CS145: INTRODUCTION TO DATA MINING

2: Vector Data: Prediction

Instructor: Yizhou Sun
yzsun@cs.ucla.edu

October 8, 2018
TA Office Hour Time Change

- Junheng Hao: Tuesday 1-3pm
- Yunsheng Bai: Thursday 1-3pm
## Methods to Learn

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

A matrix of $n \times p$:
- $n$ data objects / points
- $p$ attributes / dimensions

<table>
<thead>
<tr>
<th></th>
<th>Sex</th>
<th>Race</th>
<th>Height</th>
<th>Income</th>
<th>Marital Status</th>
<th>Years of Educ.</th>
<th>Liberalness</th>
</tr>
</thead>
<tbody>
<tr>
<td>R1001</td>
<td>M</td>
<td>1</td>
<td>70</td>
<td>50</td>
<td>1</td>
<td>12</td>
<td>1.73</td>
</tr>
<tr>
<td>R1002</td>
<td>M</td>
<td>2</td>
<td>72</td>
<td>100</td>
<td>2</td>
<td>20</td>
<td>4.53</td>
</tr>
<tr>
<td>R1003</td>
<td>F</td>
<td>1</td>
<td>55</td>
<td>250</td>
<td>1</td>
<td>16</td>
<td>2.99</td>
</tr>
<tr>
<td>R1004</td>
<td>M</td>
<td>2</td>
<td>65</td>
<td>20</td>
<td>2</td>
<td>16</td>
<td>1.13</td>
</tr>
<tr>
<td>R1005</td>
<td>F</td>
<td>1</td>
<td>60</td>
<td>10</td>
<td>3</td>
<td>12</td>
<td>3.81</td>
</tr>
<tr>
<td>R1006</td>
<td>M</td>
<td>1</td>
<td>68</td>
<td>30</td>
<td>1</td>
<td>9</td>
<td>4.76</td>
</tr>
<tr>
<td>R1007</td>
<td>F</td>
<td>5</td>
<td>66</td>
<td>25</td>
<td>2</td>
<td>21</td>
<td>2.01</td>
</tr>
<tr>
<td>R1008</td>
<td>F</td>
<td>4</td>
<td>61</td>
<td>43</td>
<td>1</td>
<td>18</td>
<td>1.27</td>
</tr>
<tr>
<td>R1009</td>
<td>M</td>
<td>1</td>
<td>69</td>
<td>67</td>
<td>1</td>
<td>12</td>
<td>3.25</td>
</tr>
</tbody>
</table>
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):**
  Note: $n$ is sample size and $N$ is population size.
  - Weighted arithmetic mean:
  - 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:** \( \text{mean} - \text{mode} = 3 \times (\text{mean} - \text{median}) \)

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

\[
\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}
\]
Symmetric vs. Skewed Data

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

- Symmetric data has a bell-shaped curve where mean, median, and mode are the same.
- Positively skewed data has a tail extending to the right, with mean > median > mode.
- Negatively skewed data has a tail extending to the left, with mean < median < mode.
Measuring the Dispersion of Data

- Quartiles, outliers and boxplots
  - **Quartiles**: Q₁ (25th percentile), Q₃ (75th percentile)
  - **Inter-quartile range**: IQR = Q₃ – Q₁
  - **Five number summary**: min, Q₁, median, Q₃, max
  - **Outlier**: usually, a value higher/lower than 1.5 x IQR of Q₃ or Q₁

- Variance and standard deviation (*sample: s, population: σ*)
  - **Variance**: (algebraic, scalable computation)
    
    \[
    s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{1}{n-1} \left[ \sum_{i=1}^{n} x_i^2 - \frac{1}{n} \left( \sum_{i=1}^{n} x_i \right)^2 \right]
    \]

  - \( \sigma^2 = E[(X - E(X))^2] = E(X^2) - (E(X))^2 \)
  - **Standard deviation** \( s \) (or \( \sigma \)) is the square root of variance \( s^2 \) (or \( \sigma^2 \))
Graphic Displays of Basic Statistical Descriptions

- **Histogram**: x-axis are values, y-axis represent 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 negatively correlated
Uncorrelated Data
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\} $\Rightarrow$ GPA
  - \{income; credit score; profession\} $\Rightarrow$ loan
  - \{college; major; GPA\} $\Rightarrow$ future income
  - ...


### Example of House Price

<table>
<thead>
<tr>
<th>Living Area (sqft)</th>
<th># of Beds</th>
<th>Price (1000$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2104</td>
<td>3</td>
<td>400</td>
</tr>
<tr>
<td>1600</td>
<td>3</td>
<td>330</td>
</tr>
<tr>
<td>2400</td>
<td>3</td>
<td>369</td>
</tr>
<tr>
<td>1416</td>
<td>2</td>
<td>232</td>
</tr>
<tr>
<td>3000</td>
<td>4</td>
<td>540</td>
</tr>
</tbody>
</table>

\[ x = (x_1, x_2)' \]

\[ y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \]
Illustration
Formalization

