Lecture 2: Regression

Jan-Willem van de Meent

(credit: Yijun Zhao, Marc Toussaint, Bishop)
Administrativa

Instructor

Jan-Willem van de Meent

Email: j.vandemeent@northeastern.edu
Phone: +1 617 373-7696
Office Hours: 478 WVH, Wed 1.30pm - 2.30pm

Teaching Assistants

Yuan Zhong

E-mail: yzhong@ccs.neu.edu
Office Hours: WVH 462, Wed 3pm - 5pm

Kamlendra Kumar

E-mail: kumark@zimbra.ccs.neu.edu
Office Hours: WVH 462, Fri 3pm - 5pm
Administrativa

Course Website
http://www.ccs.neu.edu/course/cs6220f16/sec3/

Piazza
https://piazza.com/northeastern/fall2016/cs622003/home

Project Guidelines (Vote next week)
http://www.ccs.neu.edu/course/cs6220f16/sec3/project/
Question
What would you like to get out of this course?
Linear Regression
Regression Examples

- Features: \{age, major, gender, race\} \rightarrow GPA
- Features: \{income, credit score, profession\} \rightarrow Loan Amount
- Features: \{college, major, GPA\} \rightarrow Future Income
Example: Boston Housing Data

<table>
<thead>
<tr>
<th>CRIM</th>
<th>ZN</th>
<th>INDUS</th>
<th>CHAS</th>
<th>NOX</th>
<th>RM</th>
<th>AGE</th>
<th>DIS</th>
<th>RAD</th>
<th>TAX</th>
<th>PRTARIO</th>
<th>B</th>
<th>LSTAT</th>
<th>MEDV</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00632</td>
<td>18.0</td>
<td>2.31</td>
<td>0.0</td>
<td>0.538</td>
<td>6.575</td>
<td>65.2</td>
<td>4.0900</td>
<td>1.0</td>
<td>296.0</td>
<td>15.3</td>
<td>396.90</td>
<td>4.98</td>
<td>24.0</td>
</tr>
<tr>
<td>0.02731</td>
<td>0.0</td>
<td>7.07</td>
<td>0.0</td>
<td>0.469</td>
<td>6.421</td>
<td>78.9</td>
<td>4.9671</td>
<td>2.0</td>
<td>242.0</td>
<td>17.8</td>
<td>396.90</td>
<td>9.14</td>
<td>21.6</td>
</tr>
<tr>
<td>0.02729</td>
<td>0.0</td>
<td>7.07</td>
<td>0.0</td>
<td>0.469</td>
<td>7.185</td>
<td>61.1</td>
<td>4.9671</td>
<td>2.0</td>
<td>242.0</td>
<td>17.8</td>
<td>392.83</td>
<td>4.03</td>
<td>34.7</td>
</tr>
<tr>
<td>0.03237</td>
<td>0.0</td>
<td>2.18</td>
<td>0.0</td>
<td>0.458</td>
<td>6.998</td>
<td>45.8</td>
<td>6.0622</td>
<td>3.0</td>
<td>222.0</td>
<td>18.7</td>
<td>394.63</td>
<td>2.94</td>
<td>33.4</td>
</tr>
<tr>
<td>0.06905</td>
<td>0.0</td>
<td>2.18</td>
<td>0.0</td>
<td>0.458</td>
<td>7.147</td>
<td>54.2</td>
<td>6.0622</td>
<td>3.0</td>
<td>222.0</td>
<td>18.7</td>
<td>396.90</td>
<td>5.33</td>
<td>36.2</td>
</tr>
<tr>
<td>0.02985</td>
<td>0.0</td>
<td>2.18</td>
<td>0.0</td>
<td>0.458</td>
<td>6.430</td>
<td>58.7</td>
<td>6.0622</td>
<td>3.0</td>
<td>222.0</td>
<td>18.7</td>
<td>394.12</td>
<td>5.21</td>
<td>28.7</td>
</tr>
<tr>
<td>0.08829</td>
<td>12.5</td>
<td>7.87</td>
<td>0.0</td>
<td>0.524</td>
<td>6.012</td>
<td>66.6</td>
<td>5.5605</td>
<td>5.0</td>
<td>311.0</td>
<td>15.2</td>
<td>395.60</td>
<td>12.43</td>
<td>22.9</td>
</tr>
<tr>
<td>0.14455</td>
<td>12.5</td>
<td>7.87</td>
<td>0.0</td>
<td>0.524</td>
<td>6.172</td>
<td>96.1</td>
<td>5.9505</td>
<td>5.0</td>
<td>311.0</td>
<td>15.2</td>
<td>396.90</td>
<td>19.15</td>
<td>27.1</td>
</tr>
<tr>
<td>0.21124</td>
<td>12.5</td>
<td>7.87</td>
<td>0.0</td>
<td>0.524</td>
<td>5.631</td>
<td>100.0</td>
<td>6.0821</td>
<td>5.0</td>
<td>311.0</td>
<td>15.2</td>
<td>386.63</td>
<td>29.93</td>
<td>16.5</td>
</tr>
<tr>
<td>0.17004</td>
<td>12.5</td>
<td>7.87</td>
<td>0.0</td>
<td>0.524</td>
<td>6.004</td>
<td>85.9</td>
<td>6.5921</td>
<td>5.0</td>
<td>311.0</td>
<td>15.2</td>
<td>386.71</td>
<td>17.10</td>
<td>18.9</td>
</tr>
<tr>
<td>0.22489</td>
<td>12.5</td>
<td>7.87</td>
<td>0.0</td>
<td>0.524</td>
<td>6.377</td>
<td>94.3</td>
<td>6.3467</td>
<td>5.0</td>
<td>311.0</td>
<td>15.2</td>
<td>392.52</td>
<td>20.45</td>
<td>15.0</td>
</tr>
</tbody>
</table>

