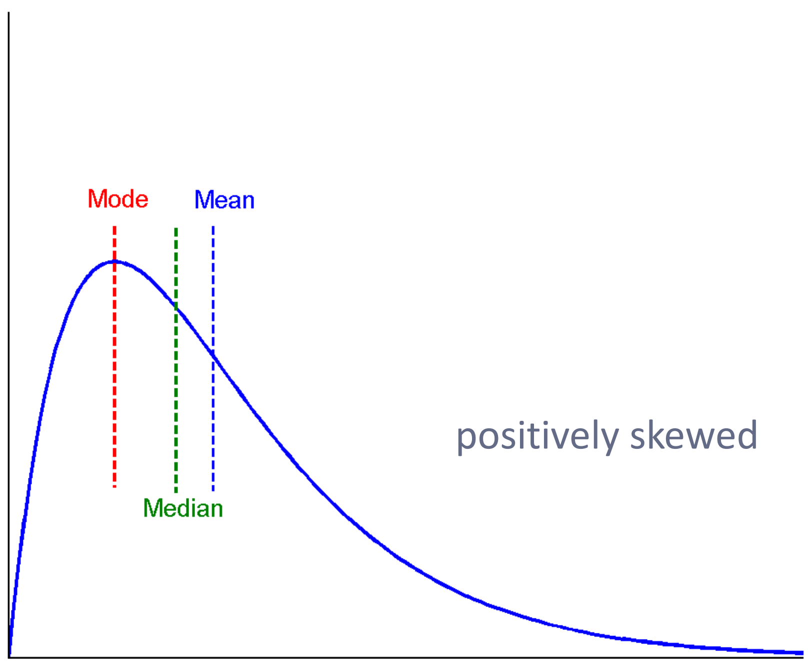
$$\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

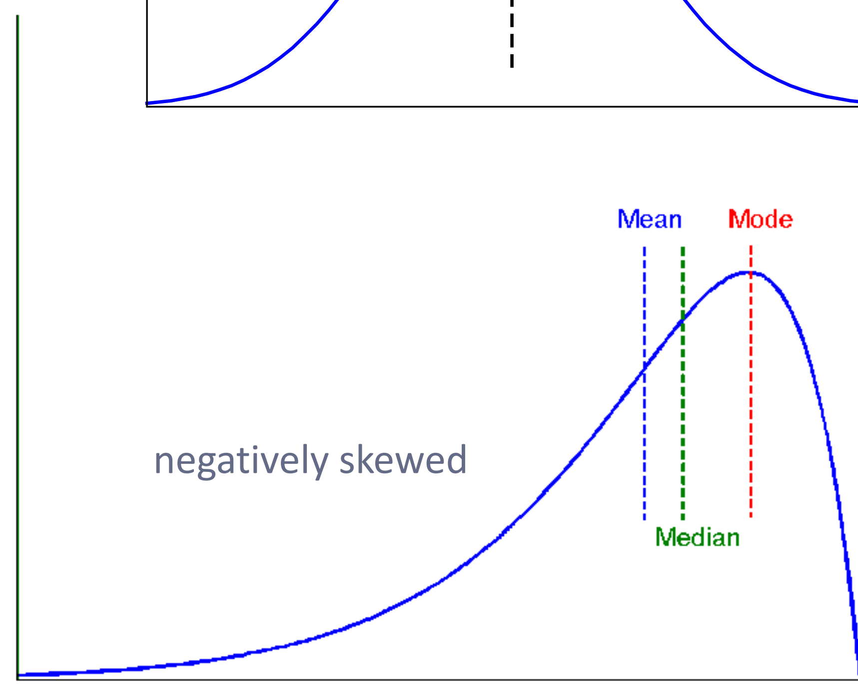
  - Trimmed mean: chopping extreme values
- Median:
  - Middle value if odd number of values, or average of the middle two values otherwise
- Mode
  - Value that occurs most frequently in the data
  - Unimodal, bimodal, trimodal
  - Empirical formula: $mean - mode = 3 \times (mean - median)$

Symmetric vs. Skewed Data

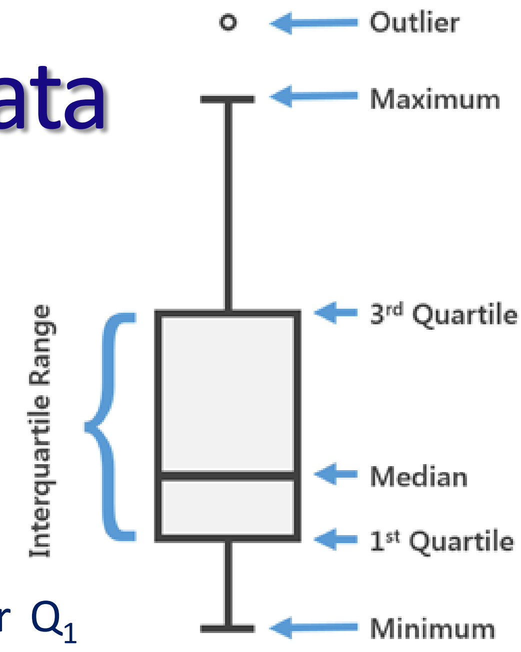
- Median, mean and mode of symmetric, positively and negatively skewed data







Measuring the Dispersion of Data



Box Plot

- Quartiles, outliers and boxplots
  - **Quartiles**: Q_1 (25th percentile), Q_3 (75th percentile)
  - **Inter-quartile range**: IQR = $Q_3 - Q_1$
  - **Five number summary**: min, Q_1, median, Q_3, max
  - **Outlier**: usually, a value higher/lower than 1.5 x IQR of Q_3 or Q_1
- Variance and standard deviation (*sample: s, population: σ)*
  - **Variance**: (algebraic, scalable computation)
    - $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{n-1}[\sum_{i=1}^{n}x_i^2 - \frac{1}{n}(\sum_{i=1}^{n}x_i)^2]$
    - $\sigma^2 = E[(X - E(X))^2] = E(X^2) - (E(X))^2$
  - **Standard deviation** *s (or σ)* is the square root of variance s^2 *(or σ^2)*

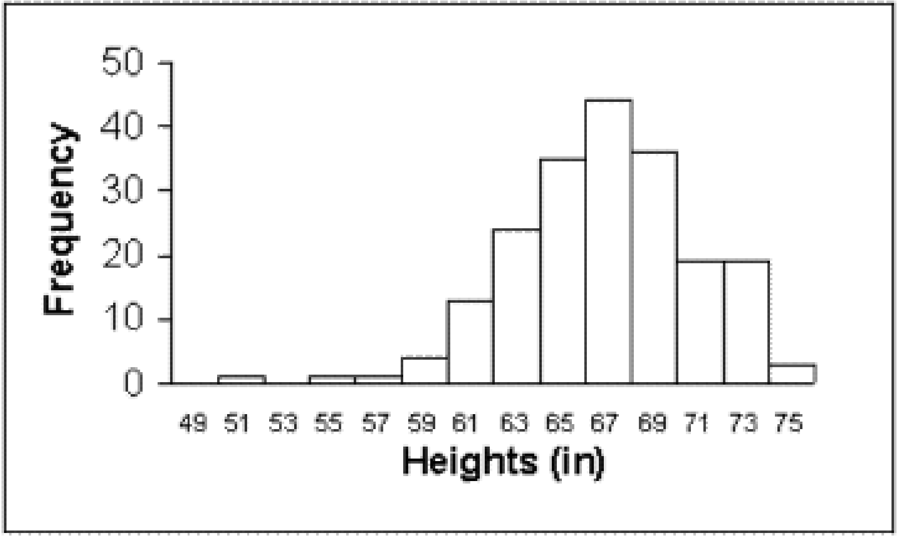
Graphic Displays of Basic Statistical Descriptions

