Computational Methods
CMSC/AMSC/MAPL 460

Least squares method: linear regression

Ramani Duraiswami,
Dept. of Computer Science
Fitting data to a model

• Practical science involves lots of fitting of data to models
• Difference between fitting and interpolation?
  – Interpolation, the fit function passes through the point
  – Fitting, the fit function satisfies some norm based criterion
• Tasks arise commonly in science
  – Fit straight lines and curves to data
  – More generally fit data to a model
• Model contains parameters
  – Job of fitting is to estimate the parameters that “best” make the model fit the data
  – “best” → define best
• Simplest example of model fitting problem
  – Linear regression
Models

• Have a certain model structure
  – E.g., “linear” “quadratic” “trigonometric” “Gaussian”
• Models have specifiable parameters
  • e.g.

<table>
<thead>
<tr>
<th>Model</th>
<th>Structure</th>
<th>Data</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Straight line:</td>
<td>$ax + by + c = 0$</td>
<td>$(x_i, y_i)$</td>
<td>$(a, b, c)$</td>
</tr>
<tr>
<td>Polynomial:</td>
<td>$y = c_0 + c_1 x + \ldots + c_n x^n$</td>
<td>$(x_i, y_i)$</td>
<td>$(c_0, c_1, \ldots, c_n)$</td>
</tr>
</tbody>
</table>

• General model
• $y(t) \simeq \beta_1 \phi_1(t) + \ldots + \beta_n \phi_n(t) +$
• Data $(y_i, t_i)$ and
• $y = \Phi \beta$
• Solve $\beta = y \mid \Phi$
Relationships among Variables

• In much science we seek relations between variables

One variable is used to “explain” another variable

\[ X \text{ Variable} \]

- Independent Variable
- Explaining Variable
- Exogenous Variable
- Predictor Variable

\[ Y \text{ Variable} \]

- Dependent Variable
- Response Variable
- Endogenous Variable
- Criterion Variable
Scatter Plots
Simple Least-Squares Regression

If we had errorless predictions:

\[ Y = a + bX \]
We will end up being reasonably confident that the true regression line is somewhere in the indicated region.
Estimated Regression Line
Estimated Regression Line
Estimated Regression Line

Equation of the Regression Line:
\[ \hat{Y} = a + bX \]

\[ e_i = y_i - \hat{y}_i \]
Remember: \( \sum_{i=1}^{N} (y_i - \bar{y}) = 0 \)

We will now also have: \( \sum_{i=1}^{N} e_i = \sum_{i=1}^{N} (y_i - \hat{y}_i) = 0 \)

\[ e_i = y_i - \hat{y}_i \]
How do we find \( a \) and \( b \)?

In Least-Squares Regression:

Find \( a, b \) to minimize the sum of squared errors/residuals

\[
\sum_{i=1}^{N} (e_i)^2 = \sum_{i=1}^{N} (y_i - [b x_i + a])^2
\]
In Least-Squares Regression:

\[
b = \frac{\sum_{i=1}^{N} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{N} (X_i - \bar{X})^2}, \quad a = \bar{Y} - b\bar{X}
\]
Models

- Have a certain model structure
  - E.g., “straight line” “parabolic” “trigonometric” “Gaussian”
- Models have specifiable parameters
- e.g.

<table>
<thead>
<tr>
<th>Model</th>
<th>Structure</th>
<th>Data</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Straight line:</td>
<td>$ax + by + c = 0$</td>
<td>$(x_i, y_i)$</td>
<td>$(a, b, c)$</td>
</tr>
<tr>
<td>Polynomial:</td>
<td>$y = c_0 + c_1 x + \ldots + c_n x^n$</td>
<td>$(x_i, y_i)$</td>
<td>$(c_0, c_1, \ldots, c_n)$</td>
</tr>
<tr>
<td>Trig.:</td>
<td>$y = c_0 + c_1 \sin x + \ldots + c_n \sin nx$</td>
<td>$(x_i, y_i)$</td>
<td>$(c_0, c_1, \ldots, c_n)$</td>
</tr>
<tr>
<td>Gaussian</td>
<td>$y = c_0 \exp\left(-\frac{(x-\mu)^2}{\sigma^2}\right)$</td>
<td>$(x_i, y_i)$</td>
<td>$(c_0, \sigma, \mu)$</td>
</tr>
<tr>
<td>“Kernel/RBF”</td>
<td>$y = \sum_i a_i k(x, x_i)$</td>
<td>$(x_i, y_i)$</td>
<td>$(a_i)$ and $(x_i)$</td>
</tr>
</tbody>
</table>