UC Irvine Machine Learning Repository

*(good source for project datasets)*

https://archive.ics.uci.edu/ml/datasets/Housing
Example: Boston Housing Data

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 african americans by town
13. **LSTAT**: % lower status of the population
14. **MEDV**: Median value of owner-occupied homes in $1000's
Example: Boston Housing Data

<table>
<thead>
<tr>
<th></th>
<th>CRIM</th>
<th>ZN</th>
<th>INDUS</th>
<th>CHAS</th>
<th>NOX</th>
<th>RM</th>
<th>AGE</th>
<th>DIS</th>
<th>RAD</th>
<th>TAX</th>
<th>PTRATIO</th>
<th>B</th>
<th>LSTAT</th>
<th>MEDV</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.00632</td>
<td>18.0</td>
<td>2.31</td>
<td>0.0</td>
<td>0.538</td>
<td>6.575</td>
<td>65.2</td>
<td>4.0900</td>
<td>1.0</td>
<td>296.0</td>
<td>15.3</td>
<td>396.90</td>
<td>4.98</td>
<td>24.0</td>
</tr>
<tr>
<td>1</td>
<td>0.02731</td>
<td>0.0</td>
<td>7.07</td>
<td>0.0</td>
<td>0.469</td>
<td>6.421</td>
<td>78.9</td>
<td>4.9671</td>
<td>2.0</td>
<td>242.0</td>
<td>17.8</td>
<td>396.90</td>
<td>9.14</td>
<td>21.6</td>
</tr>
<tr>
<td>2</td>
<td>0.02729</td>
<td>0.0</td>
<td>7.07</td>
<td>0.0</td>
<td>0.469</td>
<td>7.185</td>
<td>61.1</td>
<td>4.9671</td>
<td>2.0</td>
<td>242.0</td>
<td>17.8</td>
<td>392.83</td>
<td>4.03</td>
<td>34.7</td>
</tr>
<tr>
<td>3</td>
<td>0.03237</td>
<td>0.0</td>
<td>2.18</td>
<td>0.0</td>
<td>0.458</td>
<td>6.998</td>
<td>45.8</td>
<td>6.0622</td>
<td>3.0</td>
<td>222.0</td>
<td>18.7</td>
<td>394.63</td>
<td>2.94</td>
<td>33.4</td>
</tr>
<tr>
<td>4</td>
<td>0.06905</td>
<td>0.0</td>
<td>2.18</td>
<td>0.0</td>
<td>0.458</td>
<td>7.147</td>
<td>54.2</td>
<td>6.0622</td>
<td>3.0</td>
<td>222.0</td>
<td>18.7</td>
<td>396.90</td>
<td>5.33</td>
<td>36.2</td>
</tr>
<tr>
<td>5</td>
<td>0.02985</td>
<td>0.0</td>
<td>2.18</td>
<td>0.0</td>
<td>0.458</td>
<td>6.430</td>
<td>58.7</td>
<td>6.0622</td>
<td>3.0</td>
<td>222.0</td>
<td>18.7</td>
<td>394.12</td>
<td>5.21</td>
<td>28.7</td>
</tr>
<tr>
<td>6</td>
<td>0.08829</td>
<td>12.5</td>
<td>7.87</td>
<td>0.0</td>
<td>0.524</td>
<td>6.012</td>
<td>66.6</td>
<td>5.5605</td>
<td>5.0</td>
<td>311.0</td>
<td>15.2</td>
<td>395.60</td>
<td>12.43</td>
<td>22.9</td>
</tr>
<tr>
<td>7</td>
<td>0.14455</td>
<td>12.5</td>
<td>7.87</td>
<td>0.0</td>
<td>0.524</td>
<td>6.172</td>
<td>96.1</td>
<td>5.9505</td>
<td>5.0</td>
<td>311.0</td>
<td>15.2</td>
<td>396.90</td>
<td>19.15</td>
<td>27.1</td>
</tr>
<tr>
<td>8</td>
<td>0.21124</td>
<td>12.5</td>
<td>7.87</td>
<td>0.0</td>
<td>0.524</td>
<td>5.631</td>
<td>100.0</td>
<td>6.0821</td>
<td>5.0</td>
<td>311.0</td>
<td>15.2</td>
<td>386.63</td>
<td>29.93</td>
<td>16.5</td>
</tr>
<tr>
<td>9</td>
<td>0.17004</td>
<td>12.5</td>
<td>7.87</td>
<td>0.0</td>
<td>0.524</td>
<td>6.004</td>
<td>85.9</td>
<td>6.5921</td>
<td>5.0</td>
<td>311.0</td>
<td>15.2</td>
<td>386.71</td>
<td>17.10</td>
<td>18.9</td>
</tr>
<tr>
<td>10</td>
<td>0.22489</td>
<td>12.5</td>
<td>7.87</td>
<td>0.0</td>
<td>0.524</td>
<td>6.377</td>
<td>94.3</td>
<td>6.3467</td>
<td>5.0</td>
<td>311.0</td>
<td>15.2</td>
<td>392.52</td>
<td>20.45</td>
<td>15.0</td>
</tr>
</tbody>
</table>