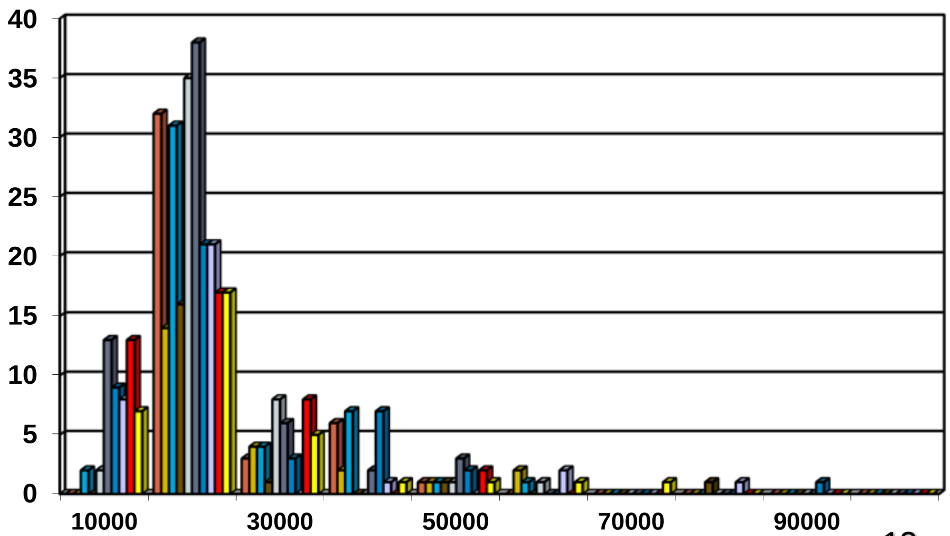
- **Histogram**: x-axis are values, y-axis repres. frequencies

- **Scatter plot**: each pair of values is a pair of coordinates and plotted as points in the plane

Histogram Analysis

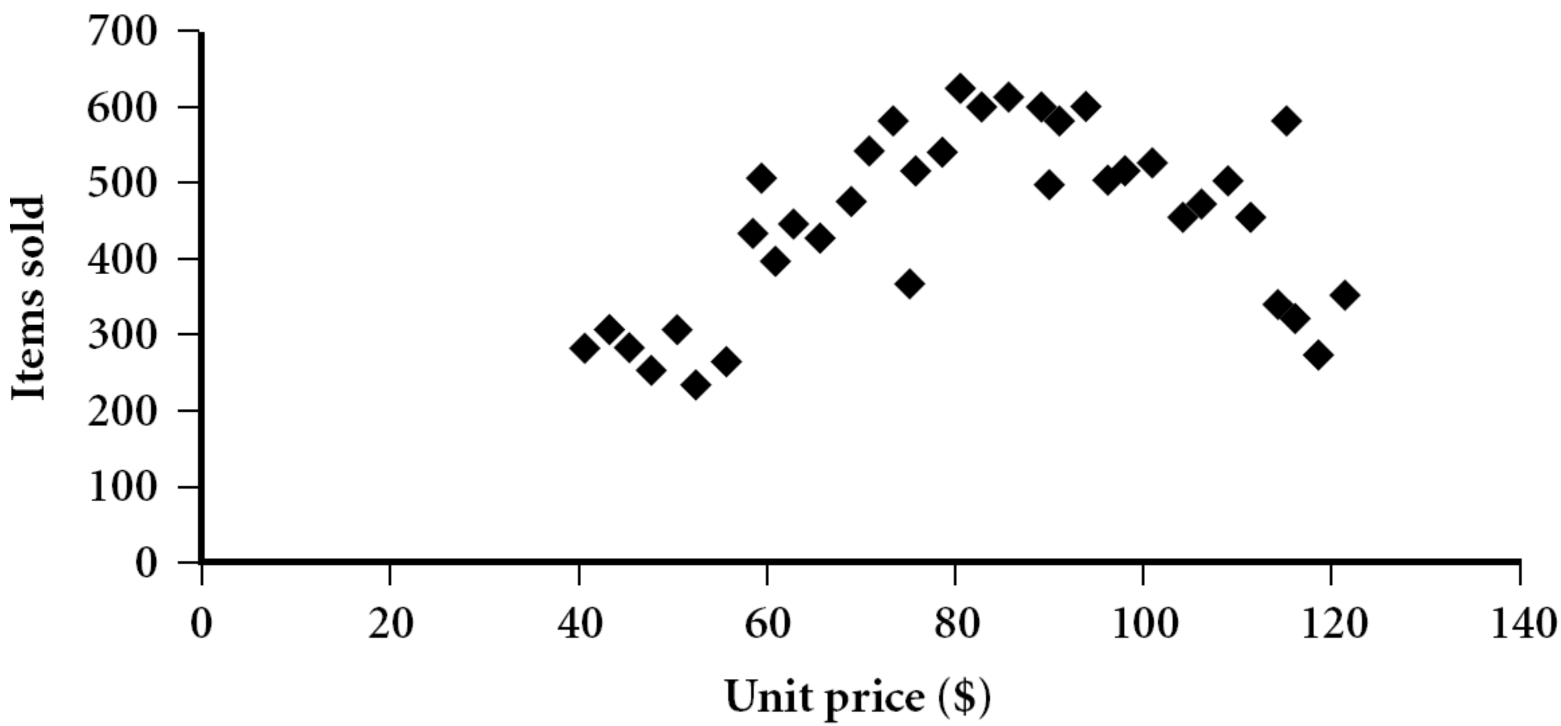
- Histogram: Graph display of tabulated frequencies, shown as bars
- It shows what proportion of cases fall into each of several categories
- Differs from a bar chart in that it is the *area* of the bar that denotes the value, not the height as in bar charts, a crucial distinction when the categories are not of uniform width
- The categories are usually specified as non-overlapping intervals of some variable. The categories (bars) must be adjacent



