Connecting the dots with common sense and linear models

Léon Bottou

NEC Labs America

COS 424 – 2/4/2010
Introduction

Useful things:
– understanding probabilities,
– understanding statistical learning theory,
– knowing countless statistical procedures,
– knowing countless machine learning algorithms.

Essential things:
– applying common sense,
– paying attention to details,
– being able to setup experiments,
– and to measure the outcome of experiments,
– and to measure plenty of other things,
**Question:**
Find $y$ given $x$.

<table>
<thead>
<tr>
<th>$x$</th>
<th>$y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.31</td>
<td>1.87</td>
</tr>
<tr>
<td>0.25</td>
<td>1.84</td>
</tr>
<tr>
<td>3.78</td>
<td>2.23</td>
</tr>
<tr>
<td>3.30</td>
<td>3.04</td>
</tr>
<tr>
<td>3.83</td>
<td>2.68</td>
</tr>
<tr>
<td>-3.29</td>
<td>0.01</td>
</tr>
<tr>
<td>-0.90</td>
<td>0.37</td>
</tr>
<tr>
<td>-3.61</td>
<td>0.37</td>
</tr>
<tr>
<td>0.64</td>
<td>2.05</td>
</tr>
<tr>
<td>-0.34</td>
<td>0.96</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Connecting the dots

Question:
Find $y$ given $x$.

<table>
<thead>
<tr>
<th>$x$</th>
<th>$y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.31</td>
<td>1.87</td>
</tr>
<tr>
<td>0.25</td>
<td>1.84</td>
</tr>
<tr>
<td>3.78</td>
<td>2.23</td>
</tr>
<tr>
<td>3.30</td>
<td>3.04</td>
</tr>
<tr>
<td>3.83</td>
<td>2.68</td>
</tr>
<tr>
<td>-3.29</td>
<td>0.01</td>
</tr>
<tr>
<td>-0.90</td>
<td>0.37</td>
</tr>
<tr>
<td>-3.61</td>
<td>0.37</td>
</tr>
<tr>
<td>0.64</td>
<td>2.05</td>
</tr>
<tr>
<td>-0.34</td>
<td>0.96</td>
</tr>
<tr>
<td>-3.53</td>
<td>-0.35</td>
</tr>
<tr>
<td>1.63</td>
<td>3.18</td>
</tr>
</tbody>
</table>

