Smart Transportation

Tarek Abdelzaher
Dept. of Computer Science
University of Illinois at Urbana Champaign

Green GPS

Subscribers

- OBDII-WiFi Adaptor ($50)
- GPS Phone

Subscribers: Premium service, High savings

Open access: Standard service, Average savings

Fuel Data

Physical Models

- $F_{\text{engine}} = \frac{V_{\text{engine}}}{V_{\text{engine}}}$
- $F_{\text{friction}} = c_{\text{friction}} \cdot \text{speed}^2$
- $F_{\text{air}} = \rho \cdot \text{speed}^2$
- $F_{\text{rudder}} = c_{\text{rudder}} \cdot \text{rudder angle}$

Fuel Error (%), Trip Length

Shortest and fastest

Green GPS
Green GPS

- An estimated 200 million cars on road in the US
- Modern cars (post 1996) equipped with OBD-II sensor interface
  - Measures various engine parameters
  - OBD-II scanners are commercially available

Cell phone + Wireless OBD = Large scale car sensor network

EPA Statistics

- 200 million light vehicles on the streets
- Each driven 12000 miles annually on average
- Average MPG is 20.3 miles/gallon
- **118 Billion Gallons of Fuel per year!**
- **Savings of 1% = One Billion Gallons**
GreenGPS: Fuel Efficient Routing

- Fuel efficient route is different from shortest or fastest route
  - Congestion → shortest may not be fuel efficient
  - MPG vs. speed is non-linear → fastest may not be fuel efficient

Initial Participatory Sensing Deployment

- OBD scanner with GPS to collect location tagged car sensor data
- 16 different compact and mid-sized sedans, e.g. Ford, Toyota, Honda
- Over 1000 miles of data collected
- Users record sensor data and GPS on SD card and upload to server
Fuel Consumption Model

- Simple model for fuel consumption derived from physics principles
- Approximate based on easily measurable parameters (e.g. stop signs, traffic lights, speed limits)

\[ gpm = k_1 \frac{\nu^2}{\Delta d} + k_2 \frac{\hat{a}^2}{\Delta d} + k_3 \cos(\theta) + k_4 A \nu^2 + k_5 \sin(\theta) \]

Sampling Regression Modeling Framework

Fuel consumption of 16 cars driven on a few roads

Predict fuel consumption of any car on any road in Urbana-Champaign
Finding Fuel-efficient Routes

- Example

Generalization Hierarchy

- Fuel consumption of different cars is different, but is their model the same?
- Derive a hierarchy for prediction using the *sampling regression* framework
Regression Modeling

- Given a set of measurements $Y = y_i$ and set of attributes $U = \{x_{ij}\}$
- Measurement $y_i$ associated with $k$ attributes $x_{i1}, \ldots, x_{ik}$
- Linear Regression Model:
  \[
  \hat{Y} = U \eta
  \]
  \[
  \epsilon = Y - \hat{Y}
  \]
- One size fits all model?

Generalization and Modeling

- Regression modeling:
  - Problem: one size does not fit all. Who says that Fords and Toyotas have the same regression model?
- Regression model per car?
  - Problem: Cannot use data collected by some cars to predict fuel consumption of others.
- Challenge: Must jointly determine both (i) regression models and (ii) their scope of applicability, to cover the whole data space with acceptable modeling error.
Data Cubes

- What are they?

Data Cell Measures

- **Measure**: A statistical function of data in each cell
  - In our example: average oil price
  - Main challenge: computing measures recursively and without reprocessing raw data
  - Measures can be classified as:
    - **Distributive** - \( f(x_1, x_2, x_3) = f(f(x_1, x_2), x_3) \) - Efficient
      - Examples: sum, count
    - **Algebraic/Compressible** - An algebraic combination of distributive functions - Efficient
      - Example: average = sum/count
    - **Holistic** - Reprocess raw data - Inefficient
      - Example: median
Regression Cubes

- Data cells correspond to:
  - Output attributes \( Y_c = \{ y_i \} \)
  - Each associated with \( k \) input attributes \( x_{i1}, \ldots, x_{ik} \).
    \[ X_c = \{ x_{ij} \} \]

- Data cells store the following measures:
  - Regression coefficients:
    \[ \hat{Y}_c = X_c \hat{\eta}_c \]
  - Regression modeling error:
    \[ E_{rr,c} = (Y_c - X_c \hat{\eta}_c)^T (Y_c - X_c \hat{\eta}_c) \]

Example of Regression Cubes

- Goal: predict fuel consumption
  - Group by make, model, or year
Example of Regression Cubes