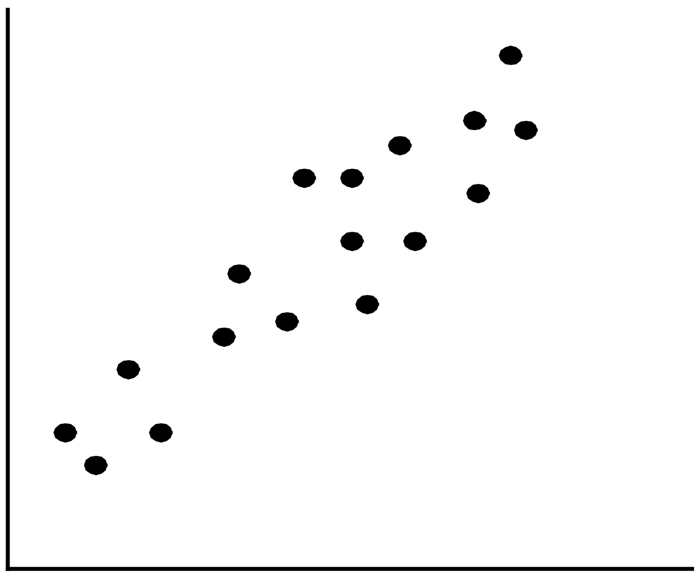


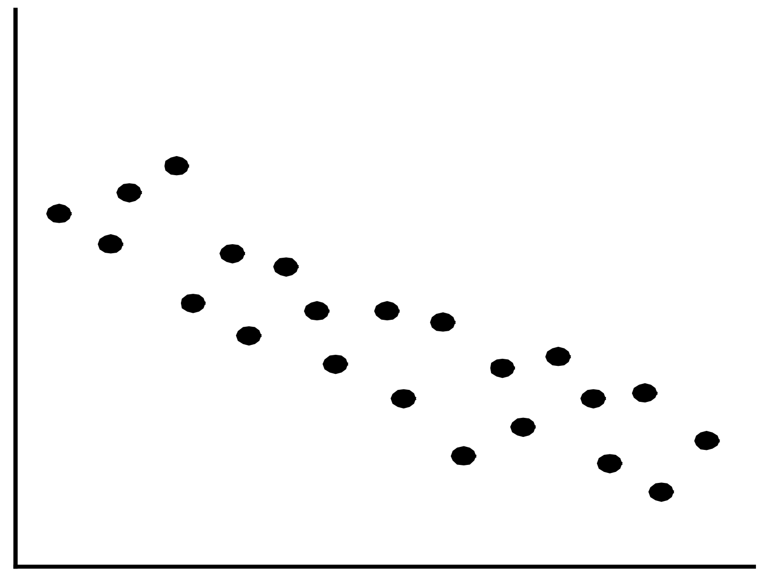
Scatter plot

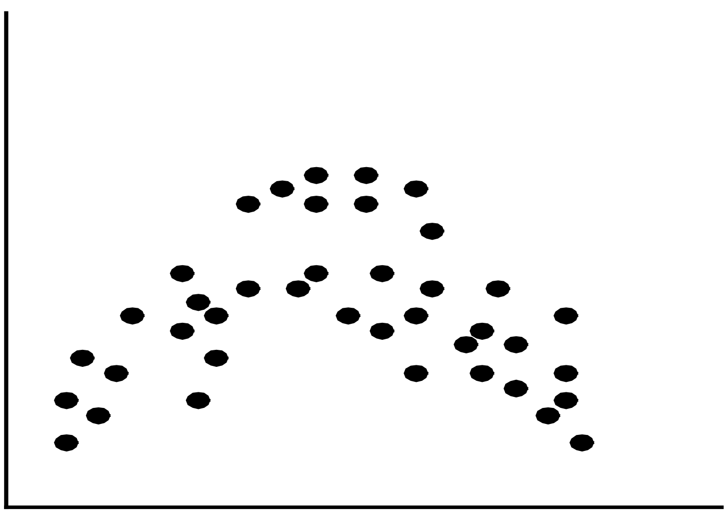
- Provides a first look at bivariate data to see clusters of points, outliers, etc
- Each pair of values is treated as a pair of coordinates and plotted as points in the plane