Data: n independent data objects

- \( y_i, i = 1, ..., n \)
- \( x_i = (x_{i1}, x_{i2}, ..., x_{ip})^T, i = 1, ..., n \)
  - A constant factor is added to model the bias term, i.e., \( x_{i0} = 1 \)
  - New \( x: x_i = (x_{i0}, x_{i1}, x_{i2}, ..., x_{ip})^T \)

Model:

- \( y: \) dependent variable
- \( x: \) explanatory variables
- \( \beta = (\beta_0, \beta_1, ..., \beta_p)^T: \) weight vector
- \( y = x^T \beta = \beta_0 + x_1\beta_1 + x_2\beta_2 + ... + x_p\beta_p \)
A 3-step Process

- Model Construction
  - Use training data to find the best parameter $\beta$, denoted as $\hat{\beta}$

- 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):
    $$\hat{y} = x^T \hat{\beta}$$
Least Square Estimation

• Cost function (Mean Square Error):
  \[ J(\beta) = \frac{1}{2} \sum_i (x_i^T \beta - y_i)^2 / n \]

• Matrix form:
  \[ J(\beta) = (X\beta - y)^T (X\beta - y) / 2n \]
  or \[ ||X\beta - y||^2 / 2n \]

\[
\begin{bmatrix}
1, x_{i1} & \ldots & x_{1f} & \ldots & x_{1p} \\
\vdots & \ddots & \vdots & \ddots & \vdots \\
1, x_{i1} & \ldots & x_{if} & \ldots & x_{ip} \\
\vdots & \ddots & \vdots & \ddots & \vdots \\
1, x_{n1} & \ldots & x_{nf} & \ldots & x_{np}
\end{bmatrix} \begin{bmatrix}
y_1 \\
\vdots \\
y_i \\
\vdots \\
y_n
\end{bmatrix}
\]

\[ X: n \times (p + 1) \text{ matrix} \quad y: n \times 1 \text{ vector} \]
Ordinary Least Squares (OLS)

• Goal: find $\hat{\beta}$ that minimizes $J(\beta)$
  
  • $J(\beta) = \frac{1}{2n} (X\beta - y)^T (X\beta - y)$
    
    $$J(\beta) = \frac{1}{2n} (\beta^T X^T X \beta - y^T X \beta - \beta^T X^T y + y^T y)$$

• Ordinary least squares
  
  • Set first derivative of $J(\beta)$ as 0

  $$\frac{\partial J}{\partial \beta} = (X^T X \beta - X^T y)/n = 0$$

  $$\Rightarrow \hat{\beta} = (X^T X)^{-1} X^T y$$

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 {

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

Where

$$J(\beta) = \frac{1}{2} \sum_i (x_i^T \beta - y_i)^2 / n = \sum_i J_i(\beta) / n$$

and

$$\frac{\partial J}{\partial \beta} = \sum_i \frac{\partial J_i}{\partial \beta} / n = \sum_i x_i (x_i^T \beta - y_i) / n$$
Stochastic Gradient Descent

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

Repeat {
  for \( i=1:n \) {
    \[ \beta^{(t+1)} := \beta^{(t)} + \eta (y_i - x_i^T \beta^{(t)}) x_i \]
  }
}

If the prediction for object \( i \) is smaller than the real value, \( \beta \) should move forward to the direction of \( x_i \)
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:
  - \( L(\beta) = \prod_i p(y_i | x_i, \beta) \)
    \[ = \prod_i \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y_i - x_i^T \beta)^2}{2\sigma^2}\right\} \]
- Maximum Likelihood Estimation
  - find \( \hat{\beta} \) that maximizes \( L(\beta) \)
  - \( \arg \max L = \arg \min J \), 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, \( \mu \): mean of the population, \( \sigma \): standard deviation

• What if some attributes are nominal?
  • Set dummy variables
    • 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 $X^TX$ is not invertible?
  • Add a small portion of identity matrix, $\lambda I$, to it
    • ridge regression or linear regression with l2 norm
      $$\sum_{i} (y_i - x_i^T \beta)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

• 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 \( f(x) : x^T \beta \)

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

• 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: \( Var(\hat{f}(x)) = E[(\hat{f}(x) - E(\hat{f}(x)))^2] \)

• How variant is the estimator? The smaller the better.

• Reconsider mean square error

\[ J(\hat{\beta})/n = \sum_i (x_i^T \hat{\beta} - y_i)^2 / n \]

• Can be considered as

\[ E[(\hat{f}(x) - f(x) - \varepsilon)^2] = bias^2 + variance + noise \]

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

http://www.holehouse.org/mlclass/10_Advice_for_applying_machine_learning.html
Example: regularization term in regression

Linear regression with regularization

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

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

Price

Size

Large \( \lambda \)
High bias (underfit)
\( \lambda = 10000 \), \( \theta_1 \approx 0, \theta_2 \approx 0, \ldots \)
\( h_\theta(x) \approx \theta_0 \)

Intermediate \( \lambda \)
“Just right”

Small \( \lambda \)
High variance (overfit)
\( \lambda \approx 0 \)
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: \( \sum_i (x_i^T \hat{\beta} - y_i)^2 / n \)
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)**
  
  • \( BIC = kln(n) - 2\ln(\hat{L}) \),
  
  • 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