Building statistical models by visualization

Tom Minka
CMU Statistics Dept
Outline

• Scatterplots
  – independence, causality

• QQ plots
  – distribution checking

• Residual plots
  – linearity, outliers

• Projections for regression
  – additivity

• Projections for classification
  – linearity
Football statistics

Is there a better representation?

<table>
<thead>
<tr>
<th></th>
<th>W</th>
<th>L</th>
<th>T</th>
<th>PF</th>
<th>PA</th>
</tr>
</thead>
<tbody>
<tr>
<td>CIN</td>
<td>2</td>
<td>14</td>
<td>0</td>
<td>279</td>
<td>456</td>
</tr>
<tr>
<td>OAK</td>
<td>11</td>
<td>5</td>
<td>0</td>
<td>450</td>
<td>304</td>
</tr>
<tr>
<td>GB</td>
<td>12</td>
<td>4</td>
<td>0</td>
<td>398</td>
<td>328</td>
</tr>
<tr>
<td>SEA</td>
<td>7</td>
<td>9</td>
<td>0</td>
<td>355</td>
<td>369</td>
</tr>
<tr>
<td>JAC</td>
<td>6</td>
<td>10</td>
<td>0</td>
<td>328</td>
<td>315</td>
</tr>
<tr>
<td>NO</td>
<td>9</td>
<td>7</td>
<td>0</td>
<td>432</td>
<td>388</td>
</tr>
<tr>
<td>KC</td>
<td>8</td>
<td>8</td>
<td>0</td>
<td>467</td>
<td>399</td>
</tr>
<tr>
<td>TB</td>
<td>12</td>
<td>4</td>
<td>0</td>
<td>346</td>
<td>196</td>
</tr>
<tr>
<td>MIN</td>
<td>6</td>
<td>10</td>
<td>0</td>
<td>390</td>
<td>442</td>
</tr>
<tr>
<td>TEN</td>
<td>11</td>
<td>5</td>
<td>0</td>
<td>367</td>
<td>324</td>
</tr>
<tr>
<td>CAR</td>
<td>7</td>
<td>9</td>
<td>0</td>
<td>258</td>
<td>302</td>
</tr>
<tr>
<td>NYJ</td>
<td>9</td>
<td>7</td>
<td>0</td>
<td>359</td>
<td>336</td>
</tr>
<tr>
<td>NE</td>
<td>9</td>
<td>7</td>
<td>0</td>
<td>381</td>
<td>346</td>
</tr>
<tr>
<td>STL</td>
<td>7</td>
<td>9</td>
<td>0</td>
<td>316</td>
<td>369</td>
</tr>
</tbody>
</table>
Visual independence test

• “Permutation test”
• Randomly pair x values with y values
• If the distribution looks different from the original, the variables are dependent
• No distributional assumptions required
Comparing distributions

- “Quantile-quantile plot”
- A pseudo-scatterplot for unpaired data
- Quantile of $x$ = fraction of points $< x$
- Plot quantile $q$ in one set against quantile $q$ in the other set, for all $q$
- Tells you how to transform one variable to have the distribution of the other
Regression models

How to do regression visually:

• Transform to make the picture simpler
• Fit a simple model
• Use residuals to suggest more complex models, outliers to remove
• Iterate
High-dimensional data

- Two basic approaches to visualization
  - Many points, few dimensions:
    - Projection
    - Slicing
  - Few points, many dimensions:
    - Parallel-coordinates
    - Iconic displays
Projections

- **PCA**
  - Maximize the spread of the projected data
- **Regression projection**
  - Project only the predictors (inputs)
  - Maximize the spread of the response (output)
Boston Housing data

- Predict the median house value in Boston census tracts, based on crime, poverty, industry, pollution, etc.
- A regression problem with many predictors
- Is an additive model appropriate?
Discriminative projections

• M-projection (Fisher, LDA)
  – Tries to separate means of classes
• V-projection
  – Tries to separate variances of classes
• MV-projection
  – Maximize KL divergence between Gaussians
  – Separates means and variances
Sonar problem

- Sonar echo is represented by energy in 60 frequency bands
- Mines vs. Rocks
- Dataset is linearly separable, but 1nn consistently beats linear classifiers
Vowels dataset

- Binary problem: “hid” vs. rest
- Knn and quadratic kernel beat linear
Online digit recognition

- Classify “8” vs. rest
- Knn beats quadratic kernel beats linear
Some good books

• “The Elements of Graphing Data”, William Cleveland, 2nd Ed.
• “Visualizing Data”, William Cleveland
• “The Visual Display of Quantitative Information”, Edward Tufte
• “Exploratory Data Analysis”, John Tukey
Summary

- Visualization is a simple and fast way to check model assumptions and learn about a domain.
- Many opportunities still exist to design better graphs, esp. for high dimensions.
- Visualization is not “art”, but a well-structured field, worthy of research attention.