Positively and Negatively Correlated Data

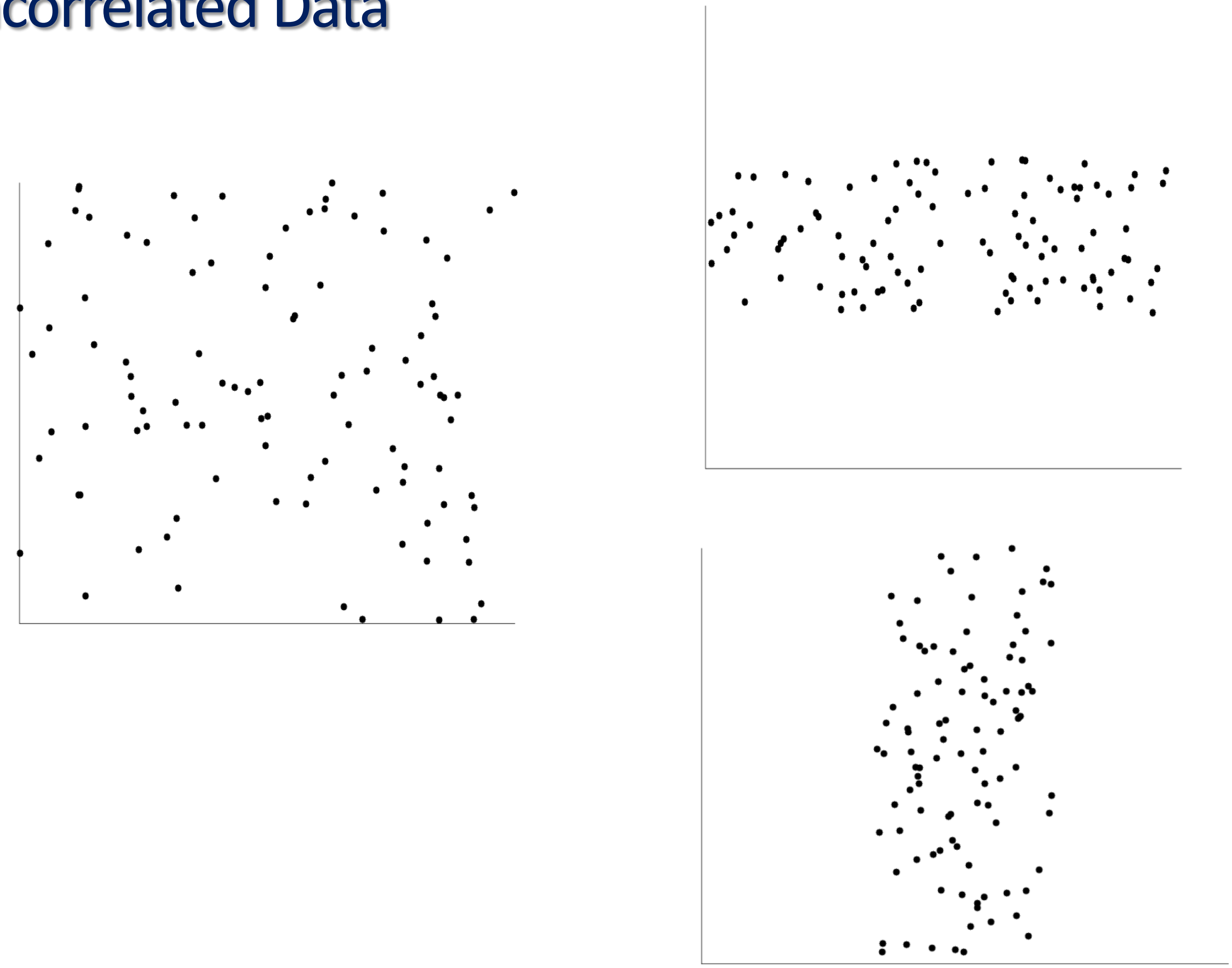




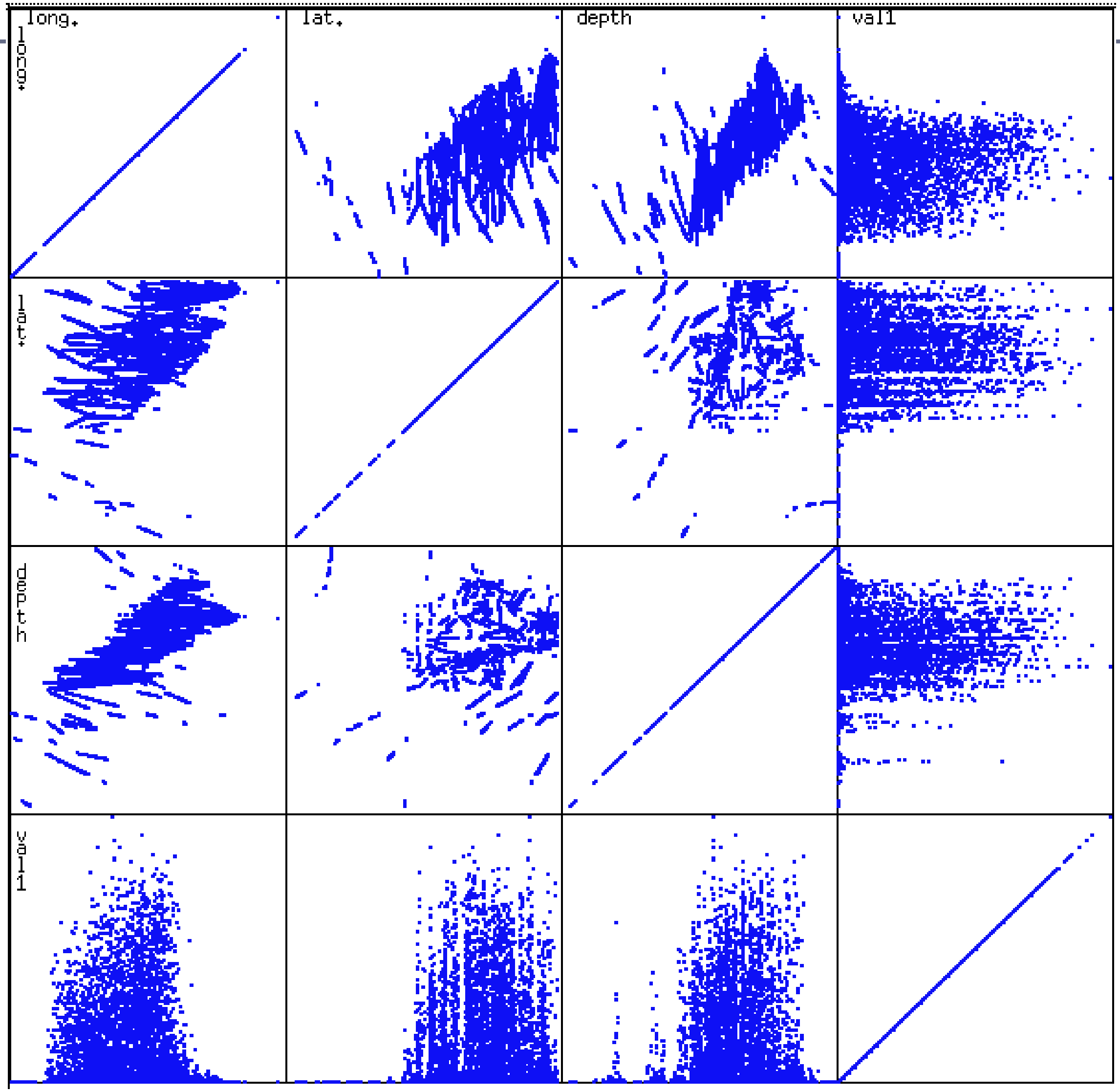


- The left half fragment is positively correlated
- The right half is negative correlated

Uncorrelated Data



Scatterplot Matrices



Matrix of scatterplots (x-y-diagrams) of the k-dim. data [total of $\binom{k}{2} + k$ unique scatterplots]

Vector Data: Prediction

- Vector Data
- Linear Regression Model
- Model Evaluation and Selection
- Summary

Linear Regression

- Ordinary Least Square Regression
  - Closed form solution
  - Gradient descent
- Linear Regression with Probabilistic Interpretation

The Linear Regression Problem

- Any Attributes to Continuous Value: $\mathbf{x} \Rightarrow y$
  - {age; major ; gender; race} ⇒ GPA
  - {income; credit score; profession} ⇒ loan
  - {college; major ; GPA} ⇒ future income
  - …