**CRIM**: per capita crime rate by town
Example: Boston Housing Data

<table>
<thead>
<tr>
<th>CRIM</th>
<th>ZN</th>
<th>INDUS</th>
<th>CHAS</th>
<th>NOX</th>
<th>RM</th>
<th>AGE</th>
<th>DIS</th>
<th>RAD</th>
<th>TAX</th>
<th>PRTATIO</th>
<th>B</th>
<th>LSTAT</th>
<th>MEDV</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00632</td>
<td>18.0</td>
<td>2.31</td>
<td>0.0</td>
<td>0.538</td>
<td>6.575</td>
<td>65.2</td>
<td>4.0900</td>
<td>1.0</td>
<td>296.0</td>
<td>15.3</td>
<td>396.90</td>
<td>4.98</td>
<td>24.0</td>
</tr>
<tr>
<td>0.02731</td>
<td>0.0</td>
<td>7.07</td>
<td>0.0</td>
<td>0.469</td>
<td>6.421</td>
<td>78.9</td>
<td>4.9671</td>
<td>2.0</td>
<td>242.0</td>
<td>17.8</td>
<td>396.90</td>
<td>9.14</td>
<td>21.6</td>
</tr>
<tr>
<td>0.02729</td>
<td>0.0</td>
<td>7.07</td>
<td>0.0</td>
<td>0.469</td>
<td>7.185</td>
<td>61.1</td>
<td>4.9671</td>
<td>2.0</td>
<td>242.0</td>
<td>17.8</td>
<td>392.83</td>
<td>4.03</td>
<td>34.7</td>
</tr>
<tr>
<td>0.03237</td>
<td>0.0</td>
<td>2.18</td>
<td>0.0</td>
<td>0.458</td>
<td>6.998</td>
<td>45.8</td>
<td>6.0622</td>
<td>3.0</td>
<td>222.0</td>
<td>18.7</td>
<td>394.63</td>
<td>2.94</td>
<td>33.4</td>
</tr>
<tr>
<td>0.06905</td>
<td>0.0</td>
<td>2.18</td>
<td>0.0</td>
<td>0.458</td>
<td>7.147</td>
<td>54.2</td>
<td>6.0622</td>
<td>3.0</td>
<td>222.0</td>
<td>18.7</td>
<td>396.90</td>
<td>5.33</td>
<td>36.2</td>
</tr>
<tr>
<td>0.02985</td>
<td>0.0</td>
<td>2.18</td>
<td>0.0</td>
<td>0.458</td>
<td>6.430</td>
<td>58.7</td>
<td>6.0622</td>
<td>3.0</td>
<td>222.0</td>
<td>18.7</td>
<td>394.12</td>
<td>5.21</td>
<td>28.7</td>
</tr>
<tr>
<td>0.08829</td>
<td>12.5</td>
<td>7.87</td>
<td>0.0</td>
<td>0.524</td>
<td>6.012</td>
<td>66.6</td>
<td>5.5605</td>
<td>5.0</td>
<td>311.0</td>
<td>15.2</td>
<td>395.60</td>
<td>12.43</td>
<td>22.9</td>
</tr>
<tr>
<td>0.14455</td>
<td>12.5</td>
<td>7.87</td>
<td>0.0</td>
<td>0.524</td>
<td>6.172</td>
<td>96.1</td>
<td>5.9505</td>
<td>5.0</td>
<td>311.0</td>
<td>15.2</td>
<td>396.90</td>
<td>19.15</td>
<td>27.1</td>
</tr>
<tr>
<td>0.21124</td>
<td>12.5</td>
<td>7.87</td>
<td>0.0</td>
<td>0.524</td>
<td>5.631</td>
<td>100.0</td>
<td>6.0821</td>
<td>5.0</td>
<td>311.0</td>
<td>15.2</td>
<td>386.63</td>
<td>29.93</td>
<td>16.5</td>
</tr>
<tr>
<td>0.17004</td>
<td>12.5</td>
<td>7.87</td>
<td>0.0</td>
<td>0.524</td>
<td>6.004</td>
<td>85.9</td>
<td>6.5921</td>
<td>5.0</td>
<td>311.0</td>
<td>15.2</td>
<td>386.71</td>
<td>17.10</td>
<td>18.9</td>
</tr>
<tr>
<td>0.22489</td>
<td>12.5</td>
<td>7.87</td>
<td>0.0</td>
<td>0.524</td>
<td>6.377</td>
<td>94.3</td>
<td>6.3467</td>
<td>5.0</td>
<td>311.0</td>
<td>15.2</td>
<td>392.52</td>
<td>20.45</td>
<td>15.0</td>
</tr>
</tbody>
</table>

