TODAY’S LECTURE

Data collection → Data processing → Exploratory analysis & Data viz → Analysis, hypothesis testing, & ML → Insight & Policy Decision

Analysis, hypothesis testing, & ML

Insight & Policy Decision

Data collection

Data processing

Exploratory analysis & Data viz
## Feature Scaling

<table>
<thead>
<tr>
<th>base</th>
<th>Area (sq. ft.)</th>
<th># Bathrooms</th>
<th># Bedrooms</th>
<th>Price (in 1000$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1600</td>
<td>2.5</td>
<td>3</td>
<td>220</td>
</tr>
<tr>
<td>1</td>
<td>1400</td>
<td>1.5</td>
<td>3</td>
<td>180</td>
</tr>
<tr>
<td>1</td>
<td>2100</td>
<td>3.5</td>
<td>4</td>
<td>350</td>
</tr>
<tr>
<td></td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>2400</td>
<td>4</td>
<td>5</td>
<td>500</td>
</tr>
<tr>
<td></td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
NORMALIZATION- ZERO MEAN UNIT STANDARD DEVIATION

\[ x_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j} \]

j: Area, Bathrooms, Bedrooms

<table>
<thead>
<tr>
<th>base</th>
<th>Area(sq. ft.)</th>
<th># Bathrooms</th>
<th># Bedrooms</th>
<th>Price (in 1000$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1600</td>
<td>2.5</td>
<td>3</td>
<td>220</td>
</tr>
<tr>
<td>1</td>
<td>1400</td>
<td>1.5</td>
<td>3</td>
<td>180</td>
</tr>
<tr>
<td>1</td>
<td>2100</td>
<td>3.5</td>
<td>4</td>
<td>350</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>1</td>
<td>2400</td>
<td>4</td>
<td>5</td>
<td>500</td>
</tr>
</tbody>
</table>
### MAX-MIN

\[ x_{ij} = \frac{x_{ij} - x_{\text{min}j}}{x_{\text{max}j} - x_{\text{min}j}} \]

j: Area, Bathrooms, Bedrooms

<table>
<thead>
<tr>
<th>base</th>
<th>Area(sq. ft.)</th>
<th># Bathrooms</th>
<th># Bedrooms</th>
<th>Price (in 1000$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1600</td>
<td>2.5</td>
<td>3</td>
<td>220</td>
</tr>
<tr>
<td>1</td>
<td>1400</td>
<td>1.5</td>
<td>3</td>
<td>180</td>
</tr>
<tr>
<td>1</td>
<td>2100</td>
<td>3.5</td>
<td>4</td>
<td>350</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>....</td>
<td>....</td>
<td>...</td>
<td>...</td>
<td>....</td>
</tr>
<tr>
<td>1</td>
<td>2400</td>
<td>4</td>
<td>5</td>
<td>500</td>
</tr>
</tbody>
</table>
All Observation Model

- **Matrix Notation**
  For all observations

\[
\begin{bmatrix}
1 & x_{11} & x_{12} & \ldots & \ldots & x_{1n} \\
1 & x_{21} & x_{22} & \ldots & \ldots & x_{2n} \\
1 & x_{31} & x_{32} & \ldots & \ldots & x_{3n} \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
1 & x_{m1} & x_{m2} & \ldots & \ldots & x_{mn}
\end{bmatrix}
\begin{bmatrix}
\theta_0 \\
\theta_1 \\
\theta_2 \\
\vdots \\
\theta_n
\end{bmatrix}
= \begin{bmatrix}
y_1 \\
y_2 \\
\vdots \\
y_m
\end{bmatrix}
\]

\[Y = X \theta\]
**GRADIENT DESCENT**

Algorithm for any* hypothesis function $h_\theta : \mathbb{R}^n \rightarrow y$, loss function $\ell : y \times y \rightarrow \mathbb{R}_+$, step size $\alpha$:

Initialize the parameter vector:

- $\theta \leftarrow 0$

Repeat until satisfied (e.g., exact or approximate convergence):

- Compute gradient: $g \leftarrow \sum_{i=1}^{m} \nabla_\theta \ell(h_\theta(x^{(i)}), y^{(i)})$
- Update parameters: $\theta \leftarrow \theta - \alpha \cdot g$

*must be reasonably well behaved
GRADIENT DESCENT - MULTIVARIATE

\[ \theta = 0 \]