Example of House Price

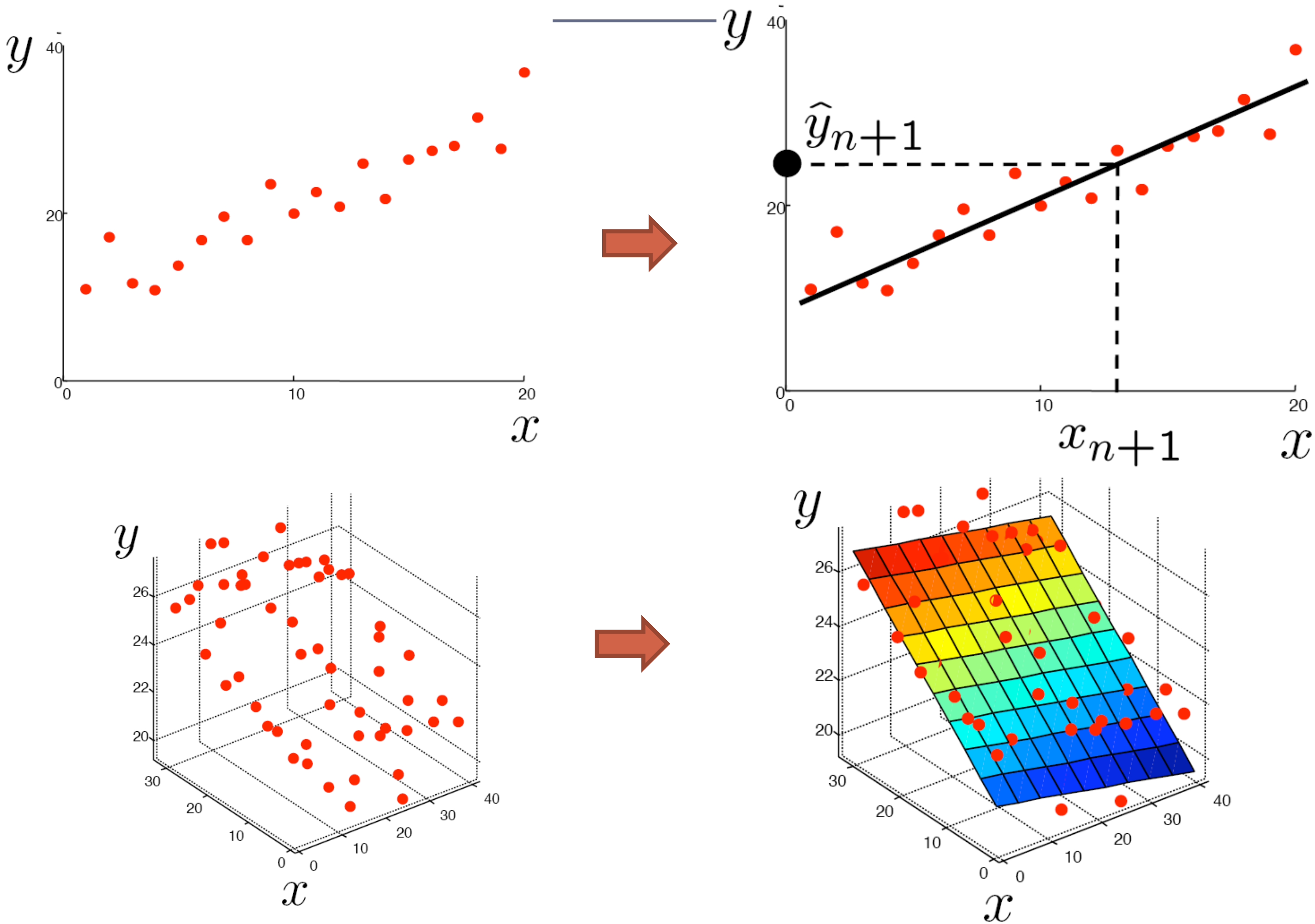
Living Area (sqft)	# of Beds	Price (1000$)
2104	3	400
1600	3	330
2400	3	369
1416	2	232
3000	4	540

x=$(x_1, x_2)'$

y

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

Illustration



Formalization

- Data: n independent data objects
  - $y_i, i = 1, \dots, n$
  - $\boldsymbol{x}_i = \left(x_{i1}, x_{i2}, \dots, x_{ip}\right)^{\mathrm{T}}, \mathrm{i} = 1, \dots, n$
    - A constant factor is added to model the bias term, $\mathrm{i.e.}, x_{i0} = 1$
    - New x: $\boldsymbol{x}_i = \left(x_{i0}, x_{i1}, x_{i2}, \dots, x_{ip}\right)^{\mathrm{T}}$
- Model:
  - y: *dependent variable*
  - $\boldsymbol{x}$: *explanatory variables*
  - $\boldsymbol{\beta} = \left(\beta_0, \beta_1, \dots, \beta_p\right)^T: weight\ vector$
  - $y = \boldsymbol{x}^T \boldsymbol{\beta} = \beta_0 + x_1\beta_1 + x_2\beta_2 + \cdots + x_p\beta_p$

A 3-step Process

- Model Construction
  - Use training data to find the best parameter $\boldsymbol{\beta}$, denoted as $\widehat{\boldsymbol{\beta}}$
- Model Selection
  - Use validation data to select the best model
    - E.g., Feature selection
- Model Usage
  - Apply the model to the unseen data (test data): $\hat{y} = \boldsymbol{x}^T\widehat{\boldsymbol{\beta}}$

Least Square Estimation

- Cost function (Mean Square Error):
  - $J(\boldsymbol{\beta}) = \frac{1}{2}\sum_i \left(\boldsymbol{x}_i^T \boldsymbol{\beta} - y_i\right)^2 / n$
- Matrix form:
  - $J(\boldsymbol{\beta}) = (\mathrm{X}\boldsymbol{\beta} - \boldsymbol{y})^T (X\boldsymbol{\beta} - \boldsymbol{y}) / 2n$

    $or \left|\left|\mathrm{X}\boldsymbol{\beta} - \boldsymbol{y}\right|\right|^2 / 2n$

$$\begin{bmatrix} 1, x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 1, x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 1, x_{n1} & \cdots & x_{nf} & \cdots & x_{np} \end{bmatrix} \qquad \begin{pmatrix} y_1 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{pmatrix}$$

$\boldsymbol{X}$: $\boldsymbol{n} \times (\boldsymbol{p} + \mathbf{1})$ matrix **y: $\boldsymbol{n} \times \mathbf{1}$ vector**

Ordinary Least Squares (OLS)

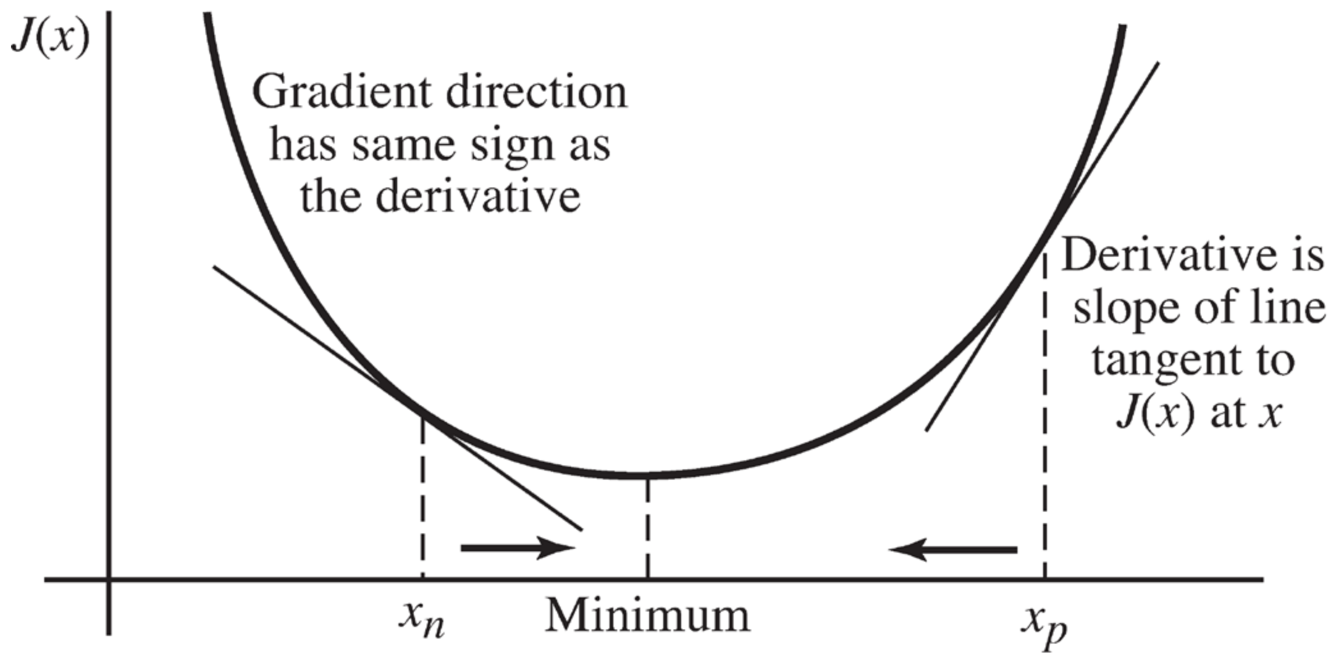
- Goal: find $\widehat{\boldsymbol{\beta}}$ that minimizes $J(\boldsymbol{\beta})$
  - $J(\boldsymbol{\beta}) = \frac{1}{2n}(\mathrm{X}\boldsymbol{\beta} - y)^T(X\boldsymbol{\beta} - y)$
    $= \frac{1}{2n}(\boldsymbol{\beta}^T X^T X\boldsymbol{\beta} - y^T X\boldsymbol{\beta} - \boldsymbol{\beta}^T X^T y + y^T y)$
- Ordinary least squares
  - Set first derivative of $J(\boldsymbol{\beta})$ as 0
  - $\frac{\partial J}{\partial \boldsymbol{\beta}} = (X^T X\boldsymbol{\beta} - X^T y)/n = 0$
  - $\Rightarrow \widehat{\boldsymbol{\beta}} = \left(X^T X\right)^{-1} X^T y$