**CHAS**: Charles River dummy variable
(= 1 if tract bounds river; 0 otherwise)
Example: Boston Housing Data

<table>
<thead>
<tr>
<th>CRIM</th>
<th>ZN</th>
<th>INDUS</th>
<th>CHAS</th>
<th>NOX</th>
<th>RM</th>
<th>AGE</th>
<th>DIS</th>
<th>RAD</th>
<th>TAX</th>
<th>PTRATIO</th>
<th>B</th>
<th>LSTAT</th>
<th>MEDV</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0632</td>
<td>18.0</td>
<td>2.31</td>
<td>0.0</td>
<td>0.538</td>
<td>6.575</td>
<td>65.2</td>
<td>4.0900</td>
<td>1.0</td>
<td>296.0</td>
<td>15.3</td>
<td>396.90</td>
<td>4.98</td>
<td>24.0</td>
</tr>
<tr>
<td>0.02731</td>
<td>0.0</td>
<td>7.07</td>
<td>0.0</td>
<td>0.469</td>
<td>6.421</td>
<td>78.9</td>
<td>4.9671</td>
<td>2.0</td>
<td>242.0</td>
<td>17.8</td>
<td>396.90</td>
<td>9.14</td>
<td>21.6</td>
</tr>
<tr>
<td>0.02729</td>
<td>0.0</td>
<td>7.07</td>
<td>0.0</td>
<td>0.469</td>
<td>7.185</td>
<td>61.1</td>
<td>4.9671</td>
<td>2.0</td>
<td>242.0</td>
<td>17.8</td>
<td>392.83</td>
<td>4.03</td>
<td>34.7</td>
</tr>
<tr>
<td>0.03237</td>
<td>0.0</td>
<td>2.18</td>
<td>0.0</td>
<td>0.458</td>
<td>6.998</td>
<td>45.8</td>
<td>6.0622</td>
<td>3.0</td>
<td>222.0</td>
<td>18.7</td>
<td>394.63</td>
<td>2.94</td>
<td>33.4</td>
</tr>
<tr>
<td>0.06905</td>
<td>0.0</td>
<td>2.18</td>
<td>0.0</td>
<td>0.458</td>
<td>7.147</td>
<td>54.2</td>
<td>6.0622</td>
<td>3.0</td>
<td>222.0</td>
<td>18.7</td>
<td>396.90</td>
<td>5.33</td>
<td>36.2</td>
</tr>
<tr>
<td>0.02985</td>
<td>0.0</td>
<td>2.18</td>
<td>0.0</td>
<td>0.458</td>
<td>6.430</td>
<td>58.7</td>
<td>6.0622</td>
<td>3.0</td>
<td>222.0</td>
<td>18.7</td>
<td>394.12</td>
<td>5.21</td>
<td>28.7</td>
</tr>
<tr>
<td>0.08829</td>
<td>12.5</td>
<td>7.87</td>
<td>0.0</td>
<td>0.524</td>
<td>6.012</td>
<td>66.6</td>
<td>5.5605</td>
<td>5.0</td>
<td>311.0</td>
<td>15.2</td>
<td>395.60</td>
<td>12.43</td>
<td>22.9</td>
</tr>
<tr>
<td>0.14455</td>
<td>12.5</td>
<td>7.87</td>
<td>0.0</td>
<td>0.524</td>
<td>6.172</td>
<td>96.1</td>
<td>5.9505</td>
<td>5.0</td>
<td>311.0</td>
<td>15.2</td>
<td>396.90</td>
<td>19.15</td>
<td>27.1</td>
</tr>
<tr>
<td>0.21124</td>
<td>12.5</td>
<td>7.87</td>
<td>0.0</td>
<td>0.524</td>
<td>5.631</td>
<td>100.0</td>
<td>6.0821</td>
<td>5.0</td>
<td>311.0</td>
<td>15.2</td>
<td>386.63</td>
<td>29.93</td>
<td>16.5</td>
</tr>
<tr>
<td>0.17004</td>
<td>12.5</td>
<td>7.87</td>
<td>0.0</td>
<td>0.524</td>
<td>6.004</td>
<td>85.9</td>
<td>6.5921</td>
<td>5.0</td>
<td>311.0</td>
<td>15.2</td>
<td>386.71</td>
<td>17.10</td>
<td>18.9</td>
</tr>
<tr>
<td>0.22489</td>
<td>12.5</td>
<td>7.87</td>
<td>0.0</td>
<td>0.524</td>
<td>6.377</td>
<td>94.3</td>
<td>6.3467</td>
<td>5.0</td>
<td>311.0</td>
<td>15.2</td>
<td>392.52</td>
<td>20.45</td>
<td>15.0</td>
</tr>
</tbody>
</table>