Repeat{

\[ \theta_j := \theta_j - \alpha \frac{1}{m} \sum_{i=1}^{m} (h(\theta(x_i) - y_i)x_{ji} \]

(update \( \theta_j \) for all \( j = 1 \ldots n \) simultaneously)

\[ \theta_0 := \theta_0 - \alpha \frac{1}{m} \sum_{i=1}^{m} (h(\theta(x_i) - y_i)x_{0i} \]

\[ \theta_1 := \theta_1 - \alpha \frac{1}{m} \sum_{i=1}^{m} (h(\theta(x_i) - y_i)x_{1i} \]

\[ \theta_2 := \theta_2 - \alpha \frac{1}{m} \sum_{i=1}^{m} (h(\theta(x_i) - y_i)x_{2i} \]

\[ \vdots \]

\[ \theta_n := \theta_n - \alpha \frac{1}{m} \sum_{i=1}^{m} (h(\theta(x_i) - y_i)x_{ni} \]

\[ \frac{1}{m} \sum_{i=1}^{m} (h(\theta(x_i) - y_i)x_{ji} = \frac{\partial}{\partial \theta_j} f(\theta) \]
GRADIENT DESCENT - MULTIVARIATE

\[ \theta = 0 \]

Repeat{

\[ \theta_j := \theta_j - \alpha \frac{1}{m} \sum_{i=1}^{m} (h(\theta(x_i) - y_i)x_{ji} \]

(update \( \theta_j \) for all \( j = 1 \ldots n \) simultaneously)

\[ \frac{1}{m} \sum_{i=1}^{m} (h(\theta(x_i) - y_i)x_{ji} = \frac{\partial}{\partial \theta_j} f(\theta) \]
STOCHASTIC GRADIENT DESCENT
- MULTIVARIATE

\[ \theta = 0 \]

Repeat{

\[ \theta_j := \theta_j - \alpha \frac{1}{m} \sum_{i=1}^{m} (h(\theta(x_i) - y_i)x_{ji} \]

(update \( \theta_j \) for all \( j = 1 \ldots n \) simultaneously

}\}

STOCHASTIC GRADIENT DESCENT

\[ \theta = 0 \]

Repeat{

\( i = \text{random index between 1 and } m \)

\[ \theta_j := \theta_j - \alpha (h(\theta(x_i) - y_i)x_{ji} \]

(update \( \theta_j \) for all \( j = 1 \ldots n \)

}\}
GRADIENT DESCENT

\[ \hat{y} = \theta_0 + \theta_1 x \]

\[ \theta_0 = 0.1 \quad \theta_1 = 0.1 \]

\[
\begin{array}{c|c|c|c|c|c|c}
 x & y & \hat{y} = \theta_0 + \theta_1 x & \frac{1}{2}(\hat{y} - y)^2 & \frac{\partial(SSE)}{\partial \theta_0} & \frac{\partial(SSE)}{\partial \theta_1} \\
\hline
0.2 & 0.44 & 0.12 & 0.0512 & -0.32 & -0.064 \\
0.31 & 0.123 & 0.131 & 0.000032 & 0.008 & 0.00248 \\
0.45 & 0.75 & 0.145 & 0.183 & -0.605 & -0.27225 \\
0.26 & 0.39 & 0.175 & 0.0231 & -0.215 & -0.16125 \\
\hline
\end{array}
\]

\[ \text{SSE} \]

\[ \frac{\partial(SSE)}{\partial \theta_0} \]

\[ \frac{\partial(SSE)}{\partial \theta_1} (\hat{y} - y)x \]

\[ -1.132 \quad -0.495 \]
**GRADIENT DESCENT**

\[ \hat{y} = \theta_0 + \theta_1 x \]

\[ \theta_0 = 0.1 \quad \theta_1 = 0.1 \]

<table>
<thead>
<tr>
<th>x</th>
<th>y</th>
<th>( \hat{y} = \theta_0 + \theta_1 x )</th>
<th>( \frac{1}{2}(\hat{y} - y)^2 )</th>
<th>( \frac{\partial (SSE)}{\partial \theta_0} )</th>
<th>( \frac{\partial (SSE)}{\partial \theta_1} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.31</td>
<td>0.123</td>
<td>0.131</td>
<td>0.0000032</td>
<td>0.008</td>
<td>0.00248</td>
</tr>
</tbody>
</table>