z	$\frac{\partial z}{\partial x}$
$\mathbf{Ax}$	$\mathbf{A}^T$
$\mathbf{x}^T\mathbf{A}$	$\mathbf{A}$
$\mathbf{x}^T\mathbf{x}$	$2\mathbf{x}$
$\mathbf{x}^T\mathbf{Ax}$	$\mathbf{Ax} + \mathbf{A}^T\mathbf{x}$

More about matrix calculus: https://atmos.washington.edu/~dennis/MatrixCalculus.pdf

Gradient Descent

- Minimize the cost function by moving down in the steepest direction



Arrows point in minus gradient direction towards the minimum

Batch Gradient Descent

- Move in the direction of steepest descend

Repeat until converge {

$$\boldsymbol{\beta}^{(t+1)} := \boldsymbol{\beta}^{(\mathrm{t})} - \eta \frac{\partial J}{\partial \boldsymbol{\beta}} |_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(\mathrm{t})}}, \qquad \text{e.g., } \eta = 0.01$$

}

Where $J(\boldsymbol{\beta}) = \frac{1}{2} \sum_i \left(\boldsymbol{x}_i^T \boldsymbol{\beta} - y_i\right)^2 / n = \sum_i J_i(\boldsymbol{\beta})/n$ *and*

$$\frac{\partial J}{\partial \boldsymbol{\beta}} = \sum_i \frac{\partial J_i}{\partial \boldsymbol{\beta}} / n = \sum_i \boldsymbol{x}_i \left(\boldsymbol{x}_i^T \boldsymbol{\beta} - y_i\right) / n$$

Stochastic Gradient Descent

- When a new observation, *i*, comes in, update weight immediately (extremely useful for large-scale datasets):

Repeat {

for i=1:n {

$$\boldsymbol{\beta}^{(t+1)} := \boldsymbol{\beta}^{(\mathrm{t})} + \eta(y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}^{(t)})\boldsymbol{x}_i$$

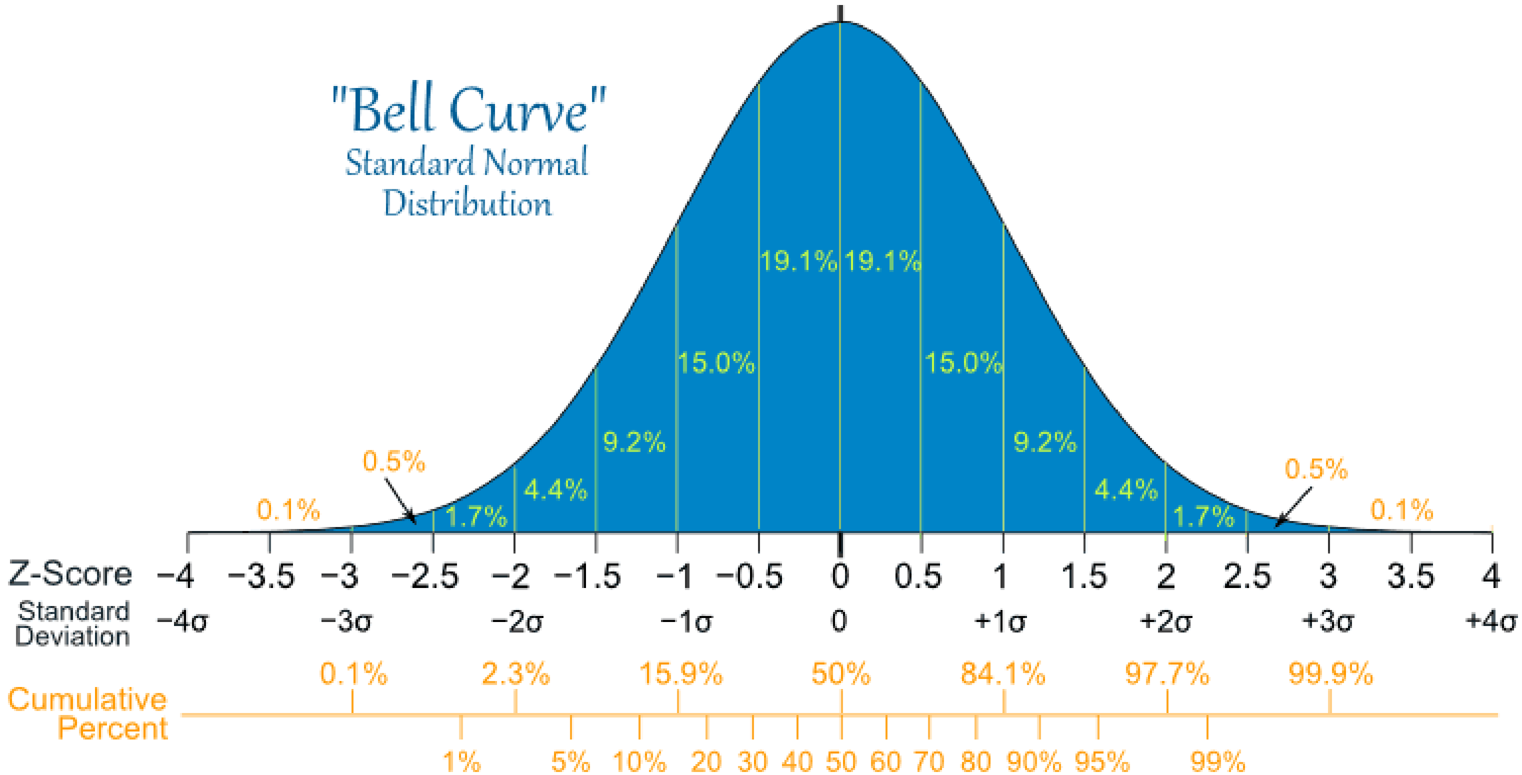
}

}

If the prediction for object *i* is smaller than the real value, $\boldsymbol{\beta}$ should move forward to the direction of $\boldsymbol{x}_i$