- These models are all separable
Linear models

- All models considered are “linear” or “separable”
- (does not mean straight lines)
- Means that we can separate the “structure” and “parameters” as a matrix-vector product
- “structure” forms a matrix and “parameters” a vector
- Goal of model fitting: find the parameters
- Question: Is the Gaussian model a linear one?
  - Can be made separable by taking logs
    \[ y(t) \approx Ke^{\lambda t} \]
    \[ \log y \approx \beta_1 t + \beta_2, \text{ with } \beta_1 = \lambda, \beta_2 = \log K \]
  - However if there are sums of gaussians or exponentials the model is not separable
    \[ y(t) \approx \beta_1 e^{-\left(\frac{t-\mu_1}{\sigma_1}\right)^2} + \ldots \beta_n e^{-\left(\frac{t-\mu_n}{\sigma_n}\right)^2} \]
Linear Systems

\[ Ax = b \]

Square system:
- unique solution
- Gaussian elimination

Rectangular system ??
- underconstrained: infinity of solutions
- overconstrained: no solution

Minimize \( |Ax-b|^2 \)
Least Squares for more complex models

- Number of equations and unknowns may not match
- Data may have noise
- Look for solution by minimizing some cost function
- Simplest and most intuitive cost function: \( \|Ax - b\|_2 \)
- Define for each data point \( x_i \) a residual \( r_i \)
- Minimize \( \sum_i r_i^2 \) with respect to \( x_l \)

\[
\sum_i (A_{ij}x_j - b_i)^2 = \sum_k (A_{ik}x_k - b_i)^2
\]

\[
\frac{\partial}{\partial x_l} (A_{ij}x_j - b_i)(A_{ik}x_k - b_i) = 0
\]

\[
(A_{ij}\delta_{jl})(A_{ik}x_k - b_i) + (A_{ij}x_j - b_i)(A_{ik}\delta_{kl}) = 0
\]

\[
A_{il}(A_{ik}x_k - b_i) + (A_{ij}x_j - b_i)A_{il} = 2(A_{il}A_{ik}x_k - A_{il}b_i) = 0
\]

\[
A_{il}A_{ik}x_k = A_{il}b_i
\]
Other Norms

• Here we fit using the “least-squares” or $L_2$ norm
• Could minimize the residual in other norms
• For example we may have more confidence in some data, and want to be sure that their residual is lower
  – Attach a weight to each residual
    $$\| r \|^2_w = \sum_{i=1}^{m} w_i r_i^2$$
• Or we may like the $1$-norm or infinity norm better
  $$\| r \|_1 = \sum_{i=1}^{m} |r_i| \quad \| r \|_{\infty} = \max_i |r_i|$$
Normal equations

• The system $A^t A \, x = A^t b$
  is called the Normal equations

• Can solve least squares problems using these

• For A size $m \times n$ and $x$ of size $n$ and $b$ of size $m$ what are the dimensions of the normal equations?
  – $n \times n$

• Solve via LU decomposition

• Is this a good idea?
  – Somewhat expensive as we have to form $A^t A$ which involves matrix multiplication and then solution
  – More importantly it is poorly conditioned
  – $\text{cond}(A^t A) = (\text{cond}(A))^2$
• Example

\[ X = \begin{pmatrix} 1 & 1 \\
\delta & 0 \\
0 & \delta \end{pmatrix} \]

\[ X^T X = \begin{pmatrix} 1 + \delta^2 & 1 \\
1 & 1 + \delta^2 \end{pmatrix} \]