 ..........
STOCHASTIC GRADIENT DESCENT
STOCHASTIC GRADIENT DESCENT
- MINI BATCH

\[ \theta = 0 \]

Repeat {

\[ i_1, \ldots, i_l = \text{random index between 1 and } m \]

\[ \theta_j := \theta_j - \alpha \frac{1}{l} \sum_{i=1}^{l} (h(\theta(x_i) - y_i)x_{ij} \]

(update \( \theta_j \) for all \( j = 1 \ldots n \))

}
CLASSIFICATION PROBLEM
GENERAL CLASSIFICATION PROBLEM

• Can we predict categorical response/output $Y$, from a set of predictors $\theta_1, \theta_2, \ldots, \theta_n$.

• For example, an individual’s choice of transportation:
  • Predictors: income, cost, and time
  • Response: car, bike, bus or train.

• From this classification model, an inference task:
  • how do people value price and time when considering a transportation choice?
WHY NOT LINEAR REGRESSION

• For categorical responses, with more than two values, if order and scale, don’t make sense, it is not a regression problem.

\[ Y = \begin{cases} 
1 & \text{if stroke} \\
2 & \text{if drugoverdose} \\
3 & \text{if epilepticseizure} 
\end{cases} \]

• For binary responses, it is a little better

\[ Y = \begin{cases} 
0 & \text{if stroke} \\
1 & \text{if drugoverdose} 
\end{cases} \]
BINARY RESPONSES

• We could use linear regression and interpret response, Y, as a probability

• for example, if $\hat{y} > 0.5$ (predict drug overdose)
CLASSIFICATION AS PROBABILITY ESTIMATION

• Instead of modeling classes 0 or 1, model conditional class probability $p(Y=1 \mid X = x)$

• classify based on this probability.

• Use of discriminant functions (think of scoring).

• One way to do this is, logistic regression.
LOGISTIC REGRESSION

• Basic idea: Build a linear model related to p(x)

• Linear regression directly (i.e. \( p(x) = \theta_0 + \theta_1 x \)) doesn’t work. Why?

• Instead build a linear model of log-odds:

\[
\log \frac{p(x)}{1 - p(x)} = \theta_0 + \theta_1 x
\]
LOGISTIC REGRESSION - ODDS

• Odds are equivalent to ratios of probabilities.

• For example, “Two to one odds that US wins the Women’s soccer world cup”, means
  
  the probability that US wins the soccer world cup is double the probability that they lose.

• So, if odds = 2, probability, \( p(x) = ?? \)

• if odds = 1/2, \( p(x) = ?? \)
LOGISTIC REGRESSION

• Suppose an individual has a 16% chance of defaulting on their credit card payment. What are the odds that (s)he will default?

• On average, what fraction of people with an odds of 0.37 of defaulting on their credit card payment will in fact default?
LOGISTIC REGRESSION
- PREDICTIONS

\[
\log \frac{p(x)}{1 - p(x)} = \theta_0 + \theta_1 x
\]

\[\implies p(x) = \frac{1}{1 + e^{-\theta^T x}}\]

Sigmoid function, given by

\[
h_\theta(x) = \frac{1}{1 + e^{-\theta^T x}}
\]

\[y = h_\theta(x), z = -\theta^T x\]
ESTIMATION OF PARAMETERS

• Bernoulli probability (think of flipping a coin weighted by \( p(x) \))

• Estimate the parameters to maximize the likelihood of the observed training data under this binomial model.

• Minimize the negative of the log likelihood of the model,

• Optimization problem:

\[
\min_{\theta_0, \theta_1} \sum_{i: y_i = 1} y_i f(x_i) - \log(1 + e^{f(x_i)})
\]

\[
f(x_i) = \theta_0 + \theta_1 x_i
\]
• Hypothesis function / Classifier

\[ 0 \leq h_\theta(x) \leq 1 \]