Answer:
Connect the dots. Read the curve.
## Connecting the dots – take two

**Question:** Find $y$ given $x$.

<table>
<thead>
<tr>
<th></th>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$x_3$</th>
<th>$x_4$</th>
<th>$x_5$</th>
<th>$x_6$</th>
<th>$x_7$</th>
<th>$x_8$</th>
<th>...</th>
<th>$x_{13,123}$</th>
<th>$x_{13,124}$</th>
<th>$x_{13,125}$</th>
<th>$y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.39</td>
<td>0.50</td>
<td>5.84</td>
<td>-4.36</td>
<td>-0.01</td>
<td>7.20</td>
<td>-7.40</td>
<td>-7.16</td>
<td>...</td>
<td></td>
<td>-5.48</td>
<td>0.77</td>
<td>5.03</td>
<td>5.46</td>
</tr>
<tr>
<td>7.34</td>
<td>1.92</td>
<td>-5.66</td>
<td>-5.33</td>
<td>-6.15</td>
<td>-3.14</td>
<td>4.53</td>
<td>6.37</td>
<td>...</td>
<td></td>
<td>-2.30</td>
<td>6.45</td>
<td>5.10</td>
<td>5.18</td>
</tr>
<tr>
<td>2.27</td>
<td>4.57</td>
<td>4.18</td>
<td>-6.07</td>
<td>-5.47</td>
<td>-6.97</td>
<td>2.67</td>
<td>-3.93</td>
<td>...</td>
<td></td>
<td>2.77</td>
<td>7.46</td>
<td>4.84</td>
<td>6.97</td>
</tr>
<tr>
<td>1.09</td>
<td>-2.17</td>
<td>-6.38</td>
<td>5.66</td>
<td>-2.65</td>
<td>-2.81</td>
<td>-0.69</td>
<td>2.76</td>
<td>...</td>
<td></td>
<td>0.42</td>
<td>5.88</td>
<td>0.29</td>
<td>-7.13</td>
</tr>
<tr>
<td>2.85</td>
<td>1.79</td>
<td>6.22</td>
<td>1.34</td>
<td>-1.83</td>
<td>3.01</td>
<td>3.99</td>
<td>-1.75</td>
<td>...</td>
<td></td>
<td>0.03</td>
<td>1.55</td>
<td>-3.32</td>
<td>-5.42</td>
</tr>
<tr>
<td>-5.67</td>
<td>2.53</td>
<td>-3.47</td>
<td>-0.46</td>
<td>3.21</td>
<td>-2.73</td>
<td>6.65</td>
<td>-0.77</td>
<td>...</td>
<td></td>
<td>-1.41</td>
<td>-3.93</td>
<td>3.14</td>
<td>5.37</td>
</tr>
<tr>
<td>3.80</td>
<td>-0.00</td>
<td>1.89</td>
<td>3.24</td>
<td>2.30</td>
<td>-1.45</td>
<td>7.63</td>
<td>-2.12</td>
<td>...</td>
<td></td>
<td>6.47</td>
<td>2.04</td>
<td>3.58</td>
<td>-4.96</td>
</tr>
<tr>
<td>7.54</td>
<td>2.47</td>
<td>6.39</td>
<td>4.95</td>
<td>-2.51</td>
<td>-6.46</td>
<td>0.49</td>
<td>-0.61</td>
<td>...</td>
<td></td>
<td>5.10</td>
<td>1.90</td>
<td>1.79</td>
<td>3.20</td>
</tr>
<tr>
<td>-7.99</td>
<td>4.93</td>
<td>-2.13</td>
<td>-7.11</td>
<td>-5.10</td>
<td>2.13</td>
<td>6.31</td>
<td>7.00</td>
<td>...</td>
<td></td>
<td>1.71</td>
<td>-2.35</td>
<td>-7.87</td>
<td>-4.70</td>
</tr>
<tr>
<td>-6.80</td>
<td>7.33</td>
<td>-0.99</td>
<td>4.17</td>
<td>-7.81</td>
<td>-7.64</td>
<td>4.01</td>
<td>-3.37</td>
<td>...</td>
<td></td>
<td>7.29</td>
<td>-2.41</td>
<td>7.66</td>
<td>-6.70</td>
</tr>
<tr>
<td>-0.78</td>
<td>5.34</td>
<td>-5.94</td>
<td>-1.76</td>
<td>3.79</td>
<td>2.92</td>
<td>0.75</td>
<td>7.04</td>
<td>...</td>
<td></td>
<td>-3.87</td>
<td>-1.46</td>
<td>-3.37</td>
<td>-3.66</td>
</tr>
<tr>
<td>7.54</td>
<td>2.47</td>
<td>6.39</td>
<td>4.95</td>
<td>-2.51</td>
<td>-6.46</td>
<td>0.49</td>
<td>-0.61</td>
<td>...</td>
<td></td>
<td>5.10</td>
<td>1.90</td>
<td>1.79</td>
<td>3.20</td>
</tr>
<tr>
<td>-7.99</td>
<td>4.93</td>
<td>-2.13</td>
<td>-7.11</td>
<td>-5.10</td>
<td>2.13</td>
<td>6.31</td>
<td>7.00</td>
<td>...</td>
<td></td>
<td>1.71</td>
<td>-2.35</td>
<td>-7.87</td>
<td>-4.70</td>
</tr>
<tr>
<td>-6.80</td>
<td>7.33</td>
<td>-0.99</td>
<td>4.17</td>
<td>-7.81</td>
<td>-7.64</td>
<td>4.01</td>
<td>-3.37</td>
<td>...</td>
<td></td>
<td>7.29</td>
<td>-2.41</td>
<td>7.66</td>
<td>-6.70</td>
</tr>
</tbody>
</table>

**Idea:** (1) understand how we do the 2D case. (2) generalize!
A Simple Linear Model

Polynomial: \( f(x) = w_0 + w_1 x + w_2 x^2 + \cdots + w_n x^n \)

