Visualization and Feature Importance

Exploratory Data Analysis
Images, word clouds
Univariate vs. multivariate
Feature importance

Figure 1: Example Twitter profile pictures for users scoring high in a personality trait.
Visualization matters

- Check data quality
- Understand the data
- Understand the model
  - To aid in model development
  - To explain results to users
Exploratory Data Analysis (EDA)

- Look at the data!!!
- Look at some images; read some posts
- Counts
  - Present/missing
- Means/standard deviations
- Histograms
- Correlations of features with outputs
<table>
<thead>
<tr>
<th>1900s</th>
<th>1910s</th>
<th>1920s</th>
<th>1930s</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1940s</td>
<td>1950s</td>
<td>1960s</td>
<td>1970s</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1980s</td>
<td>1990s</td>
<td>2000s</td>
<td>2010s</td>
</tr>
</tbody>
</table>

Shiry Ginosaur et al.
Variable explanation/importance

- **Interpretation**
  - Find items closest to the cluster center
  - Find words closest to a vector embedding

- **Method specific or agnostic variable importance**

- **argmax_x f(x) for hidden nodes or outputs**
  - Which input (Image, document …) maximizes the p(Y=y)?
\[
\text{argmax}_x f(x)
\]

(a) Extraverted.  
(b) Conscientious.

Figure 1: Example Twitter profile pictures for users scoring high in a personality trait.
Variable Importance

- \( y = 1000 \, x_1 + x_2 \)
- Which is more important: \( x_1 \) or \( x_2 \)?
- How should you measure importance?
- Possible answers:
  - Standardize \( x_1 \) and \( x_2 \)
  - Change each of the features over its usual range and see how much \( y \) changes
  - Remove each of the features and see how the prediction changes - *with or without retraining the model*
Variable Importance: Regression

- Univariate and multivariate are different
  - Since features are usually highly redundant

- True model: \( y = x_1 + x_5 \)

- Fit: \( y = c_1 x_1 + c_2 x_2 + c_3 x_3 + c_4 x_4 + c_5 x_5 \)
  - with \( x_1 = x_2 = x_3 = x_4 \)

- Giving: \( y = \frac{1}{4} x_1 + \frac{1}{4} x_2 + \frac{1}{4} x_3 + \frac{1}{4} x_4 + c_5 x_5 \)

- How important is \( x_1 \)?
  - \( \frac{1}{4} \) or 1?
Kinds of generic variable importance

◆ The accuracy loss from leaving out a variable when building a model
  - What is the importance of $x_1$ in
    \[ y = c_1 x_1 + c_2 x_2 + c_3 x_3 + c_4 x_4 + c_5 x_5 \]
    with $x_1 = x_2 = x_3 = x_4$

◆ The accuracy loss from pegging a variable to its average value in a trained model
Random Forest Variable Importance

- Find test set error, $\text{Err}$
- Permute a variable $x_j$, find new test set error, $\text{Err}_t$
- Variable importance is the difference, $(\text{Err} - \text{Err}_t)$
  divided by the standard error

From the R package for Random forests
For interpretation

- Find correlation of each feature $x_j$ with $y$
  - But beware on nonlinear relations
  - Pearson vs. Spearman correlations

Pearson = -0.799, Spearman = -1

https://support.minitab.com/
Look at the data!

- Frequency
- Correlation (Pearson)
  
  \[ corr(x,y) = \frac{E[(x- \mu_x)(y-\mu_y)]}{\sigma_x \sigma_y} \]
Words reflect who says them
Words reflect who says them
Well adjusted (anti-neurotic) words

- church
- holla
- san_diego
- heads
- headed_to
- fullest
- miami
- psalm
- practice
- lets_go
- opportunity
- spring_break
- the_pool
- the_gym
- grind
- volleyball
- proverbs
- smh
- on_my_way
- greatness
- cali
- skiing
- morning_fb
- swag
- dat
- mountains
- beautiful
- beautiful!
- day
- soccer
- blessed
- basketball
- blessings
- death
- workout
- fam
- heaven
- bless
- celtics
- champs
- success
- blessed
- faith
- our_team
- fam
- fam
- death
- holla
- beautiful
- beautiful!
What you should know

◆ Start by looking at distributions
  ● Look for outliers
  ● Label clusters with frequent items close to the center

◆ $\text{argmax}_x f(x)$ for feature detectors or outputs
  ● Images, words/documents ...

◆ Correlations (Pearson or Spearman)
  ● E.g., word clouds

◆ Univariate vs. multivariate variable importance