Probabilistic Interpretation

- Review of normal distribution
- $X \sim N(\mu, \sigma^2) \Rightarrow f(X = x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$



Probabilistic Interpretation

- Model: $y_i = x_i^T \beta + \varepsilon_i$
  - $\varepsilon_i \sim N(0, \sigma^2)$
  - $y_i | x_i, \beta \sim N(x_i^T \beta, \sigma^2)$
    - $E(y_i | x_i) = x_i^T \beta$
- Likelihood:
  - $L(\boldsymbol{\beta}) = \prod_i p(y_i | x_i, \beta)$

$$= \prod_i \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{\left(y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}\right)^2}{2\sigma^2}\}$$

- Maximum Likelihood Estimation
  - find $\widehat{\boldsymbol{\beta}}$ that maximizes $\mathrm{L}(\boldsymbol{\beta})$
  - $\arg\max L = \arg\min J$, Equivalent to OLS!

Other Practical Issues

- Handle different scales of numerical attributes
  - Z-score: $z = \frac{x-\mu}{\sigma}$
    - x: raw score to be standardized, μ: mean of the population, σ: standard deviation
- What if some attributes are nominal?
  - Set dummy variables
    Type equation here.
    - E.g., $x = 1, if\ sex = F; x = 0, if\ sex = M$
    - Nominal variable with multiple values?
      - Create more dummy variables for one variable
- What if some attribute are ordinal?
  - replace x_{if} by their rank $r_{if} \in \{1, \ldots, M_f\}$
  - map the range of each variable onto [0, 1] by replacing *i*-th object in the *f*-th variable by $z_{if} = \frac{r_{if}-1}{M_f-1}$

Other Practical Issues

- What if $X^T X$ is not invertible?
  - Add a small portion of identity matrix, λI, to it
    - ridge regression or linear regression with l2 norm regularization

$$\sum_i (y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

- What if non-linear correlation exists?
  - Transform features, say, x to x^2

Vector Data: Prediction

- Vector Data
- Linear Regression Model
- Model Evaluation and Selection
- Summary

Model Selection Problem

- Basic problem:
  - how to choose between competing linear regression models

- Model too simple:
  - “underfit” the data; poor predictions; high bias; low variance
- Model too complex:
  - “overfit” the data; poor predictions; low bias; high variance
- Model just right:
  - balance bias and variance to get good predictions

Bias and Variance

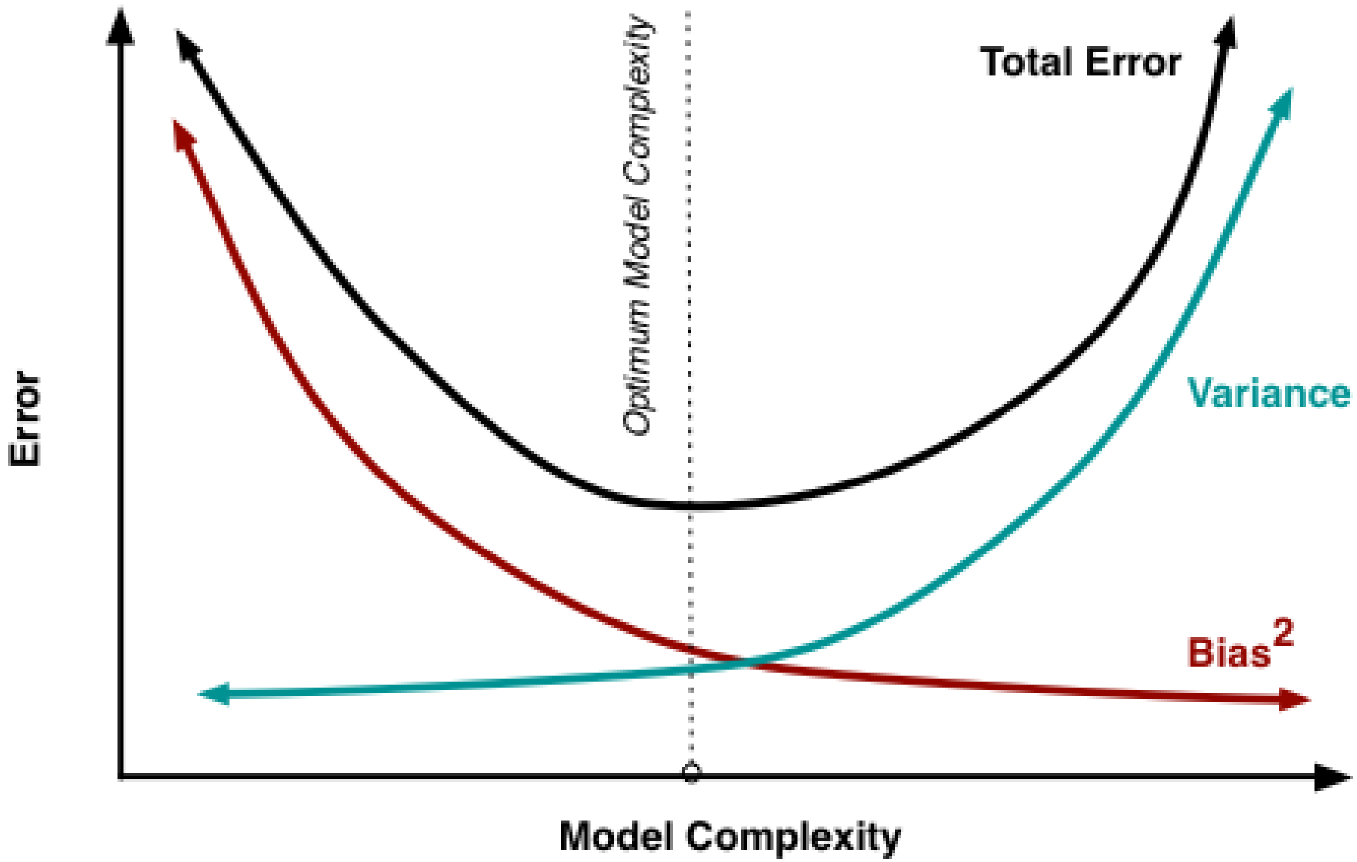
True predictor $f(x)$: $x^T\boldsymbol{\beta}$

