Lesson 3: Modelling the Web with Advanced Statistical Descriptive Text Models

Unit 3: Fitting a curve on a (log-log) plot

Rene Pickhardt

Introduction to Web Science Part 2
Emerging Web Properties
Completing this unit you should

• Know the axioms for a distance measure and how they relate to norms.

• Know at least two distance measures on functions spaces.

• Understand why changing to the CDF makes sense when looking at distance between functions.

• Understand the principle of the Kolomogorov-Smirnov test for fitting curves
Can we fit a function to this data?

• On a log log plot the rank / frequency diagram appears roughly as a straight line

1. Power functions appear as straight lines on log log plots

2. Distance of both functions should be smaller than c

$$\| f_{\text{fit}} - f_{\text{observed}} \| < c$$
Which curve is fitting the data best? Why?

Wordrank frequency diagram on Wikipedia data sets (log-log scale)
Which of the black lines is longest – Why?

Wordrank frequency diagram on Wikipedia data sets (top ranks)
Again! Do not get fooled by log

Wordrank frequency diagram on Wikipedia data sets (top ranks)

\[ d = 115627.15 \]

\[ d = 1065.54 \]

\[ d = 28.1 \]
Distances are best seen on linear plots

- It makes sense to look at the top ranked words to find the greatest distance
Let’s be a little more systematic

• We looked at maximum point wise distance

• We used this as a distance measure between functions

• Are there others / better distance measures for functions?

• How can distance measures be characterized anyway?
How to define distance between functions?

• Recall our goal:
  – Find \( f_{fit} \) such that \( ||f_{fit} - f_{observed}|| < c \)

• We define the distance of two functions as:
  \[
  d(f_{fit}, f_{observed}) := ||f_{fit} - f_{observed}||
  \]

• But how to calculate \( ||g|| \) for some function?
The uniform norm (aka sup norm)

• Let $f : M \rightarrow \mathbb{R}$ be a function

$$\| f \|_\infty := \sup_{x \in M} \| f(x) \|_{\mathbb{R}} = \sup \{ |f(x)| : x \in M \}$$

• $\| f \|_\infty$ is a norm i.e. it has the following properties
  – Positive definite
  – Homogeneous
  – Triangle inequality
Positive definite ($\|f\|_\infty = 0 \Rightarrow f = 0$)

- Let $f : M \rightarrow \mathbb{R}$ be a function

$$\|f\|_\infty := \sup_{x \in M} \|f(x)\|_{\mathbb{R}}$$

**Proof:**

$$\|f\|_\infty = 0 \iff \sup_{x \in M} \|f(x)\|_{\mathbb{R}} = 0$$

$$\Rightarrow \|f(x)\|_{\mathbb{R}} = 0 \ \forall x$$

$$\Rightarrow f(x) = 0 \ \forall \ \Rightarrow \ f = 0$$
Homogeneous ($\|\alpha f\|_\infty = \alpha \|f\|_\infty$, $\alpha \in \mathbb{R}$)

- Let $f : M \to \mathbb{R}$ be a function

$$\|f\|_\infty := \sup_{x \in M} \|f(x)\|_\mathbb{R} = \sup \{ |f(x)| : x \in M \}$$

- Proof:

$$\|\alpha f\|_\infty = \sup_{x \in M} \|\alpha f(x)\|_\mathbb{R}$$

$$= \sup_{x \in M} |\alpha| \|f(x)\|_\mathbb{R}$$

$$= |\alpha| \sup_{x \in M} \|f(x)\|_\mathbb{R} = |\alpha| \|f\|_\infty$$
Triangle inequality \[ \| f + g \|_\infty \leq \| f \|_\infty + \| g \|_\infty \]

\[
\| f + g \|_\infty = \sup_{x \in M} \| f(x) + g(x) \|_\mathbb{R}
\]

\[
\leq \sup_{x \in M} \| f(x) \|_\mathbb{R} + \| g(x) \|_\mathbb{R}
\]

\[
\leq \sup_{x \in M} \| f(x) \|_\mathbb{R} + \sup_{x \in M} \| g(x) \|_\mathbb{R}
\]

\[
= \| f \|_\infty + \| g \|_\infty
\]
“-0.9” now seems to be the best exponent

<table>
<thead>
<tr>
<th>$f_{fit}$</th>
<th>$d(f_{obs}, f_{fit})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$C/x^{0.9}$</td>
<td>115 k</td>
</tr>
<tr>
<td>$C/x^{1.0}$</td>
<td>148 k</td>
</tr>
<tr>
<td>$C/x^{1.1}$</td>
<td>177 k</td>
</tr>
</tbody>
</table>

$$C/x^a = C \times x^{-a}$$

Wordrank frequency diagram on Wikipedia data sets (log-log scale)
Biggest distance occurs at rank 5 for all fits.
Biggest distance occurs at rank 5 for all fits

Wordrank frequency diagram on Wikipedia data sets (top ranks)

Maybe our distance measure was not suitable
Can we do better?