Data Cells:

\((*,*,*) = X, Y\)

Example of Regression Cubes

Data Cells:

\((\text{Toyota},*,*) = X_{c1} Y_{c1}\)
\((\text{Ford},*,*) = X_{c2} Y_{c2}\)
\((\text{Honda},*,*) = X_{c3} Y_{c3}\)
The Challenge of Regression Cubes

- Main challenge: compute cuboid measures, the model and error, recursively (without reprocessing raw data)

- Model parameters and estimation error at cell $c$
  - Not distributive

\[
\hat{Y}_c = X_c \hat{\eta}_c
\]

\[
Err_c = (Y_c - X_c \hat{\eta}_c)^T (Y_c - X_c \hat{\eta}_c)
\]

Compressed Representation

- Compressed representation of a cell $c$:
  - $\rho_c = Y_c^T Y_c$: scalar value
  - $\Theta_c = X_c^T X_c$: vector of size $k$
  - $\nu_c = X_c^T Y_c$: $k$ by $k$ matrix
  - $n_c$: number of samples

\[
\rho_c = \sum_{i=1}^{m} \rho_i \quad \nu_c = \sum_{i=1}^{m} \nu_i \quad \Theta_c = \sum_{i=1}^{m} \Theta_i \quad n_c = \sum_{i=1}^{m} n_{c_i}
\]

- These matrices are distributive measures
Compressible Measures

- Model coefficients:

\[ \hat{\eta}_c = (X_c^T X_c)^{-1} X_c^T Y_c = \Theta_c^{-1} \nu_c \]

- Error:

\[ \text{Err}_c = (Y_c - X_c \hat{\eta}_c)^T (Y_c - X_c \hat{\eta}_c) = \\
Y_c^T Y_c - (X_c \hat{\eta}_c)^T Y_c - Y_c^T X_c \hat{\eta}_c + (X_c \hat{\eta}_c)^T X_c \hat{\eta}_c = \\
\rho_c - \hat{\eta}_c^T \nu_c - \nu_c^T \hat{\eta}_c + \hat{\eta}_c^T \Theta_c \hat{\eta}_c \]

- Model coefficients and regression error are compressible measures

Sparse Sampling Challenge

- Challenge: High dimensions \( \rightarrow \) sparse samples

- Need:
  - A method to identify when samples are too sparse for a specific model
  
  - A technique to reduce the number of model parameters: Model reduction
  
  - A technique to group samples together and increase the number of samples: Query expansion
Data Cell Confidence

- Measure of confidence:
  - Probability at which the actual coefficients are far from the estimate

\[
P_r[||\hat{\eta}_c - \eta_c|| > \delta] = \frac{k\sigma^2}{\delta^2\lambda_{\text{min}}(X_c^TX_c)}
\]

- Reliable Cell:

\[
\frac{k\hat{\sigma}^2}{\delta^2\lambda_{\text{min}}(\Theta_c)} < 0.05
\]

Model Reduction

- Independently find \( L \) for each cell, such that:
  - The cell is reliable
  - Corresponding error is minimized
  - Exponential number of possible \( L \)s

Our heuristic:

<table>
<thead>
<tr>
<th>( L )</th>
<th>Error</th>
<th>Reliable</th>
</tr>
</thead>
<tbody>
<tr>
<td>{v}</td>
<td>0.031</td>
<td>yes</td>
</tr>
<tr>
<td>{m}</td>
<td>0.152</td>
<td>yes</td>
</tr>
<tr>
<td>{A}</td>
<td>0.043</td>
<td>yes</td>
</tr>
<tr>
<td>{S}</td>
<td>0.056</td>
<td>yes</td>
</tr>
</tbody>
</table>

Attributes:
- Velocity (\( v \))
- Mass (\( m \))
- Frontal area (\( A \))
- Stop signs (\( S \))
Model Reduction

- Independently find $L$ for each cell, such that:
  - The cell is reliable
  - Corresponding error is minimized
  - Exponential number of possible $L$s
- Our heuristic:

<table>
<thead>
<tr>
<th>Attributes</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Velocity ($v$)</td>
<td></td>
</tr>
<tr>
<td>Mass ($m$)</td>
<td></td>
</tr>
<tr>
<td>Frontal area ($A$)</td>
<td></td>
</tr>
<tr>
<td>Stop signs ($S$)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$L = {v}$</th>
<th>0.031</th>
<th>yes</th>
</tr>
</thead>
<tbody>
<tr>
<td>$L = {m}$</td>
<td>0.152</td>
<td>yes</td>
</tr>
<tr>
<td>$L = {A}$</td>
<td>0.043</td>
<td>yes</td>
</tr>
<tr>
<td>$L = {S}$</td>
<td>0.056</td>
<td>yes</td>
</tr>
</tbody>
</table>