Estimated predictor $\hat{f}(x)$: $x^T\widehat{\boldsymbol{\beta}}$

- Bias: $E(\hat{f}(x)) - f(x)$
  - How far away is the expectation of the estimator to the true value? The smaller the better.
- Variance: $Var\left(\hat{f}(x)\right) = E[\left(\hat{f}(x) - E\left(\hat{f}(x)\right)\right)^2]$
  - How variant is the estimator? The smaller the better.
- Reconsider mean square error
  - $J(\widehat{\boldsymbol{\beta}})/n = \sum_i \left(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}} - y_i\right)^2/n$
  - Can be considered as
    - $E[\left(\hat{f}(x) - f(x) - \varepsilon\right)^2] = bias^2 + variance + noise$

Note $E(\varepsilon) = 0, Var(\varepsilon) = \sigma^2$

Bias-Variance Trade-off

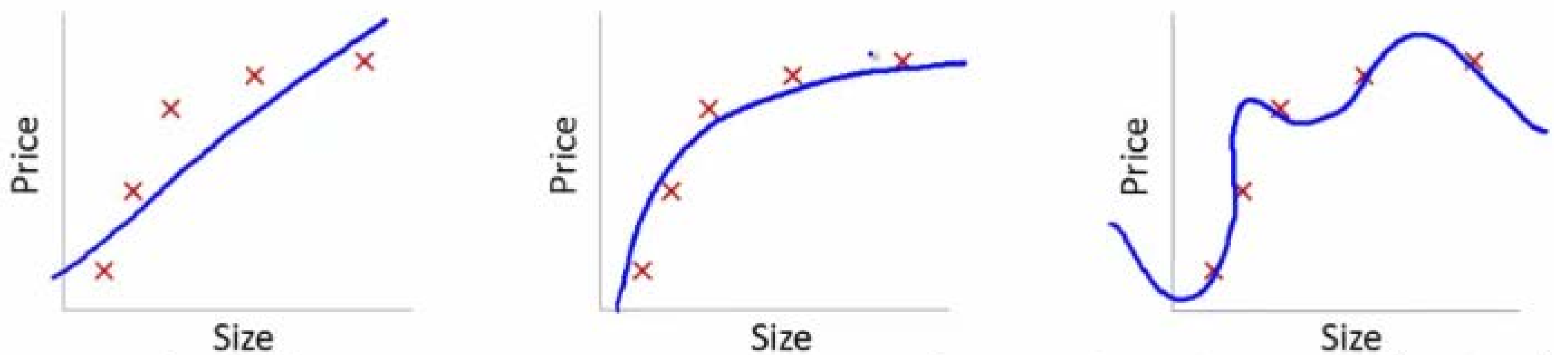


Example: degree d in regression

1. $h_\theta(x) = \theta_0 + \theta_1 x$
2. $h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2$
3. $h_\theta(x) = \theta_0 + \theta_1 x + \cdots + \theta_3 x^3$

$\vdots$

10. $h_\theta(x) = \theta_0 + \theta_1 x + \cdots + \theta_{10} x^{10}$



$\theta_0 + \theta_1 x$	$\theta_0 + \theta_1 x + \theta_2 x^2$	$\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$
High bias (underfit)	"Just right"	High variance (overfit)

http://www.holehouse.org/mlclass/10_Advice_for_applying_machine_learning.html

Example: regularization term in regression

Linear regression with regularization

Model: $h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$

$$J(\theta) = \frac{1}{2m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2 + \frac{\lambda}{2m}\sum_{j=1}^{m}\theta_j^2$$

Large λ | Intermediate λ | Small λ

High bias (underfit) | "Just right" | High variance (overfit)

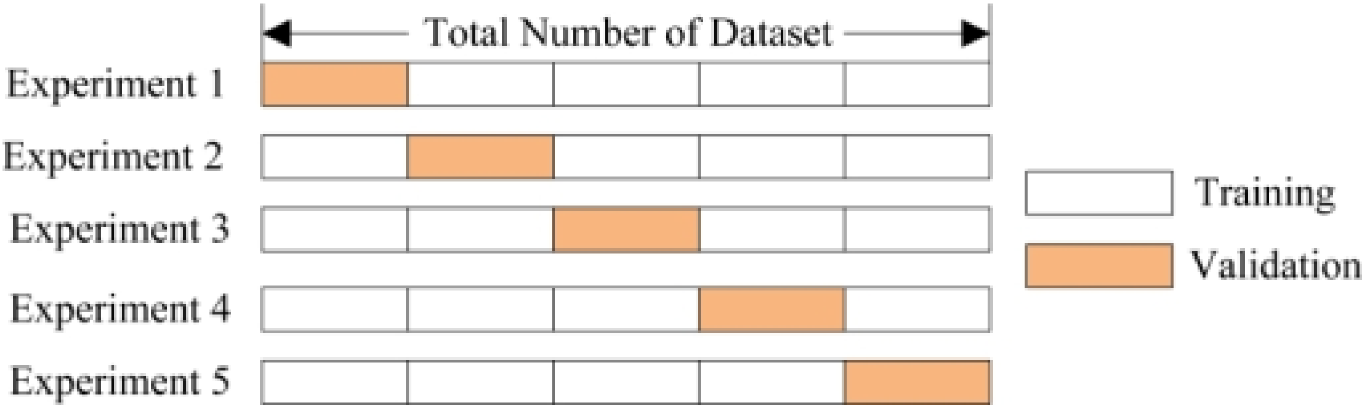
$\lambda = 10000.\ \theta_1 \approx 0, \theta_2 \approx 0, \ldots$

$h_\theta(x) \approx \theta_0$

$\lambda \approx 0$

Cross-Validation

- Partition the data into K folds
  - Use K-1 fold as training, and 1 fold as testing
  - Calculate the average accuracy best on K training-testing pairs
    - Accuracy on validation/test dataset!
      - **Mean square error** can again be used: $\sum_i \left(\boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}} - y_i\right)^2 / n$



AIC & BIC*

- AIC and BIC can be used to test the quality of statistical models
  - AIC (Akaike information criterion)
    - $AIC = 2k - 2\ln(\hat{L})$,
    - where k is the number of parameters in the model and $\hat{L}$ is the likelihood under the estimated parameter
  - BIC (Bayesian Information criterion)
    - B$IC = kln(n) - 2\ln(\hat{L})$,
    - Where n is the number of objects

Stepwise Feature Selection

- Avoid brute-force selection
  - 2^p
- Forward selection
  - Starting with the best single feature
  - Always add the feature that improves the performance best
  - Stop if no feature will further improve the performance
- Backward elimination
  - Start with the full model
  - Always remove the feature that results in the best performance enhancement
  - Stop if removing any feature will get worse performance

Vector Data: Prediction

- Vector Data
- Linear Regression Model
- Model Evaluation and Selection
- Summary

Summary

- What is vector data?
  - Attribute types
  - Basic statistics
  - Visualization
- Linear regression
  - OLS
  - Probabilistic interpretation
- Model Evaluation and Selection
  - Bias-Variance Trade-off
  - Mean square error
  - Cross-validation, AIC, BIC, step-wise feature selection