Slight generalization:

\[
\begin{align*}
x \quad \rightarrow \quad \Phi(x) &= \begin{bmatrix}
\phi_0(x) \\
\phi_1(x) \\
\vdots \\
\phi_n(x)
\end{bmatrix} \\
\rightarrow \quad f(x) &= [w_0, w_1, \ldots, w_n] \times \\
&= \begin{bmatrix}
\phi_0(x) \\
\phi_1(x) \\
\vdots \\
\phi_n(x)
\end{bmatrix}
\end{align*}
\]

Equivalently: \( f(x) = w^\top \Phi(x) \)

Lets choose a basis \( \Phi \) and use the data to determine \( w \).
Linear Least Squares

Input: \( x_i \)
Output: \( w^\top \Phi(x_i) \)
Desired Output: \( y_i \)

Difference: \( y_i - w^\top \Phi(x_i) \)

Minimize: \( C(w) = \sum_{i=1}^{n} (y_i - w^\top \Phi(x_i))^2 \)

Quadratic convex function in \( w \).
The minimum exists and is unique.
But it could be reached for multiple values of \( w \).
At the optimum, 
\[
\frac{dC}{dw} = \sum_{i=1}^{n} 2 \left( y_i - w^\top \Phi(x_i) \right) \Phi(x_i)^\top = 0
\]

Therefore we must solve the system of equations:

\[
\begin{bmatrix}
\sum_{i=1}^{n} \Phi(x_i) \Phi(x_i)^\top
\end{bmatrix} \times w = \begin{bmatrix}
\sum_{i=1}^{n} y_i \Phi(x_i)
\end{bmatrix}
\]

Shorthand form:

\[
\left( X^\top X \right) w = \left( X^\top Y \right)
\]
Almost the same as \[ w = \left( X^\top X \right)^{-1} ( X^\top Y ). \]

You should never solve a system by inverting a matrix.

Who said \( X^\top X \) is invertible?

Consider the case where \( \phi_1(x) = \phi_8(x) \)
  – the matrix \( X^\top X \) is singular.
  – but the minimum is unchanged.
  – the minimum is reached by many \( w \),
    as long as \( w_1 + w_8 \) remains constant.

Among the \( w \) that minimize \( C(w) \),
compute the one with the smallest norm.
Numerical Procedures

Diagonalization of $X^\top X$

$$Q^\top D Q w = X^\top Y \iff w = Q^\top D^+ Q X^\top Y$$

Traditional methods: SVD or QR decomposition of $X$

$$V D U^\top U D V^\top w = V D U^\top Y \iff w = V D^+ U^\top Y$$

$$R^\top Q^\top Q R w = R^\top Q^\top Y \iff R w = Q^\top Y$$

and solve using back-substitution.

Simple and Fast: Regularization + Cholesky

$$\min C(w) + \varepsilon w^2 \iff (X^\top X + \varepsilon I) w = (X^\top Y)$$

$$\iff U U^\top w = (X^\top Y)$$

and solve using two rounds of back-substitution.
Polynomial degree 1

\[ \Phi(x) = 1, x \]
Polynomial degree 2

\( \Phi(x) = 1, x, x^2 \)
Polynomial degree 3

\[ \Phi(x) = 1, x, x^2, x^3 \]
Polynomial degree 6

\[ \Phi(x) = 1, x, x^2, x^3, x^4, x^5, x^6 \]
Polynomial degree 9

\[ \Phi(x) = 1, x, x^2, \ldots, x^9 \]
Polynomial degree 12

\[ \Phi(x) = 1, x, x^2, \ldots, x^{12} \]
Polynomial degree 20

$\Phi(x) = 1, x, x^2, \ldots, x^{20}$
Polynomials of the form $x^k$ quickly become very steep. There are much better polynomial bases: e.g. Chebyshev, Hermite, …
Mean squared error for polynomial models

Training set MSE:
\[ \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{f}(x_i))^2 \]

True MSE:
\[ \frac{1}{8} \int_{-4}^{+4} \sigma_{\text{true}}^2 + (f_{\text{true}}(x) - \hat{f}(x))^2 \, dx \]

Is MSE a good measure of the error?

Why integrating on \([-4, +4]\)?
About Error Measures

Domain
– should be related to the input data distribution.

Metric
– Uniform metric: $L_\infty$
– Averaged with a $L_p$ norm, e.g. MSE.