Model Reduction

- Independently find $L$ for each cell, such that:
  - The cell is reliable
  - Corresponding error is minimized
  - Exponential number of possible $L$s
- Our heuristic:

<table>
<thead>
<tr>
<th>Attributes</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Velocity ($v$)</td>
<td></td>
</tr>
<tr>
<td>Mass ($m$)</td>
<td></td>
</tr>
<tr>
<td>Frontal area ($A$)</td>
<td></td>
</tr>
<tr>
<td>Stop signs ($S$)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$L = {v}$</th>
<th>0.021</th>
<th>no</th>
</tr>
</thead>
<tbody>
<tr>
<td>$L = {v, m}$</td>
<td>0.030</td>
<td>yes</td>
</tr>
<tr>
<td>$L = {v, A}$</td>
<td>0.028</td>
<td>yes</td>
</tr>
</tbody>
</table>
Model Reduction

- Independently find $L$ for each cell, such that:
  - The cell is reliable
  - Corresponding error is minimized
  - Exponential number of possible $L$s
- Our heuristic:

<table>
<thead>
<tr>
<th>L</th>
<th>Error</th>
<th>Reliable</th>
</tr>
</thead>
<tbody>
<tr>
<td>$L = {v}$</td>
<td>0.021</td>
<td>no</td>
</tr>
<tr>
<td>$L = {v, m}$</td>
<td>0.030</td>
<td>yes</td>
</tr>
<tr>
<td>$L = {v, A}$</td>
<td>0.028</td>
<td>yes</td>
</tr>
<tr>
<td>$L = {v, S}$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$L = {m}$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$L = {A}$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$L = {S}$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Attributes**
- Velocity ($v$)
- Mass ($m$)
- Frontal area ($A$)
- Stop signs ($S$)
Model Reduction

- Independently find $L$ for each cell, such that:
  - The cell is reliable
  - Corresponding error is minimized
  - Exponential number of possible $L$s

- Our heuristic:

  $L = \{v\}$
  $L = \{m\}$
  $L = \{A\}$
  $L = \{S\}$
  $L = \{v, m\}$
  $L = \{v, A\}$
  $L = \{v, S\}$
  $L = \{v, S, m\}$
  $L = \{v, S, A\}$
  Reduced Model: $\{v, S\}$

Model Performance

All driven paths are split into smaller segments to capture variations in fuel consumption on individual streets
Model Performance

- The sampling regression cube improves prediction accuracy significantly
- A regression cube without model reduction is even worse than a single model!

Accuracy Results
## Generalization Hierarchy Evaluation Results

- Evaluate model performance using this framework

## Fuel Savings Evaluation

- Experiments on five cars, each does *four round-trips* between 2 landmarks in Urbana-Champaign on *fastest* and *shortest* routes

### Car Details

<table>
<thead>
<tr>
<th>Car Make</th>
<th>Year</th>
<th>Model</th>
<th>Landmarks</th>
<th>Route</th>
<th>Savings %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Honda Accord</td>
<td>2001</td>
<td>2001</td>
<td>H1 to Mall</td>
<td>Shortest</td>
<td>31.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>H1 to Gym</td>
<td>Shortest</td>
<td>19.7</td>
</tr>
<tr>
<td>Ford Taurus 2001</td>
<td></td>
<td></td>
<td>H2 to Restaurant</td>
<td>Shortest</td>
<td>26</td>
</tr>
<tr>
<td>Toyota Celica 2001</td>
<td></td>
<td></td>
<td>H2 to Work</td>
<td>Fastest</td>
<td>10.1</td>
</tr>
<tr>
<td>Nissan Sentra 2009</td>
<td></td>
<td></td>
<td>H3 to CUPHD</td>
<td>Fastest</td>
<td>8.4</td>
</tr>
<tr>
<td>Honda Civic 2002</td>
<td></td>
<td></td>
<td>Grad to Work</td>
<td>Fastest</td>
<td>18.7</td>
</tr>
</tbody>
</table>

Average fuel savings across 5 cars
Finding Fuel-efficient Routes

- GreenGPS routes (green) may be different from fastest (blue), shortest (red), and Garmin EcoRoute (purple)

Fuel Savings Evaluation

- GreenGPS savings compared to fastest route, shortest route, and Garmin EcoRoute