Least Squares Data Fitting

Stephen Boyd

EE103
Stanford University

November 4, 2015
Outline

Least-squares data fitting

Validation
we believe a scalar $y$ and an $n$-vector $x$ are related by model

$$y \approx f(x)$$

- $x$ is called the *independent variable*
- $y$ is called the *outcome or response variable*
- $f : \mathbb{R}^n \to \mathbb{R}$ gives the relation between $x$ and $y$
- often $x$ is a feature vector, and $y$ is something we want to predict
- we don’t know $f$, which gives the ‘true’ relationship between $x$ and $y$
Model

- choose model \( \hat{f} : \mathbb{R}^n \rightarrow \mathbb{R} \), a guess or approximation of \( f \), based on some observed data

\[(x_1, y_1), \ldots, (x_N, y_N)\]

called observations, examples, samples, or measurements

- model form:

\[
\hat{f}(x) = \theta_1 f_1(x) + \cdots + \theta_p f_p(x)
\]

- \( f_i : \mathbb{R}^n \rightarrow \mathbb{R} \) are basis functions that we choose
- \( \theta_i \) are model parameters that we choose
- \( \hat{y}_i = \hat{f}(x_i) \) is (the model's) prediction of \( y_i \)
- we’d like \( \hat{y}_i \approx y_i \), i.e., model is consistent with observed data

Least-squares data fitting
Least-squares data fitting

- *prediction error or residual* is \( r_i = \hat{y}_i - y_i \)
- express \( y \), \( \hat{y} \), and \( r \) as \( N \)-vectors
- \( \text{rms}(r) \) is *RMS prediction error*
- *least-squares data fitting*: choose \( \theta_i \) to minimize RMS prediction error
- define \( N \times p \) matrix \( A \), \( A_{ij} = f_j(x_i) \) so \( \hat{y} = A\theta \)
- least-squares data fitting: choose \( \theta \) to minimize \( \| r \|^2 = \| A\theta - y \|^2 \)
- \( \hat{\theta} = (A^TA)^{-1}A^T y \) (if columns of \( A \) are independent)
- \( \| A\hat{\theta} - y \|^2 / N \) is *minimum mean-square (fitting) error*
Fitting a constant model

- simplest possible model: \( p = 1, \, f_1(x) = 1 \), so model \( \hat{f}(x) = \theta_1 \) is a constant function
- \( A = 1 \), so
  \[
  \hat{\theta}_1 = (1^T 1)^{-1} 1^T y = \frac{1}{N} 1^T y = \text{avg}(y)
  \]
- the mean of \( y \) is the least-square fit by a constant
- MMSE is \( \text{std}(y)^2 \); RMS error is \( \text{std}(y) \)
- more sophisticated models are judged against the constant model
Fitting univariate functions

- when \( n = 1 \), we seek to approximate a function \( f : \mathbb{R} \to \mathbb{R} \)
- we can plot the data \((x_i, y_i)\) and the model function \( \hat{y} = \hat{f}(x) \)
Straight-line fit

- $p = 2$, with $f_1(x) = 1$, $f_2(x) = x$
- model has form $\hat{f}(x) = \theta_1 + \theta_2 x$
- matrix $A$ has form

$$A = \begin{bmatrix}
1 & x_1 \\
1 & x_2 \\
\vdots & \vdots \\
1 & x_N
\end{bmatrix}$$

- can work out $\hat{\theta}_1$ and $\hat{\theta}_2$ explicitly:

$$\hat{f}(u) = \text{avg}(y) + \rho \frac{\text{std}(y)}{\text{std}(x)} (u - \text{avg}(x))$$

(but QR works fine . . . )
Example

\[ \hat{f}(x) \]

Least-squares data fitting
Asset $\alpha$ and $\beta$

- $x_i$ is return of whole market, $y_i$ is return of a particular asset
- write straight-line model as

$$\hat{y} = (r^{rf} + \alpha) + \beta(x - \text{avg}(x))$$

- $r^{rf}$ is the risk-free interest rate
- several other slightly different definitions are used
- called asset ‘$\alpha$’ and ‘$\beta$’, widely used

Least-squares data fitting
Time series trend

- $y_i$ is value of quantity at time $x_i$
- common case: $x_i = i$
- $\hat{y} = \hat{\theta}_1 + \hat{\theta}_2 x$ is called trend line
- $y - \hat{y}$ is called de-trended time series
- $\hat{\theta}_2$ is trend coefficient
World petroleum consumption

[Graph showing the trend of world petroleum consumption from 1980 to 2010. The x-axis represents the year, and the y-axis represents petroleum consumption in million barrels per day.]
World petroleum consumption, de-trended

[Graph showing the trend of petroleum consumption from 1980 to 2010]

Petroleum consumption (million barrels per day)

Year

1980 1990 2000 2010

Least-squares data fitting
Polynomial fit

- \( f_i(x) = x^{i-1}, \quad i = 1, \ldots, p \)
- model is a polynomial of degree less than \( p \)
  \[
  \hat{f}(x) = \theta_1 + \theta_2 x + \cdots + \theta_p x^{p-1}
  \]

- \( A \) is a Vandermonde matrix
  \[
  A = \begin{bmatrix}
  1 & x_1 & \cdots & x_1^{p-1} \\
  1 & x_2 & \cdots & x_2^{p-1} \\
  \vdots & \vdots & \ddots & \vdots \\
  1 & x_N & \cdots & x_N^{p-1}
  \end{bmatrix}
  \]
Examples

- $N = 20$ data points
- fits of degree $p - 1 = 3, 5, \text{ and } 10$
Regression as general data fitting

- regression model is affine function $\hat{y} = \hat{f}(x) = x^T \beta + v$
- fits general fitting form with basis functions

$$f_1(x) = 1, \quad f_i(x) = x_{i-1}, \quad i = 2, \ldots, n + 1$$

so model is

$$\hat{y} = \theta_1 + \theta_2 x_1 + \cdots + \theta_{n+1} x_n = x^T \theta_{2:n} + \theta_1$$

- $\beta = \theta_{2:n+1}$, $v = \theta_1$
General data fitting as regression

- general fitting model $\hat{f}(x) = \theta_1 f_1(x) + \cdots + \theta_p f_p(x)$
- common assumption: $f_1(x) = 1$
- same as regression model $\hat{f}(\tilde{x}) = \tilde{x}^T \beta + v$, with
  - $\tilde{x} = (f_2(x), \ldots, f_p(x))$ are ‘transformed features’
  - $v = \theta_1$, $\beta = \theta_2:p$
Outline

Least-squares data fitting

Validation
Generalization

basic idea:

▶ goal of model is *not* to predict outcome in the given data
▶ instead it is to *predict the outcome on new, unseen data*

▶ a model that makes reasonable predictions on new, unseen data has *generalization ability*
▶ a model that makes poor predictions on new, unseen data is said to suffer from *over-fit*
Validation

a simple and effective method to guess if a model will generalize

- split original data into a *training set* and a *test set*
- typical splits: 80%/20%, 90%/10%
- build (‘train’) model on training data set
- then *check the model’s predictions on the test data set*
- (can also compare RMS prediction error on train and test data)
- if they are similar, we can *guess* the model will generalize
Validation

- can be used to choose among different candidate models, e.g.
  - polynomials of different degrees
  - regression models with different sets of regressors
- we’d use one with low, or lowest, test error
Example

- polynomial models from 20 training points above
- evaluated below with 20 new test points
Example

- suggests degree 5 or 6 polynomial would be good choice