Derivatives
– Very close functions can have very different derivatives.
– Sobolev metrics.

Integrals
– Conversely, very close functions always have very close integrals.
Choose **knots** $r_1 \ldots r_k$

\[
\phi_0(x) = 1 \\
\phi_1(x) = x \\
\phi_2(x) = \max(0, x - r_1) \\
\ldots \\
\phi_j(x) = \max(0, x - r_{j-1})
\]
Piecewise Linear Models

Piecewise linear with 5 knots

Piecewise linear with 9 knots

Piecewise linear with 18 knots
MSE for Piecewise Linear Models

Training set MSE:

\[ \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{f}(x_i))^2 \]

True MSE:

\[ \frac{1}{8} \int_{-4}^{+4} \sigma_{true}^2 + (f_{true}(x) - \hat{f}(x))^2 \, dx \]
**Counting the dimensions**
- Linear functions on \( K + 1 \) segments: \( 2K + 2 \) parameters.
- Continuity constraints: \( K \) constraints.
- Other constraints: 0 (hinges), 1 (ramps), 2 (triangles).

\[
\text{dim}(\Phi) = K + 1
\]

\[
\text{dim}(\Phi) = K
\]
Piecewise Linear Variants

Piecewise ramps with 6 knots

Piecewise triangles with 7 knots
- Quadratic splines: $\Phi(x) = 1, x, x^2, \ldots \max(0, x - r_k)^2 \ldots$
- Cubic splines: $\Phi(x) = 1, x, x^2, x^3, \ldots \max(0, x - r_k)^3 \ldots$
Quadratic Splines

 Piecewise quadratic with 1 knot

 Piecewise quadratic with 6 knots

 Piecewise quadratic with 12 knots
MSE for Quadratic Splines

Training set MSE:
\[ \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{f}(x_i))^2 \]

True MSE:
\[ \frac{1}{8} \int_{-4}^{+4} \sigma_{\text{true}}^2 + (f_{\text{true}}(x) - \hat{f}(x))^2 \, dx \]
Changing the training data: more examples

Polynomial d=12

- 30 examples
- 300 examples
Changing the training data: less noise

Polynomial $d=12$

Noise $sdev=0.5$

Noise $sdev=0.1$
First Conclusions

The fancier the models, the higher the price.
– We can pay with more data.
– We can pay with better data.

In practice we do the converse.
– Changing the data is usually more costly than changing the model.
– Adapt the model “capacity” to the data.
– No shortage of methods.

The validation questions.
– We have too many options. How to choose one?
– How to estimate the quality of our work?
Estimate the quality of our work

Performance on the training data is not convincing
– Cannot distinguish between learning by rote and understanding.
– Understanding leads to more useful predictions than learning by rote.
– Therefore we need fresh data to evaluate our work.

• Testing examples set aside before starting the work.
  – Statistics work for randomly picked testing examples.
  – Real life suggests selected testing examples (e.g. time series.)

• Testing data of a different nature.
  – New perspective on the same phenomenon.
  – Often more instructive and convincing.

What about the “elegance” of a model ?
– Einstein: “Make everything as simple as possible, but not simpler.”
– How do you define “simple” ?
One should only use the testing set once! Of course...
- The more we look at the testing set, the less convincing we are.
- Public benchmarks and their problems.
The “validation set”

How to select the right model without looking at the testing set?

1. Set aside test set

2. Estimate multiple models using only the training set.

3. Compare models using the validation set and select the final model.

4. Retrain selected model using both validation and training set. (there are variants on that.)

5. Finally evaluate the performance of the selected model using the testing set. *Ideally this happens only once!*
Potential problems

All this consumes valuable examples!
– This is a serious problem when examples are rare!

What is the optimal size of the testing set?
– Large enough to measure the performance with sufficient accuracy.

What is the optimal size of the validation set?
– Large enough to justify our model selection, but not larger!
– Depends on the number of models to compare.
– Depends on the data needs of the models we compare.
– Depends on the total size of the data set.
– Trial and errors...
K-fold cross validation

1. Set aside test set

Repeat:

2. Estimate multiple models using the training set minus part Ti.

3. Measure their performances on Ti.

4. Select the final model on the basis of its average performance.

5. Retrain selected model using the full training set.

6. Finally evaluate the performance of the selected model using the testing set. Ideally this happens only once!
Potential problems

All this consumes valuable computing time!
– This is a serious problem when examples are abundant.