$d = 115627.15$

$d = 28.1$
Problems with our 1\textsuperscript{st} approach

- We measured the largest \textit{point wise} distance between observed data and fit.

- One outlier enough to skew our result

- Millions of low rank distances will not contribute to the result even if they are all off.
L1 Norm (integral) cumulate point wise error

\[ \| f \|_1 = \int_\Omega |f(x)| \, d\mu(x) \]

L1-Norm of f in our case:

\[ \| f \|_1 = \sum_{x \in \Omega} |f(x)| \]

Let us define a distance:

\[ d_1(f_{obs}, f_{fit}) := \| f_{obs} - f_{fit} \|_1 \]
Now “-1.0” seems to be the best exponent

<table>
<thead>
<tr>
<th>$f_{\text{fit}}$</th>
<th>$d_1(f_{\text{obs}}, f_{\text{fit}})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$C/x^{0.9}$</td>
<td>11 M</td>
</tr>
<tr>
<td>$C/x^{1.0}$</td>
<td>4.9 M</td>
</tr>
<tr>
<td>$C/x^{1.1}$</td>
<td>6.7 M</td>
</tr>
</tbody>
</table>

\[ C/x^a = C \times x^{-a} \]
Problems with our 2\textsuperscript{nd} approach

- One outlier is still enough to skew our result

- Result is not normalized
  - It can be an arbitrary large number
  - We don’t know is 720M a good fit
  - Will better fits exist?
3rd approach: Study the cumulative plots

CDF of word rank frequency diagram on Wikipedia data sets (log scale)
“-1.1” is now the best exponent we can find

CDF of word rank frequency diagram on Wikipedia data sets (log scale)
Now “-1.1” seems to be the best exponent

<table>
<thead>
<tr>
<th>$f_{fit}$</th>
<th>$d_{ks}(f_{obs}, f_{fit})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$C/x^{0.9}$</td>
<td>0.29</td>
</tr>
<tr>
<td>$C/x^{1.0}$</td>
<td>0.15</td>
</tr>
<tr>
<td>$C/x^{1.1}$</td>
<td>0.11</td>
</tr>
</tbody>
</table>

$$C/x^a = C \times x^{-a}$$

CDF of word rank frequency diagram on Wikipedia data sets (log scale)
3rd way is called Kolmogorov Smirnov Test

• Cancelling out positive and negative errors

• Wide spread statistical test for fitting tasks

• Implemented in many fitting libraries

• Even though it is wide spread it is still a modelling choice
Comparing the results of the thee distance measures

<table>
<thead>
<tr>
<th>fit</th>
<th>Uniform norm (point wise distance)</th>
<th>L1-norm (cumulated error)</th>
<th>Kolmogorov Smirnov (uniform norm on CDF) – Mix of 1 and 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>( C/x^{0.9} )</td>
<td>115 k</td>
<td>11 M</td>
<td>0.29</td>
</tr>
<tr>
<td>( C/x^{1.0} )</td>
<td>148 k</td>
<td>4.9 M</td>
<td>0.15</td>
</tr>
<tr>
<td>( C/x^{1.1} )</td>
<td>177 k</td>
<td>6.7 M</td>
<td>0.11</td>
</tr>
</tbody>
</table>
We can characterize our data with the help of the Zipf parameter

• The **exponent** of the best fitting function is called the **Zipf parameter**

• Obviously the parameter depends on the choice of distance measure in our "**meta-model**"
  – Beware: Modelling choices change results!

• **Non trivial task** to find the best parameter
  – We just guessed and tested 3 values
  – Next unit: Estimate the parameter directly without guessing
Thank you for your attention!

Contact:
Rene Pickhardt
Institute for Web Science and Technologies
Universität Koblenz-Landau
rpickhardt@uni-koblenz.de
Copyright:

- This Slide deck is licensed under creative commons 3.0. share alike attribution license. It was created by Rene Pickhardt. You can use share and modify this slide deck as long as you attribute the author and keep the same license. All graphics unless otherwise stated have been self made by Rene Pickhardt and are also licensed under CC-BY-SA 3.0