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

Visualization and Feature Importance

Lyle Ungar

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
Variable explanation/importance

◆ Interpretation
  ● Find items closest to the cluster center
  ● Find words closest to a vector embedding

◆ Method specific or agnostic variable importance

◆ $\text{argmax}_x f(x)$ 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: Regression

- \( y = 1000 \, x_1 + x_2 \)
- **Which is more important:** \( x_1 \) or \( x_2 \)?
- **How should you measure importance?**
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?
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

- Just 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**

\[ \text{corr}(x,y) = \frac{E[(x - \mu_x)(y - \mu_y)]}{\sigma_x \sigma_y} \]
Words reflect who says them

wwwbp.org
Words reflect who says them
Well adjusted (anti-neurotic) words
Visualization Summary

- $\text{argmax}_x f(x)$ for feature detectors or outputs
  - Images
  - words/documents …

- Regression (or other) coefficients
  - E.g. word clouds
  - Careful about the normalization

- Items close to the cluster center

- Variable importance
  - Beware univariate vs. multivariate