How accurate is k-fold cross-validation?
– More than using a single partition as validation set.
– Less than using a validation set as large as the training set.
– The statistical properties of the procedure are unclear.

Suggestions
– Avoid k-fold cross validation for very large datasets.
– Observe the variations of measured performances on the folds.

Subtleties
– Evaluating the performance of a trained model.
– Evaluation the performance of a training procedure.
Beyond Curve Fitting

\[ x \rightarrow \Phi(x) = \begin{bmatrix} \phi_0(x) \\ \phi_1(x) \\ \vdots \\ \phi_n(x) \end{bmatrix} \rightarrow f(x) = [w_0, w_1, \ldots, w_n] \times \begin{bmatrix} \phi_0(x) \\ \phi_1(x) \\ \vdots \\ \phi_n(x) \end{bmatrix} \]

Given suitable basis functions \( \Phi \), the inputs \( x \) could be anything.
- numerical variables, e.g. \( 3.1415 \)
- categorical variables, e.g. blue, green, yellow, ...
- ordered variables, e.g. small, medium, large.
- complex data structures, such as trees, graphs, etc.
- any combination of the above.

This does not mean that constructing the features \( \phi_i(x) \) will be easy.
The “adult” dataset

Predict whether income exceeds $50K/year ($y = +1$) or not ($y = -1$).

http://archive.ics.uci.edu/ml/datasets/Adult

**Input variables**
- 6 continuous variables:
  - age, years of education, hours-per-week,
  - capital-gains, capital-losses, fnlwgt(?).
- 8 categorical variables:
  - workclass, education, marital status, sex,
  - occupation, race, relationship, native country.

**Training and testing sets**
- Training set: 32561 examples
- Testing set: 16281 examples
Creating $\Phi(x)$ for the adult dataset

**Coding on 1+123 binary features $\phi_i(x)$**
- First feature is always $\phi_1(x) = 1$.
- One feature for each possible value of each categorical variable.
- Five features for each continuous variable (quantified on 5 quantiles).

copied from (Platt, 1998)

**Split**
- 28000 training + 4562 validation examples.
- 16281 testing examples.

**Results**

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Misclassification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Validation set (after training on 28K)</td>
<td>15.98 %</td>
</tr>
<tr>
<td>Testing set (after training on 32K)</td>
<td>15.47 %</td>
</tr>
</tbody>
</table>
A quadratic basis for the adult dataset

Coding on $1 + 123 + 7503$ features

– Additional features for quadratic models.

\[ \forall i \in 1 \ldots 123 \quad \forall j \in 1 \ldots i - 1 \quad \phi_{ij}(x) = \phi_i(x) \phi_j(x) \]

Remarks

– Feature count grows quickly.
– This is slow ($X$ is sparse, but $X^\top X$ is not.)

Results

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Misclassification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Validation set (after training on 28K)</td>
<td>16.40 %</td>
</tr>
<tr>
<td>Testing set (after training on 32K)</td>
<td>— %</td>
</tr>
</tbody>
</table>
Weighting the quadratic terms

Idea
Remember the regularization + cholevsky trick?

\[
\min C(w) + \varepsilon w^2 \iff (X^\top X + \varepsilon I) w = (X^\top Y)
\]

Let’s penalize more the coefficients of the quadratic terms.

\[
\min C(w) + w^\top \Lambda w \iff (X^\top X + \Lambda) w = (X^\top Y)
\]

Details
– \( \varepsilon = 10^{-5} \) for constant and linear terms.
– \( \varepsilon \in [10^{-5}, 10^5] \) for quadratic terms.
Weighting the quadratic terms

We get the linear result when $\epsilon \to \infty$.

We get the quadratic result when $\epsilon \to 0$.

After retraining with $\epsilon = 100$ on all 32K examples:
Testing set error: 14.93 %.
Coming next

Homework 1
– Due on Tue Feb 23rd.
– Something about splines.

Next lectures
– Tuesday Feb 9th: R tutorial (Sean Gerrish)
– Thursday Feb 11th: Review of probabilities