**MEDV**: Median value of owner-occupied homes in $1000's
Example: Boston Housing Data

<table>
<thead>
<tr>
<th>CRIM</th>
<th>ZN</th>
<th>INDUS</th>
<th>CHAS</th>
<th>NOX</th>
<th>RM</th>
<th>AGE</th>
<th>DIS</th>
<th>RAD</th>
<th>TAX</th>
<th>PTRATIO</th>
<th>B</th>
<th>LSTAT</th>
<th>MEDV</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00632</td>
<td>18.0</td>
<td>2.31</td>
<td>0.0</td>
<td>0.538</td>
<td>6.575</td>
<td>65.2</td>
<td>4.0900</td>
<td>1.0</td>
<td>296.0</td>
<td>15.3</td>
<td>396.90</td>
<td>4.98</td>
<td>24.0</td>
</tr>
<tr>
<td>0.02731</td>
<td>0.0</td>
<td>7.07</td>
<td>0.0</td>
<td>0.469</td>
<td>6.421</td>
<td>78.9</td>
<td>4.9671</td>
<td>2.0</td>
<td>242.0</td>
<td>17.8</td>
<td>396.90</td>
<td>9.14</td>
<td>21.6</td>
</tr>
<tr>
<td>0.02729</td>
<td>0.0</td>
<td>7.07</td>
<td>0.0</td>
<td>0.469</td>
<td>7.185</td>
<td>61.1</td>
<td>4.9671</td>
<td>2.0</td>
<td>242.0</td>
<td>17.8</td>
<td>392.83</td>
<td>4.03</td>
<td>34.7</td>
</tr>
<tr>
<td>0.03237</td>
<td>0.0</td>
<td>2.18</td>
<td>0.0</td>
<td>0.458</td>
<td>6.998</td>
<td>45.8</td>
<td>6.0622</td>
<td>3.0</td>
<td>222.0</td>
<td>18.7</td>
<td>394.63</td>
<td>2.94</td>
<td>33.4</td>
</tr>
<tr>
<td>0.06905</td>
<td>0.0</td>
<td>2.18</td>
<td>0.0</td>
<td>0.458</td>
<td>7.147</td>
<td>54.2</td>
<td>6.0622</td>
<td>3.0</td>
<td>222.0</td>
<td>18.7</td>
<td>396.90</td>
<td>5.33</td>
<td>36.2</td>
</tr>
<tr>
<td>0.02985</td>
<td>0.0</td>
<td>2.18</td>
<td>0.0</td>
<td>0.458</td>
<td>6.430</td>
<td>58.7</td>
<td>6.0622</td>
<td>3.0</td>
<td>222.0</td>
<td>18.7</td>
<td>394.12</td>
<td>5.21</td>
<td>28.7</td>
</tr>
<tr>
<td>0.08829</td>
<td>12.5</td>
<td>7.87</td>
<td>0.0</td>
<td>0.524</td>
<td>6.012</td>
<td>66.6</td>
<td>5.5605</td>
<td>5.0</td>
<td>311.0</td>
<td>15.2</td>
<td>395.60</td>
<td>12.43</td>
<td>22.9</td>
</tr>
<tr>
<td>0.14455</td>
<td>12.5</td>
<td>7.87</td>
<td>0.0</td>
<td>0.524</td>
<td>6.172</td>
<td>96.1</td>
<td>5.9505</td>
<td>5.0</td>
<td>311.0</td>
<td>15.2</td>
<td>396.90</td>
<td>19.15</td>
<td>27.1</td>
</tr>
<tr>
<td>0.21124</td>
<td>12.5</td>
<td>7.87</td>
<td>0.0</td>
<td>0.524</td>
<td>5.631</td>
<td>100.0</td>
<td>6.0821</td>
<td>5.0</td>
<td>311.0</td>
<td>15.2</td>
<td>386.63</td>
<td>29.93</td>
<td>16.5</td>
</tr>
<tr>
<td>0.17004</td>
<td>12.5</td>
<td>7.87</td>
<td>0.0</td>
<td>0.524</td>
<td>6.004</td>
<td>85.9</td>
<td>6.5921</td>
<td>5.0</td>
<td>311.0</td>
<td>15.2</td>
<td>386.71</td>
<td>17.10</td>
<td>18.9</td>
</tr>
<tr>
<td>0.22489</td>
<td>12.5</td>
<td>7.87</td>
<td>0.0</td>
<td>0.524</td>
<td>6.377</td>
<td>94.3</td>
<td>6.3467</td>
<td>5.0</td>
<td>311.0</td>
<td>15.2</td>
<td>392.52</td>
<td>20.45</td>
<td>15.0</td>
</tr>
</tbody>
</table>

$N$ data points

$D$ features
Regression: Problem Setup

Given $N$ observations

\[ \{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\} \]

learn a function

\[ y_i = f(x_i) \quad \forall i = 1, 2, \ldots, N \]

and for a new input $x^*$ predict

\[ y^* = f(x^*) \]
Assume $f$ is a linear combination of $D$ features

$$y = w_0 + w_1 x_1 + \ldots + w_D x_D = \mathbf{w}^\top \mathbf{x}$$

were $\mathbf{x}$ and $\mathbf{w}$ are defined as

$$\mathbf{x} = (1, x_1, \ldots, x_D) \quad \mathbf{w} = (w_0, w_1, \ldots, w_D)$$

for $N$ points we write

$$y = X \mathbf{w} \quad y = (y_1, \ldots, y_N) \quad X = (\mathbf{x}_1^\top, \ldots, \mathbf{x}_N^\top)$$

**Learning task:** Estimate $\mathbf{w}$
Linear Regression

Figure: 1D and 2D linear regression
Error Measure

Mean Squared Error (MSE):

$$E(w) = \frac{1}{N} \sum_{n=1}^{N} (w^T x_n - y_n)^2$$

$$= \frac{1}{N} \| Xw - y \|^2$$

where

$$X = \begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_N^T \end{bmatrix} \quad y = \begin{bmatrix} y_1^T \\ y_2^T \\ \vdots \\ y_N^T \end{bmatrix}$$
Minimizing the Error

\[ E(w) = \frac{1}{N} \| Xw - y \|^2 \]

\[ \nabla E(w) = \frac{2}{N} X^T(Xw - y) = 0 \]

\[ X^TXw = X^Ty \]

\[ w = X^\dagger y \]

where \( X^\dagger = (X^TX)^{-1}X^T \) is the 'pseudo-inverse' of \( X \).
Minimizing the Error

\[ E(w) = \frac{1}{N} \| Xw - y \|^2 \]

\[ \nabla E(w) = \frac{2}{N} X^T(Xw - y) = 0 \]

\[ X^TXw = X^Ty \]

\[ w = X^\dagger y \]

where \( X^\dagger = (X^TX)^{-1}X^T \) is the 'pseudo-inverse' of \( X \)

Matrix Cookbook (on course website)
Ordinary Least Squares

- Construct matrix $\mathbf{X}$ and the vector $\mathbf{y}$ from the dataset $\{(x_1, y_1), x_2, y_2), \ldots, (x_N, y_N)\}$ (each $x$ includes $x_0 = 1$) as follows:

$$
\mathbf{X} = \begin{bmatrix}
-x_1^T & -x_2^T & \cdots & -x_N^T
\end{bmatrix}
$$

$$
\mathbf{y} = \begin{bmatrix}
y_1^T \\
y_2^T \\
\vdots \\
y_N^T
\end{bmatrix}
$$

- Compute $\mathbf{X}^\dagger = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$

- Return $\mathbf{w} = \mathbf{X}^\dagger \mathbf{y}$
Gradient Descent

The ellipses shown above are the contours of a quadratic function. Also shown is the trajectory taken by gradient descent, which was initialized at $(48,30)$. The $x$'s in the figure (joined by straight lines) mark the successive values of $\theta$ that gradient descent went through.