Sigmoid function / logistic function

\[ h_\theta(x) = \frac{1}{1 + e^{-z}} \]

\[ y = h_\theta(x), \ z = \theta^T x \]
• Hypothesis function / Classifier

\[ 0 \leq h_\theta(x) \leq 1 \]

\[ h_\theta(x) = \frac{1}{1 + e^{-z}} = p(y = 1 | x, \theta) \]

\[ y = 1 \text{ if } h_\theta(x) \geq 0.5 \]

\[ y = 0 \text{ if } h_\theta(x) < 0.5 \]

\[ y = h_\theta(x), z = \theta^T x \]
Training set: \( \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\} \)

\[
x \in \begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_n \end{bmatrix} \quad x_0 = 1, y \in \{0, 1\}
\]

\[
h_\theta(x) = \frac{1}{1 + e^{-\theta^T x}}
\]

Parameters, \( \theta \)?
LOSS FUNCTION

Linear Regression: \( f(\theta, x) = \frac{1}{m} \sum_{i=1}^{m} \frac{1}{2}(h_{\theta}(x_i) - y_i)^2 \)

\[ x \in \begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_n \end{bmatrix} \quad x_0 = 1, y \in \{0, 1\} \]

\[ h_{\theta}(x) = \frac{1}{1 + e^{-\theta^T x}} \]

Parameters, \( \theta? \)

\[ \min_{\theta_0, \theta_1, \ldots, \theta_n} \frac{1}{m} \sum_{i=1}^{m} \left[ -y_i \log(h_{\theta}(x_i)) - (1 - y_i) \log(1 - h_{\theta}(x_i)) \right] \]
LOSS FUNCTION

\[ x \in \left[ \begin{array}{c} x_0 \\ x_1 \\ \vdots \\ x_n \end{array} \right] \quad x_0 = 1, y \in \{0,1\} \]

\[ f(\theta) = -\frac{1}{m} \left[ \sum_{i=1}^{m} y_i \log(h_\theta(x_i)) + (1 - y_i)\log(1 - h_\theta(x_i)) \right] \]

\[ \min_{\theta_0, \theta_1, \ldots, \theta_n} \frac{1}{m} \sum_{i=1}^{m} \left[ -y_i \log(h_\theta(x_i)) - (1 - y_i)\log(1 - h_\theta(x_i)) \right] \]

to predict a new observation, \( x \):

\[ \frac{1}{1 + e^{-\theta^T x}} \]

\[ p(y = 1 \mid x, \theta) \]
GRADIENT DESCENT

\[ f(\theta) = -\frac{1}{m} \left[ \sum_{i=1}^{m} y_i \log(h_\theta(x_i)) + (1 - y_i) \log(1 - h_\theta(x_i)) \right] \]

want \( \min_{\theta_0, \theta_1, \ldots, \theta_n} f(\theta) \)

Repeat \{ 
\[ \theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} f(\theta) \]
(updated all \( \theta_j \) simultaneously) 
\}

\[ \frac{\partial}{\partial \theta_j} f(\theta) = \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x_i) - y_i)x_{ij} \]

where, \( h_\theta(x) = \frac{1}{1 + e^{-\theta^T x}} \), \( \theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_n \end{bmatrix} \)
PREDICTION

![Graph showing probability of petal width (cm) for Virginica and Not Virginica species, with a decision boundary at 1.5 cm.](image-url)
PREDICTION

• For Iris data versicolor classification

  intercept: -4.220
  petal_width feature: 2.617

• Probability that a given petal width is a versicolor

  \[
  \hat{p}(1.7) = \frac{1}{1 + e^{-(\theta_0 + \theta_1 x)}} = \frac{1}{1 + e^{-4.22 + 2.617 \times 1.7}} = 0.55
  \]

  \[
  \hat{p}(2.5) = \frac{1}{1 + e^{-(\theta_0 + \theta_1 x)}} = \frac{1}{1 + e^{-4.22 + 2.617 \times 2.5}} = 0.91
  \]
MULTIPLE LOGISTIC REGRESSION

\[
\log \frac{p(x)}{1 - p(x)} = \theta_0 + \theta_1 x_1 + \ldots + \theta_n x_n
\]
• Suppose we collect data for a group of students in a statistics class with variables $x_1 =$ hours, $x_2 =$ undergrad GPA, and $y =$ receive an A. We fit a logistic regression and find the following coefficients,

$$
\theta_0 = -6, \theta_1 = 0.05, \theta_2 = 1
$$

Estimate the probability that a student who studies 40 hours and has an undergraduate GPA of 3.5, gets an A in the class.

• With estimated parameters from previous question, and GPA of 3.5 as before, how many hours would the student need to study to have a 50% chance of getting an A in the class?