When we run batch gradient descent to fit $\theta$ on our previous dataset, to learn to predict housing price as a function of living area, we obtain $\theta_0 = 71.27, \theta_1 = 0.1345$. If we plot $h_{\theta}(x)$ as a function of $x$ (area), along with the training data, we obtain the following figure:

If the number of bedrooms were included as one of the input features as well, we get $\theta_0 = 89.60, \theta_1 = 0.1392, \theta_2 = -8.738$. The above results were obtained with batch gradient descent. There is an alternative to batch gradient descent that also works very well. Consider the following algorithm:

\[ w \]
Least Mean Squares
(a.k.a. gradient descent)

- Initialize the $w(0)$ for time $t = 0$
- for $t = 0, 1, 2, \ldots$ do
  - Compute the gradient $g_t = \nabla E(w(t))$
  - Set the direction to move, $v_t = -g_t$
  - Update $w(t + 1) = w(t) + \eta v_t$
- Iterate until it is time to stop
- Return the final weights $w$
Question
When would you want to use OLS, when LMS?
Computational Complexity

**Ordinary least squares (OMS)**

\[ w = (X^\top X)^{-1}(X^\top y) \]

\[
\begin{align*}
(X^\top y) & \quad O(DN) \\
(X^\top X) & \quad O(D^2N) \\
(X^\top X)^{-1} & \quad O(D^3)
\end{align*}
\]

**Least Mean Squares (LMS)**

\[ \nabla E(w) = \frac{2}{N} X^\top (Xw - y) \]

\[
\begin{align*}
Xw & \quad O(DN) \\
X(w - y) & \quad O(N) \\
X^\top (Xw - y) & \quad O(DN)
\end{align*}
\]
# Computational Complexity

<table>
<thead>
<tr>
<th>Ordinary least squares (OMS)</th>
<th>Least Mean Squares (LMS)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mathbf{w} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}(\mathbf{X}^\mathsf{T}\mathbf{y})$</td>
<td>$\nabla E(\mathbf{w}) = \frac{2}{N}\mathbf{X}^\mathsf{T}(\mathbf{X}\mathbf{w} - \mathbf{y})$</td>
</tr>
<tr>
<td>$\mathbf{X}^\mathsf{T}\mathbf{y}$</td>
<td>$\mathbf{X}\mathbf{w}$</td>
</tr>
<tr>
<td>$\mathbf{X}^\mathsf{T}\mathbf{X}$</td>
<td>$\mathbf{X}(\mathbf{w} - \mathbf{y})$</td>
</tr>
<tr>
<td>$(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}$</td>
<td>$\mathbf{X}^\mathsf{T}(\mathbf{X}\mathbf{w} - \mathbf{y})$</td>
</tr>
</tbody>
</table>

OMS is expensive when D is large
Effect of step size

- Use 0.1 (practical observation)
- Use variable size: $\eta = k5^E_k$

Yijun Zhao
Linear Regression
Choosing Stepsize

Set step size proportional to $\nabla f(x)$?

- Large gradient → Large step?
- Small gradient → Small step?

NO!

We need methods independent of $|\nabla f(x)|$, invariant of scaling of $f$ and $x$!
Choosing Stepsize

Set step size proportional to $\nabla f(x)$?

NO!

We need methods indep. of $|\nabla f(x)|$, invariant of scaling of $f$ and $x$!

Two commonly used techniques
1. Stepsize adaptation
2. Line search
Stepsize Adaptation

Input: initial $x \in \mathbb{R}^n$, functions $f(x)$ and $\nabla f(x)$, initial stepsize $\alpha$, tolerance $\theta$

Output: $x$

1: repeat
2: $y \leftarrow x - \alpha \frac{\nabla f(x)}{|\nabla f(x)|}$
3: if [then step is accepted] $f(y) \leq f(x)$
4: $x \leftarrow y$
5: $\alpha \leftarrow 1.2\alpha$  \hspace{1cm} // increase stepsize
6: else [step is rejected]
7: $\alpha \leftarrow 0.5\alpha$  \hspace{1cm} // decrease stepsize
8: end if
9: until $|y - x| < \theta$  \hspace{1cm} [perhaps for 10 iterations in sequence]

("magic numbers")
Second Order Methods

Compute **Hessian** matrix of second derivatives

\[
H = \begin{bmatrix}
\frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\
\frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\
\vdots & \vdots & \ddots & \vdots \\
\frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2}
\end{bmatrix}
\]
Second Order Methods

Broyden-Fletcher-Goldfarb-Shanno (BFGS) method:

**Input:** initial \( x \in \mathbb{R}^n \), functions \( f(x), \nabla f(x) \), tolerance \( \theta \)

**Output:** \( x \)

1. initialize \( H^{-1} = I_n \)
2. repeat
3. compute \( \Delta = -H^{-1} \nabla f(x) \)
4. perform a line search \( \min_\alpha f(x + \alpha \Delta) \)
5. \( \Delta \leftarrow \alpha \Delta \)
6. \( y \leftarrow \nabla f(x + \Delta) - \nabla f(x) \)
7. \( x \leftarrow x + \Delta \)
8. update \( H^{-1} \leftarrow \left( I - \frac{y \Delta^T}{\Delta^T y} \right)^T H^{-1} \left( I - \frac{y \Delta^T}{\Delta^T y} \right) + \frac{\Delta \Delta^T}{\Delta^T y} \)
9. until \( \|\Delta\|_\infty < \theta \)

Memory-limited version: L-BFGS
Stochastic Gradient Descent

What if $N$ is really large?

Batch gradient descent (evaluates all data)

$$w_t = w_{t-1} - \alpha_t \nabla_w E(y; w)|_{w=w_{t-1}}$$

Minibatch gradient descent (evaluates subset)

$$w_t = w_{t-1} - \alpha_t \nabla_w E(y_t; w)|_{w=w_{t-1}} \quad y_t \subset y$$

Converges under Robbins-Monro conditions

$$\sum_{t=1}^{\infty} \alpha_t = \infty \quad \sum_{t=1}^{\infty} \alpha_t^2 < \infty \quad \alpha_t = \frac{\alpha_0}{(\tau+t)^\kappa}$$
Probabilistic Interpretation
Normal Distribution

Right Skewed

Left Skewed

Random

"Bell Curve"
Normal Distribution

Density: \[ f(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}(x-\mu)^2/\sigma^2\right) \]
Central Limit Theorem

If $y_1, \ldots, y_n$ are
1. Independent identically distributed (i.i.d.)
2. Have finite variance $0 < \sigma_y^2 < \infty$

$$f(\bar{y}) = \text{Normal}(\bar{y} ; \mu_y, \sigma_y^2/N) \quad \bar{y} = \frac{1}{N} \sum_{n=1}^{N} y_n$$
Multivariate Normal

Density: \[ f(x; \mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^D |\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right) \]
Regression: Probabilistic Interpretation

\[ y_n = ax_n + b + \sigma \epsilon_n \quad \epsilon \sim \text{Normal}(0, 1) \]
Regression: Probabilistic Interpretation

\[ \mu_n = w^T x_n \quad y_n \sim \text{Normal}(\mu_n, \Sigma) \]
Regression: Probabilistic Interpretation

Joint probability of $N$ independent data points

\[ p(y_1, \ldots, y_N) = \prod_{n=1}^{N} p(y_n) \]

\[ = \frac{1}{\sqrt{2\pi\sigma^2}} \prod_{n=1}^{N} \exp\left(-\frac{1}{2} \frac{(x-\mu)^2}{\sigma^2}\right) \]

\[ = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2} \sum_{n=1}^{N} \frac{(x-\mu)^2}{\sigma^2}\right) \]
Regression: Probabilistic Interpretation

**Log** joint probability of $N$ independent data points

\[
\log p(y_1, \ldots, y_N) = \sum_{n=1}^{N} \log p(y_n)
\]

\[
= -\frac{1}{2} \left[ N \log(2\pi\sigma^2) + \sum_{n=1}^{N} \frac{(y_n - \mu_n)^2}{\sigma^2} \right]
\]
Regression: Probabilistic Interpretation

**Log** joint probability of $N$ independent data points

\[ \log p(y_1, \ldots, y_N) = \sum_{n=1}^{N} \log p(y_n) \]

\[ = -\frac{1}{2} \left[ N \log(2\pi\sigma^2) + \sum_{n=1}^{N} \frac{(y_n - \mathbf{w}^\top \mathbf{x}_n)^2}{\sigma^2} \right] \]
Regression: Probabilistic Interpretation

\textbf{Log} joint probability of \( N \) independent data points

\[ \log p(y_1, \ldots, y_N) = \sum_{n=1}^{N} \log p(y_n) \]

\[ = -\frac{1}{2} \left[ N \log(2\pi\sigma^2) + \sum_{n=1}^{N} \frac{(y_n - w^\top x_n)^2}{\sigma^2} \right] \]

\[ = -\frac{N}{2} \left[ \text{const} + E(w) \right] \]
Regression: Probabilistic Interpretation

**Log** joint probability of $N$ independent data points

$$
\log p(y_1, \ldots, y_N) = \sum_{n=1}^{N} \log p(y_n)
$$

$$
= -\frac{1}{2} \left[ N \log(2\pi\sigma^2) + \sum_{n=1}^{N} \frac{(y_n - \mathbf{w}^\top \mathbf{x}_n)^2}{\sigma^2} \right]
$$

$$
= -\frac{N}{2} \left[ \text{const} + E(\mathbf{w}) \right]
$$

$$
\arg\max_{\mathbf{w}} p(y_1, \ldots, y_N; \mathbf{w}) = \arg\min_{\mathbf{w}} E(\mathbf{w})
$$

*Maximum Likelihood*
Basis function regression

Linear regression

\[ y = w_0 + w_1 x_1 + \ldots + w_D x_D = \mathbf{w}^T \mathbf{x} \]

Basis function regression

\[ y = w_0 + w_1 \phi_1(\mathbf{x}) + \ldots + w_D \phi_D(\mathbf{x}) \]

Polynomial regression

\[ \mathbf{x}_d := \phi_d(\mathbf{x}) \quad \phi_d(\mathbf{x}) := \mathbf{x}_d^d \]
Polynomial Regression

1.1. Example: Polynomial Curve Fitting

\[ t(x) = \sum_{m=0}^{M} a_m x^m \]

The RMS error defined by

\[ E_{\text{RMS}} = \sqrt{\frac{\sum_{i=1}^{N} (w_i t - \hat{t}_i)^2}{N}} \]

allows us to compare different sizes of data sets on an equal footing, and the square root ensures that \( E_{\text{RMS}} \) is measured on the same scale (and in the same units) as the target variable \( t \). Graphs of the training and test set RMS errors are shown, for various values of \( M \), in Figure 1.5. The test set error is a measure of how well we are doing in predicting the values of \( t \) for new data observations of \( x \). We note from Figure 1.5 that small values of \( M \) give relatively large values of the test set error, and this can be attributed to the fact that the corresponding polynomials are rather inflexible and are incapable of capturing the oscillations in the function \( \sin(2\pi x) \). Values of \( M \) in the range \( 3 \leq M \leq 8 \) give small values for the test set error, and these also give reasonable representations of the generating function \( \sin(2\pi x) \), as can be seen, for the case of \( M = 3 \), from Figure 1.4.
Polynomial Regression

1.1. Example: Polynomial Curve Fitting

Figure 1.4 Plots of polynomials having various orders $M$, shown as red curves, fitted to the data set shown in Figure 1.2. The $(\text{RMS})$ error defined by

$$E_{\text{RMS}} = \sqrt{\frac{2}{N} E(w^{\star})}$$

in which the division by $N$ allows us to compare different sizes of data sets on an equal footing, and the square root ensures that $E_{\text{RMS}}$ is measured on the same scale (and in the same units) as the target variable $t$. Graphs of the training and test set RMS errors are shown, for various values of $M$, in Figure 1.5. The test set error is a measure of how well we are doing in predicting the values of $t$ for new data observations of $x$. We note from Figure 1.5 that small values of $M$ give relatively large values of the test set error, and this can be attributed to the fact that the corresponding polynomials are rather inflexible and are incapable of capturing the oscillations in the function $\sin(2\pi x)$. Values of $M$ in the range $3 \leq M \leq 8$ give small values for the test set error, and these also give reasonable representations of the generating function $\sin(2\pi x)$, as can be seen, for the case of $M = 3$, from Figure 1.4.

Underfit
Polynomial Regression

1. Example: Polynomial Curve Fitting

![Plots of polynomials having various orders](image)

Figure 1.4: Plots of polynomials having various orders $M$, shown as red curves, fitted to the data set shown in Figure 1.2.

(RMS) error defined by

$$\text{ERMS} = \sqrt{\frac{\sum (w \star t)^2}{N}}$$

in which the division by $N$ allows us to compare different sizes of data sets on an equal footing, and the square root ensures that $\text{ERMS}$ is measured on the same scale (and in the same units) as the target variable $t$. Graphs of the training and test set RMS errors are shown, for various values of $M$, in Figure 1.5. The test set error is a measure of how well we are doing in predicting the values of $t$ for new data observations of $x$.

We note from Figure 1.5 that small values of $M$ give relatively large values of the test set error, and this can be attributed to the fact that the corresponding polynomials are rather inflexible and are incapable of capturing the oscillations in the function $\sin(2\pi x)$.

Values of $M$ in the range $3 \leq M \leq 8$ give small values for the test set error, and these also give reasonable representations of the generating function $\sin(2\pi x)$, as can be seen, for the case of $M = 3$, from Figure 1.4.
Regularization

$L2$ regularization (ridge regression) minimizes:

$$E(w) = \| Xw - y \|^2 + \lambda \| w \|^2$$

where $\lambda \geq 0$ and $\| w \|^2 = w^T w$

$L1$ regularization (LASSO) minimizes:

$$E(w) = \| Xw - y \|^2 + \lambda |w|_1$$

where $\lambda \geq 0$ and $|w|_1 = \sum_{i=1}^{D} |\omega_i|$
Regularization

$L_2$ Regularization Example

Yijun Zhao

Linear Regression
Regularization

$L2$: closed form solution

\[ w = (X^TX + \lambda I)^{-1}X^Ty \]

$L1$: No closed form solution. Use quadratic programming:

minimize \( \| Xw - y \|^2 \)

s.t. \( \| w \|_1 \leq s \)
Review: Bias-Variance Trade-off

Maximum likelihood estimator

\[ \hat{f} := \arg\max_{\tilde{f}} p(y | \tilde{f}) \]

Bias-variance decomposition

(expected value over possible data points)

\[ \mathbb{E}[(y - \hat{f}(x))^2] = \text{Bias}[\hat{f}(x)]^2 + \text{Var}[\hat{f}(x)] + \sigma^2 \]

\[ \text{Bias}[\hat{f}(x)] = \mathbb{E}[\hat{f}(x) - f(x)] \]

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

\[ \sigma^2 = \mathbb{E}[y^2] - \mathbb{E}[f(x)]^2 \]
Bias-Variance Trade-off

Often:
- low bias $\Rightarrow$ high variance
- low variance $\Rightarrow$ high bias

Trade-off:
K-fold Cross-Validation

1. **Divide** dataset into K “folds”
2. **Train** on all except $k$-th fold
3. **Test** on $k$-th fold
4. **Minimize** test error w.r.t. $\lambda$
K-fold Cross-Validation

- Choices for $K$: 5, 10, $N$ (leave-one-out)
- Cost of computation: $K \times$ number of $\lambda$
A learning curve plots the performance of the algorithm as a function of the size of training data.

**Expected Error**

- **$E_{\text{out}}$**: Generalization error
- **$E_{\text{in}}$**: In-sample error

**Number of Data Points, $N$**
Learning Curve

Simple Model

Complex Model
Loss Functions

squared loss: $\frac{1}{2}(w^\top x - y)^2 \quad y \in \mathbb{R}$

logistic loss: $\log (1 + \exp(-y w^\top x)) \quad y \in \{-1, +1\}$

hinge loss: $\max\{0, 1 - y w^\top x\} \quad y \in \{-1, +1\}$